EVIDENCE IS LAYERED: EXPERT-GUIDED QUERIES FOR GENERALIZABLE IMAGE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained vision transformers (ViTs) provide transferable representations for generalizable image forgery detection, yet how forensic evidence is distributed across their layers and how to retrieve it effectively remain poorly understood. We investigate these questions with *LayerProbe*, a controlled study of 14 frozen ViT backbones spanning standalone and MLLM vision encoders, using layer-wise linear probing and greedy layer fusion. The results reveal that the best readout depth varies across backbones, while selected layers with lower single-layer scores can still complement the strongest layer. Motivated by these findings, we propose ForenLens, a frozen-backbone detector that uses queries to retrieve complementary forensic evidence from a compact, backbone-specific layer set selected by *LayerProbe*. ForenLens uses a shared cross-attention reader to extract layer-specific responses before combining them through query-adaptive routing, preserving layer identity and learning where each query reads for each image. Complementing this adaptive layer routing, training-only forensic guidance shapes what expert queries retrieve by supervising their spatial attention. On the OpenMMSec out-of-distribution split and seven external benchmarks, ForenLens achieves state-of-the-art performance with 93.2% mean AP and 87.1% mean accuracy, while remaining robust to post-processing. Our code will be available upon acceptance.

1 INTRODUCTION

Advances in image generation and editing have made realistic synthetic and manipulated images widely accessible. These tools support creative applications, but can also be used to fabricate visual evidence and spread misinformation. Reliable forgery detection must therefore work beyond the models and data seen during training. This remains difficult: detectors that perform well in distribution often fail on unseen generators, retrained variants, or images altered by common post-processing operations (Yan et al., 2025a; Li et al., 2025b; Cavia et al., 2024).

One line of detection research focuses on generator fingerprints, upsampling artifacts, or reconstruction discrepancies (Wang et al., 2020; Tan et al., 2024; Wang et al., 2023; Ricker et al., 2024). These cues can help distinguish real and forged images, but their reliability may change across generators and processing conditions. Another line of research builds on pretrained vision encoders, particularly vision transformers (ViTs). These encoders provide representations learned from large-scale visual data, which can be reused or adapted for forgery detection. For example, UnivFD uses features from a frozen CLIP ViT-L/14, while Effort and SICA adapt pretrained encoders under constraints designed to preserve useful representations (Ojha et al., 2023; Yan et al., 2025b; Du et al., 2026).

These methods demonstrate the forensic value of pretrained vision encoders. However, their overall detection performance does not reveal what each backbone layer actually contributes or whether different layers provide complementary evidence. We therefore examine both individual detection performance and cross-layer complementarity.

To answer these questions, we conduct *LayerProbe*, a controlled study of 14 frozen ViT backbones covering standalone vision encoders and the vision encoders of multimodal large language models (MLLMs). Keeping the backbones frozen lets us examine the evidence already present in their representations. We use linear probes to measure detection performance at each layer, then evaluate the gain from adding other layers to the strongest-layer baseline (Section 3). Two findings emerge. First, the best readout depth depends on the backbone: standalone vision encoders often peak before

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.

the final layer, while MLLM vision encoders tend to retain strong detection performance at greater depths (Figure 1). Second, selected layers with lower single-layer scores consistently improve the strongest-layer baseline when combined with it (Figure 2). Therefore, a layer's score in isolation is not enough to judge its value in a multi-layer readout.

LayerProbe identifies complementary layers in each backbone, but how to retrieve useful forensic evidence from these layers remains unsolved. To retrieve this evidence, we adopt learnable queries (Li et al., 2023), a small set of vectors that sit outside the frozen ViT and attend to a layer's patch tokens, collecting spatially localized forensic cues. Consequently, these queries naturally extend across layers and can be guided by external spatial priors, while their independence from the frozen ViT makes them applicable to diverse ViT backbones at a modest training cost.

Building on the above, we propose ForenLens, which uses these queries to retrieve forensic evidence from the layers selected by *LayerProbe*. Each query reads all selected layers and adaptively combines them into one evidence token for each image. To encourage different queries to focus on different cues, we supervise the spatial attention of expert-guided queries with training-only guidance maps from four forensic operators. We further introduce free queries, kept distinct from the expert-guided ones through an orthogonality constraint, to capture evidence beyond the expert priors.

We evaluate ForenLens on the 72 unseen forgery types from OpenMMSec OOD split (Du et al., 2026) and 51 test sets from seven external benchmarks. Across the eight benchmark groups, ForenLens achieves the highest mean AP and mean accuracy, at 93.2% and 87.1%, respectively. Ablation results support the contributions of query-based cross-layer retrieval and expert guidance, while experiments across 14 ViT backbones support the use of the same readout design for all of them. Additionally, post-processing tests show that ForenLens remains robust to common image degradations, and visualizations reveal the layers and image regions emphasized by different queries.

Our main contributions are as follows:

- We design *LayerProbe*, a controlled study combining layer-wise linear probing and greedy layer fusion across 14 frozen ViT backbones. It reveals backbone-dependent readout depths and complementary evidence from individually weaker layers.
- We introduce ForenLens, a frozen-backbone detector that uses queries with training-only expert guidance to retrieve complementary evidence across layers.
- Extensive evaluations demonstrate the state-of-the-art generalization of ForenLens, supported by ablations, experiments on backbones, and robustness analyses.

2 Related Work

Image synthesis and manipulation. Image synthesis has progressed from GANs (Goodfellow et al., 2014; Karras et al., 2020) to diffusion and latent-diffusion models (Ho et al., 2020; Rombach et al., 2022), with diffusion transformers, rectified flows, and visual autoregressive models expanding its capabilities (Peebles & Xie, 2023; Esser et al., 2024; Tian et al., 2024). Face swapping and instruction-guided editing enable diverse modifications to existing images (Li et al., 2020; Brooks et al., 2023). These varied processes leave different forensic traces, further affected by post-processing, motivating detection across generators and manipulation types.

Generalizable image forgery detection. Generalizable detection draws on synthesis artifacts (Wang et al., 2020; Tan et al., 2024; Fei et al., 2026), reconstruction-based cues and supervision (Wang et al., 2023; Ricker et al., 2024; Chen et al., 2024a), and transferable CLIP representations (Cozzolino et al., 2024). UnivFD uses frozen CLIP features (Ojha et al., 2023), while Effort and SICA constrain encoder adaptation to preserve useful pretrained structure (Yan et al., 2025b; Du et al., 2026). Related approaches use forgery-aware adapters and prompt learning (Liu et al., 2024; Wang et al., 2026b). RINE weights intermediate CLIP representations, bringing evidence from multiple encoder layers into the detector (Koutlis & Papadopoulos, 2024). DoU encourages feature diversity (He et al., 2026), while hybrid or patch-based approaches integrate semantic and low-level cues or emphasize local traces (Yan et al., 2025a; Cheng et al., 2025; Zhou et al., 2026; Wang et al., 2026a). Training-data design and image transformations further reduce spurious biases (Guillaro et al., 2025; Li et al., 2025a), while MLLM-based methods, including AIGI-Holmes, MARE, and Deep-VRM, bring semantic reasoning into forensic analysis (Zhou et al., 2025; Xu et al., 2026; Lin et al., 2026).

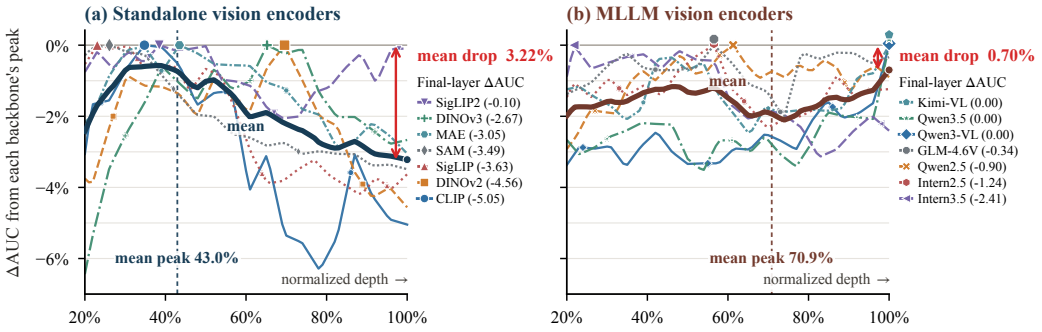

Figure 1: **The two backbone families retain forensic evidence differently with depth.** Thin curves track probe AUC relative to each backbone's own peak for standalone (left) and MLLM (right) vision encoders. Thick curves show family means, and vertical dashed lines mark the mean of individual peak depths. Arrows indicate the mean AUC drop from each backbone's peak to its final layer.

3 Empirical Motivation

Pretrained ViTs underpin generalizable forensic detectors (Ojha et al., 2023; Cozzolino et al., 2024; Yan et al., 2025b; Du et al., 2026), but detectors typically use a fixed readout from these backbones. Although some detectors exploit intermediate features, this alone does not reveal how forensic readability is reorganized across layers or whether different layers retain conditionally useful, non-redundant information. We use single-layer probes to measure detection performance at each depth and greedy layer fusion to test whether other layers add complementary evidence.

We hypothesize that semantic encoding redistributes forensic readability across layers. This yields two predictions: the best readout location should depend on the backbone's representational role, and a layer weak in isolation should still complement the strongest layer.

We apply a common *LayerProbe* protocol to 14 frozen ViT backbones. The seven standalone vision encoders are CLIP (Radford et al., 2021), DINOv2 and DINOv3 (Oquab et al., 2024; Siméoni et al., 2026), SigLIP and SigLIP2 (Zhai et al., 2023; Tschannen et al., 2025), MAE (He et al., 2022), and SAM (Kirillov et al., 2023). The seven MLLM vision encoders are taken from Qwen2.5-VL and Qwen3-VL (Bai et al., 2025b;a), Qwen3.5 (Qwen Team, 2026), InternVL2.5 and InternVL3.5 (Chen et al., 2024b; Wang et al., 2025), Kimi-VL (Kimi Team, 2025), and GLM-4.6V (GLM-V Team, 2025). At each Transformer block, an independent linear probe reads per-channel means and standard deviations computed across patch tokens. All the probes are fitted on the same 8,000 OpenMMSec training images and evaluated by AUC on a disjoint 8,000-image in-domain (ID) set.

Tracing evidence through the hierarchy. If the final layer, or any fixed normalized depth, were universally sufficient, cross-layer readout would be unnecessary. Our findings challenge this assumption. As seen in Figure 1, the strongest layer ranges from early layers (e.g., SigLIP) to the final layer. Standalone encoders peak at 43.0% on average and finish 3.22% (absolute) lower in AUC, while MLLM encoders peak at 70.9% and finish only 0.70% lower.

The contrast is systematic: standalone vision encoders usually peak before the final layer and then lose forensic readability, whereas MLLM vision encoders remain near peak much deeper, often up to the final layer. This pattern suggests that low-level artifacts encoded early in the hierarchy remain valuable forensic evidence, but the two families preserve these cues differently during semantic abstraction. Standalone vision encoders progressively form self-contained visual representations and attenuate these cues along the way; MLLM vision encoders retain fine-grained token-level content for subsequent multimodal reasoning, carrying more low-level forensic evidence into deeper layers and thereby sustaining forensic discriminability near the final layer.

Looking beyond the strongest layer. Architecture-dependent Top-1 selection alone does not establish complementarity: other layers may be weaker copies of the same signal. To test whether weaker layers provide non-redundant evidence, we conduct a greedy fusion experiment. Let ℓ_k be the k-th selected layer and A_k the validation AUC after fusing $\{\ell_1, \ldots, \ell_k\}$. Starting from ℓ_1, the procedure repeatedly refits the same probe and adds the layer with the largest AUC gain. If layers differed only in strength, it would select layers with high single-layer ranks and yield little improvement.

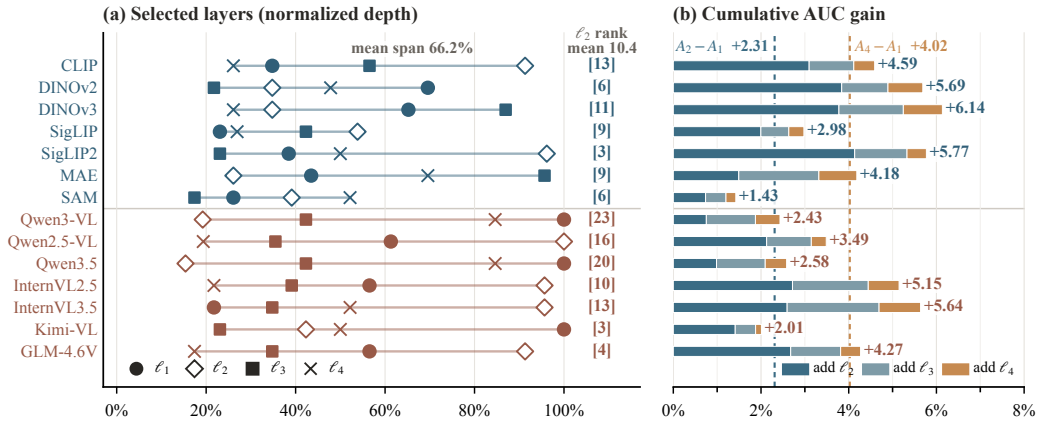

Figure 2: **Layers weak in isolation provide complementary evidence across the hierarchy.** (a) Locations of the four greedily selected layers ℓ_1–ℓ_4 across all 14 backbones. Brackets give the single-layer AUC rank of ℓ_2 (1 is best). (b) Cumulative validation-AUC improvement over A_1, decomposed into the marginal gains when each layer is added.

Figure 2 shows the opposite. The second selected layer ℓ_2 ranks only 10.4 on average in isolation, yet improves the fused AUC for all 14 backbones, with a mean $A_2 - A_1$ gain of 2.31% (absolute); the normalized depth span of the four selected layers averages 66.2% across backbones. This rank–gain mismatch and broad span show that single-layer quality differs from conditional usefulness: ℓ_2 is valuable not because it is strong in isolation, but because it supplies evidence missing from ℓ_1. Semantic encoding thus distributes partially non-redundant forensic components across layers rather than around a single optimum.

By jointly examining individual-layer detection performance and the incremental gains from greedy layer fusion, *LayerProbe* shows that the best readout depth varies across backbones and that complementary forensic evidence spans distant layers. A detector should therefore recover complementary evidence from a compact layer set without erasing layer-specific structure, deciding both where to read and what to retrieve.

4 METHOD

The above empirical study reveals complementary forensic evidence across ViT layers, motivating ForenLens. It uses a shared reader to retrieve evidence separately from the frozen ViT layers selected by *LayerProbe*, and then combines the responses through query-adaptive routing (Figure 3). Training-only expert guidance encourages complementary query roles, while free queries capture cues beyond these priors. To explain this design, we first describe cross-layer readout (Section 4.1), then expert guidance (Section 4.2), and finally decision readout and training (Section 4.3).

4.1 QUERY-BASED CROSS-LAYER READOUT

To retrieve complementary evidence across layers, we preserve layer-specific responses and learn how each query combines them for a given image.

Selected-layer patch memory. We use the ID-only greedy *LayerProbe* procedure (Section 3) to select a compact layer set $\mathcal{K} = \{\ell_1, \ldots, \ell_K\}$ (Top-K in greedy order) for each backbone. This set is fixed before detector training and reused at test time. The frozen ViT provides their patch memories $\boldsymbol{H}^{(\ell)} \in \mathbb{R}^{P \times d}$ in one forward pass per image, where P is the patch count and d the hidden width.

Shared cross-attention reader. Inspired by BLIP-2 (Li et al., 2023), we use learnable queries $\mathcal{Q}$ with embeddings $\boldsymbol{q}_q$. Each query reads all K selected layers separately through the same cross-attention:

$$(\boldsymbol{t}_{q\ell}, \boldsymbol{a}_{q\ell}) = \mathrm{MHA}\left(\boldsymbol{q}_q, \boldsymbol{H}^{(\ell)}, \boldsymbol{H}^{(\ell)}\right) \tag{1}$$

Here $\boldsymbol{t}_{q\ell} \in \mathbb{R}^d$ is the layer-specific response and $\boldsymbol{a}_{q\ell} \in \mathbb{R}^P$ is its spatial attention averaged over heads. Sharing the reader places responses from different layers in a common space.

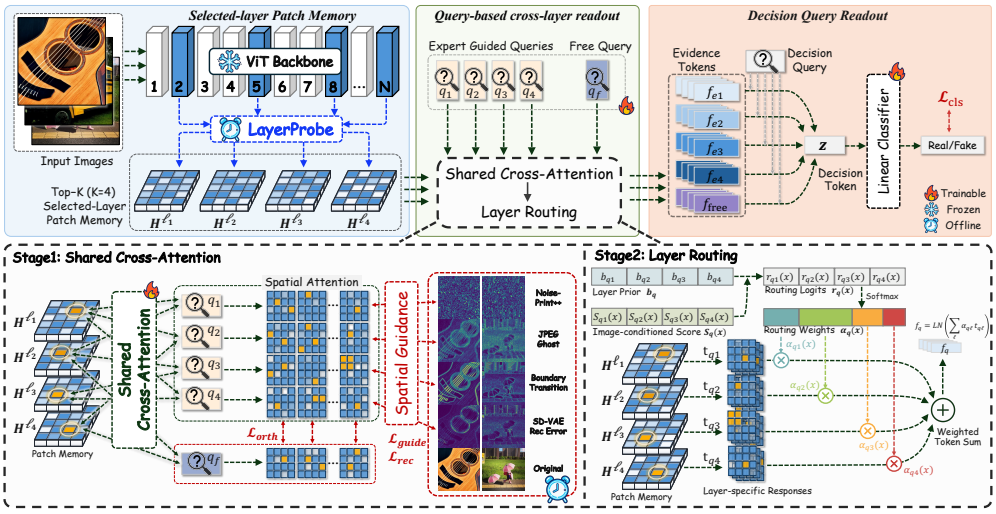

Figure 3: **Overview of the ForenLens pipeline.** The upper row shows selected-layer patch memories, cross-layer query readout, and decision readout; the lower panels expand shared cross-attention, layer routing, and expert guidance. The auxiliary losses are shown schematically: $\mathcal{L}_{\text{guide}}$ supervises layer-aggregated spatial attention, while $\mathcal{L}_{\text{orth}}$ and $\mathcal{L}_{\text{rec}}$ act on evidence tokens.

Query-adaptive layer routing. For each query, we combine a learned layer prior $b_{q\ell}$ with an image-conditioned score parameterized by its learned routing vector $\boldsymbol{u}_q$. The resulting weights fuse its layer-specific responses:

$$r_{q\ell} = b_{q\ell} + \frac{\langle \mathrm{LN}(\boldsymbol{t}_{q\ell}), \boldsymbol{u}_q \rangle}{\sqrt{d}}, \qquad \alpha_{q\ell} = \mathrm{softmax}_{\ell \in \mathcal{K}}(r_{q\ell}), \qquad \boldsymbol{f}_q = \mathrm{LN}\left(\sum_{\ell \in \mathcal{K}} \alpha_{q\ell} \boldsymbol{t}_{q\ell} \right) \tag{2}$$

The prior captures a stable query-specific layer preference, while the inner product adjusts it using the current image response. Layer identity is retained until this fusion, which yields one evidence token $\boldsymbol{f}_q$ per query. The same weights also aggregate spatial attention:

$$\bar{\boldsymbol{a}}_q = \sum_{\ell \in \mathcal{K}} \alpha_{q\ell} \boldsymbol{a}_{q\ell} \tag{3}$$

The routed attention $\bar{\boldsymbol{a}}_q$ provides spatial support for $\boldsymbol{f}_q$ and is the quantity supervised by the guidance introduced next.

4.2 Expert Guidance and Query Specialization

Classification supervision alone may lead queries in the shared reader to rely on the same cue. To encourage complementary retrieval, we partition the query bank as $\mathcal{Q} = \mathcal{E} \cup \mathcal{F}$ into expert-guided and free queries. Expert-guided queries receive spatial forensic priors during training, while free queries receive no operator-specific targets and can seek evidence beyond these priors. Both groups use the same cross-layer reader and learn their layer preferences from data.

Expert-guided queries. For expert guidance, we adopt Noiseprint++ (Guillaro et al., 2023), JPEG Ghost (Farid, 2009), a lightweight boundary-transition operator, and SD-VAE reconstruction error (Ricker et al., 2024) to provide spatial priors during training. Appendix D details their guidance-map construction.

For image i and expert e, the operator produces a pixel-space guidance map, aligned to the ViT patch grid as $\mathbf{g}_{ie} \in \mathbb{R}^P$. This map serves only as an auxiliary training target for the routed spatial attention $\bar{\boldsymbol{a}}_{ie}$ in Eq. (3), not as an input to the reader. A calibrated confidence c_{ie} and validity mask m_{ie} suppress weak or undefined guidance maps. Let $\mathcal{E}_{\text{act}}$ contain experts with at least one valid target in the current minibatch. We normalize the confidence-weighted error within each expert and then average across experts (using $\epsilon > 0$ to prevent division by zero):

$$\mathcal{L}_{\text{guide}} = \frac{1}{|\mathcal{E}_{\text{act}}|} \sum_{e \in \mathcal{E}_{\text{act}}} \frac{\sum_i m_{ie} c_{ie} \left[1 - \cos\left(\bar{\boldsymbol{a}}_{ie}, \mathbf{g}_{ie} \right) \right]}{\sum_i m_{ie} c_{ie} + \epsilon} \tag{4}$$

The loss supervises the cross-layer attention mixture, allowing spatial responses and routing weights to adapt jointly to the assigned forensic cue. Normalizing within each expert prevents differences in target availability or confidence scale from determining its contribution to the objective.

For the SD-VAE query, a scalar head on its evidence token predicts the spatial mean of the reconstruction-error map. We denote the confidence-weighted Smooth-L1 loss over valid samples by $\mathcal{L}_{\mathrm{rec}}$. This scalar target anchors the token to mean reconstruction error, complementing the spatial attention guidance.

Free queries. An unguided query may still duplicate expert evidence under classification supervision. We discourage redundancy between free-query and expert-query tokens through an orthogonality loss using the stop-gradient operator sg:

$$\mathcal{L}_{\mathrm{orth}} = \frac{1}{|\mathcal{F}||\mathcal{E}|} \sum_{f\in\mathcal{F}} \sum_{e\in\mathcal{E}} \left|\cos\left(\boldsymbol{f}_f, \mathrm{sg}[\boldsymbol{f}_e]\right)\right| \tag{5}$$

The stop-gradient blocks gradients from $\mathcal{L}_{\mathrm{orth}}$ through expert tokens, directing separation toward the free representation.

4.3 Decision readout and training objective

Decision readout. Each query contributes one evidence token $\boldsymbol{f}_q$, and their relevance may vary by image. A final decision query $\boldsymbol{q}_{\mathrm{dec}}$ attends over $\boldsymbol{F} = [\boldsymbol{f}_q]_{q\in\mathcal{Q}}$, and a linear head $\boldsymbol{W}$ maps the result to binary logits:

$$\boldsymbol{z} = \mathrm{LN}\left(\boldsymbol{q}_{\mathrm{dec}} + \mathrm{MHA}\left(\boldsymbol{q}_{\mathrm{dec}}, \boldsymbol{F}, \boldsymbol{F}\right)\right), \qquad \hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{z} \tag{6}$$

This attention operates on the evidence retrieved in Section 4.1, combining query contributions according to the current image.

Training objective. We optimize the queries, shared reader, layer router, decision readout, classifier, and auxiliary reconstruction head, while keeping the ViT frozen. We use two-class cross-entropy as $\mathcal{L}_{\mathrm{cls}}$ and combine it with the role-specific losses in Section 4.2 using non-negative weights λ_{guide}, λ_{orth}, and λ_{rec}:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{guide}}\mathcal{L}_{\mathrm{guide}} + \lambda_{\mathrm{orth}}\mathcal{L}_{\mathrm{orth}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} \tag{7}$$

At inference, all forensic guidance operators are removed, and ForenLens performs only frozen feature extraction, cross-layer query readout, and decision readout.

5 Experiments

5.1 Experimental setup

Datasets. We train ForenLens on the OpenMMSec training split and evaluate it on out-of-distribution (OOD) split (Du et al., 2026). External evaluation covers AIGCDetectBenchmark (Zhong et al., 2023), the pixel and semantic tracks of AIGI-Now (Chen et al., 2026), AIGIBench (Li et al., 2025b), Chameleon (Yan et al., 2025a), Holmes-Set (Zhou et al., 2025), and WildRF (Cavia et al., 2024).

Baselines. We compare artifact-oriented detectors, including CNNDetection (Wang et al., 2020), NPR (Tan et al., 2024), AIDE (Yan et al., 2025a), SAFE (Li et al., 2025a), and GenSign (Fei et al., 2026); pretrained-representation methods, including UnivFD (Ojha et al., 2023), Effort (Yan et al., 2025b), PGC (Zhou et al., 2026), RINE (Koutlis & Papadopoulos, 2024), SICA (Du et al., 2026), and DoU (He et al., 2026); and hybrid or reconstruction models, including Mesorch (Zhu et al., 2025), MARE (Xu et al., 2026), and DRCT (Chen et al., 2024a). Average Precision (AP) is the primary metric in the main text, and the appendix retains the complete four-metric results (ROC-AUC, AP, Accuracy, and F1).

Details of dataset composition, aggregation, backbone and selected-layer configurations, and training hyperparameters are provided in Appendix D.1.

5.2 Cross-benchmark generalization

We evaluate ForenLens without target-domain adaptation, using only in-domain OpenMMSec data for training and layer selection and a fixed classification threshold of 0.5 for Accuracy and F1.

Table 1: **Cross-benchmark Average Precision (%).** OpenMMSec equally averages four domains; external groups equally average their test sets. *Avg* equally weights all 8 groups. Best and second-best displayed values are bold and underlined; ties share the same style. † marks OpenMMSec values transcribed from SICA; italicized ‡ methods use author checkpoints without OpenMMSec retraining. MARE denotes its retrained FDM.

Method	Venue	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
CNNDet.	CVPR'20	55.9	62.9	73.9	48.6	62.9	51.6	80.6	77.0	64.2
UnivFD	CVPR'23	62.3†	86.3	67.2	81.3	82.1	70.0	85.6	82.1	77.1
NPR	CVPR'24	63.1	84.1	93.5	67.1	58.0	42.7	89.5	76.6	71.8
RINE	ECCV'24	73.3	98.3	**98.5**	52.4	57.5	39.7	99.0	72.3	73.9
Mesorch	AAAI'25	64.4†	79.2	87.7	57.0	64.1	44.3	91.4	82.3	71.3
AIDE	ICLR'25	48.6	83.4	64.1	52.1	51.2	44.8	73.6	50.2	58.5
Effort	ICML'25	73.1†	88.5	42.8	38.1	38.6	27.7	62.2	49.3	52.5
SAFE	KDD'25	54.4	83.4	95.3	48.6	51.4	52.5	91.7	54.2	66.5
SICA	ICML'26	79.0†	98.3	93.4	62.8	69.5	50.8	98.4	94.6	80.8
PGC	ICML'26	75.3	97.8	94.3	47.1	60.1	41.8	98.8	73.8	73.6
MARE	MM'26	58.5	71.1	75.9	43.7	73.1	57.0	80.5	86.0	68.2
DRCT‡	ICML'24	51.3	81.5	95.5	67.6	81.3	71.0	98.2	84.5	78.9
DoU‡	CVPR'26	58.1	**99.2**	90.6	44.1	58.4	46.5	93.9	74.2	70.6
GenSign (ProGAN20)‡	CVPR'26	53.6	98.6	70.2	52.9	61.7	45.2	80.0	71.7	66.7
GenSign (SD1.4)‡	CVPR'26	52.8	98.1	97.6	50.1	45.7	38.1	99.3	68.1	68.7
ForenLens (CLIP)	–	78.8	96.6	97.8	65.6	74.8	65.7	95.4	93.7	83.5
ForenLens (SigLIP2)	–	**81.8**	98.2	95.6	**89.7**	**93.5**	**87.9**	**99.4**	**99.2**	**93.2**

Table 2: **Ablation on query composition across backbones (AP, %).**

Noiseprint ++	JPEG Ghost	Boundary transition	SD-VAE	Free query	Mean AP CLIP	Mean AP SigLIP2
×	✓	✓	✓	✓	80.9	92.6
✓	×	✓	✓	✓	80.6	91.9
✓	✓	×	✓	✓	80.5	92.0
✓	✓	✓	×	✓	82.3	92.7
✓	✓	✓	✓	×	81.4	92.0
✓	✓	✓	✓	✓	**83.5**	**93.2**

Table 3: **Ablation on query reader, selected-layer budget, and loss controls (AP, %).**

Reader	Top-K	$\mathcal{L}_{\text{guide}}$	$\mathcal{L}_{\text{orth}}$	$\mathcal{L}_{\text{rec}}$	Mean AP CLIP	Mean AP SigLIP2
Pooled	4	–	–	–	79.5	88.9
Query	1	✓	✓	✓	79.5	81.8
Query	2	✓	✓	✓	80.6	**93.6**
Query	3	✓	✓	✓	81.0	93.4
Query	4	×	✓	✓	80.5	90.5
Query	4	✓	×	✓	81.4	92.2
Query	4	✓	✓	×	79.1	91.9
Query	4	✓	✓	✓	**83.5**	93.2

Table 1 shows that ForenLens (SigLIP2) achieves the highest mean AP of 93.2%, leads six of eight groups, and is the only configuration exceeding 80% AP on every group. Compared with SICA, the best baseline by mean AP in this table, ForenLens improves AP on AIGI-Now semantic and Chameleon with both backbones, reaching 65.6% and 65.7% with CLIP and 89.7% and 87.9% with SigLIP2. The CLIP configuration also achieves 83.5% mean AP, exceeding all baseline configurations in the table. ForenLens also ranks first across all four metrics in the comparison using the same SigLIP2 backbone (Appendix B). These results show strong aggregate performance across backbones and more balanced transfer with SigLIP2, supporting the use of complementary evidence from selected layers for forensic generalization.

5.3 Ablation Study

Impact of query composition. To determine whether the performance gain merely stems from introducing an arbitrary query bank, we analyze the impact of expert-guided and free-query composition in Table 2. For both backbones, removing any one of the four expert-guided queries or the free query degrades AP. The consistent pattern shows that expert-guided queries provide complementary, structured artifact coverage, while the free query supplies content and contextual evidence beyond their predefined targets.

Impact of query reader. To assess the effectiveness of our query-based readout framework, we compare ForenLens with the Pooled baseline in Table 3. It raises the eight-group mean AP from 79.5% to 83.5% with CLIP and from 88.9% to 93.2% with SigLIP2. The consistent gains support learning structured forensic readout from selected layers rather than reducing them to a single pooled representation.

Table 4: **Backbone ablation with a fixed four-guided, one-free query set (Average Precision, %).**

Backbone	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
Standalone vision encoders									
CLIP	78.8	96.6	97.8	65.6	74.8	65.7	95.4	93.7	83.5
DINOv2	73.7	94.1	79.7	85.6	79.4	67.6	95.0	90.8	83.2
DINOv3	76.1	94.5	78.8	86.5	86.0	74.7	95.6	95.3	86.0
MAE	59.1	77.9	63.9	71.5	58.2	43.6	80.9	76.1	66.4
SAM	64.0	84.8	97.8	81.1	57.4	42.3	96.7	85.0	76.1
SigLIP	71.2	92.9	81.9	90.4	71.1	58.2	95.7	94.1	81.9
SigLIP2	**81.8**	**98.2**	95.6	89.7	**93.5**	**87.9**	99.4	**99.2**	**93.2**
MLLM vision encoders									
Qwen2.5-VL	71.9	93.1	**99.9**	95.5	72.6	64.6	97.6	92.1	85.9
Qwen3-VL	73.2	92.4	99.8	91.7	79.3	64.9	**99.6**	97.7	87.3
Qwen3.5	70.9	93.8	**99.9**	93.0	70.8	51.8	98.8	96.2	84.4
InternVL2.5	70.5	88.7	94.2	84.8	83.5	73.1	97.0	91.8	85.4
InternVL3.5	65.2	83.2	87.8	89.3	77.9	62.6	91.9	91.0	81.1
Kimi-VL	70.6	93.5	**99.9**	**99.5**	66.2	48.3	97.6	90.0	83.2
GLM-4.6V	71.4	92.1	99.5	90.7	63.6	39.5	98.5	90.4	80.7

Impact of cross-layer complementarity. Following the *LayerProbe* selection order, increasing K from 1 to 2 with the same routing mechanism improves mean AP by 1.1% and 11.8% (absolute) for CLIP and SigLIP2, respectively. The pronounced SigLIP2 gain highlights the complementary evidence recovered by jointly reading the first two selected layers. Further additions continue to improve CLIP, whereas SigLIP2 remains close to its $K = 2$ performance. These different trends highlight backbone-dependent cross-layer complementarity, with substantial OOD gains achievable from a compact set of selected layers.

Impact of training losses. We assess the auxiliary objectives by removing each loss individually (Table 3). All three removals lower AP on both backbones. Removing $\mathcal{L}_{\text{guide}}$ reduces AP by 3.0% and 2.7% (absolute) for CLIP and SigLIP2, respectively, showing comparable benefits from spatial role guidance. The orthogonality and reconstruction objectives yield larger marginal gains on CLIP, indicating that the contribution of representation-level supervision is more backbone-dependent. This pattern suggests that stronger frozen features still benefit from explicit query-role guidance, even when the additional gains from representation-level constraints are smaller.

Ablation on backbones. To assess the impact of backbone choice on forensic generalization, we evaluate ForenLens across all 14 frozen backbones. As shown in Table 4, SigLIP2 achieves the highest mean AP of 93.2%, substantially outperforming CLIP. Among MLLM vision encoders, Qwen3-VL leads with 87.3% AP, surpassing the more recent Qwen3.5. This comparison suggests that advances in general visual-semantic modeling need not translate into better forensic detection. Notably, five of the seven MLLM vision encoders outperform all standalone vision encoders on both AIGI-P and AIGI-S, which emphasize pixel-artifact and semantic discrimination, respectively. Their advantage on the pixel track is consistent with our earlier hypothesis that MLLM vision encoders retain fine-grained forensic evidence during semantic encoding. Together with their strong semantic-track performance, this observation suggests that useful low-level and semantic information can coexist within the frozen hierarchies of MLLM vision encoders. Overall, these results reveal distinct forensic strengths across backbones, while SigLIP2 provides the strongest aggregate performance.

5.4 Mechanism and Visualization Analysis

Query-adaptive layer routing. To examine query-specific depth preferences, we analyze mean routing weights over the four *LayerProbe*-selected CLIP layers (Figure 4, left). The Noiseprint++, JPEG Ghost, and SD-VAE queries favor the earlier selected layers (7 and 9), which jointly receive 85.7–88.6% of their routing weight. The free query instead places 94.1% on the later selected layers (14 and 22), while the boundary-transition query maintains a broader cross-layer distribution. This pattern is consistent with our hypothesis that expert-guided queries emphasize earlier artifact-sensitive representations, while the free query draws complementary semantic and contextual evidence from later representations.

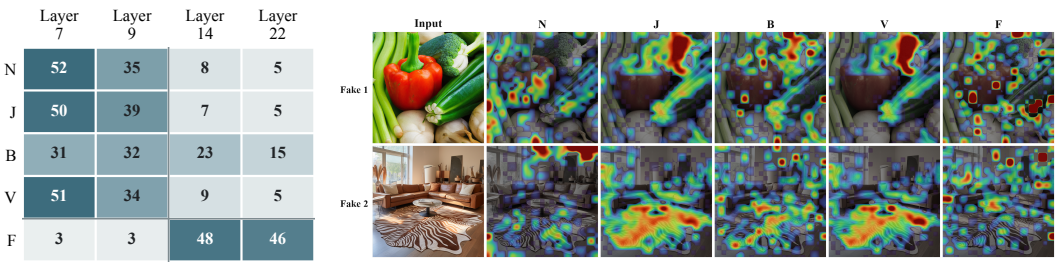

Figure 4: **Query roles specialize across layers and space.** Left: Mean routing mass (%) across Layers 7, 9, 14, and 22 from the CLIP checkpoint. Rows denote the Noiseprint++ (N), JPEG Ghost (J), boundary-transition (B), SD-VAE (V), and free (F) queries. Each row sums to 100% before rounding. Right: Depth-aggregated spatial attention on two fake images. Attention overlays use a separate display scale for each query.

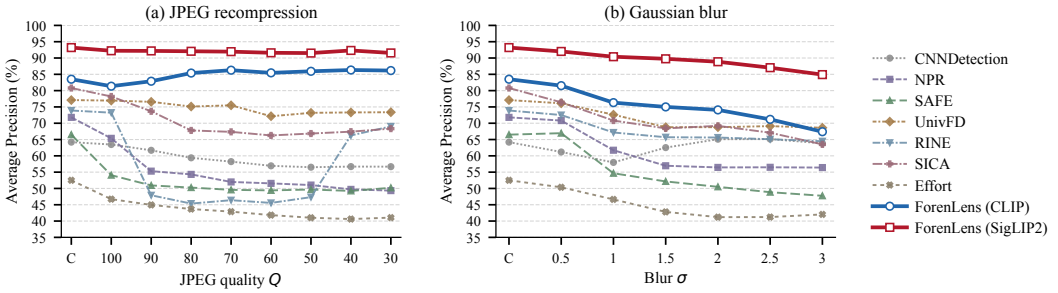

Figure 5: **Post-processing robustness (eight-group mean AP, %).** C denotes inputs without additional post-processing. Lower JPEG quality and larger blur σ indicate stronger degradation.

Query-attention visualization. We further examine spatial readout patterns through layer-aggregated query attention on two fake images (Figure 4, right). The JPEG Ghost and SD-VAE queries share broad responses over textured regions, while the remaining queries exhibit different regional emphases. These shared and distinct spatial responses complement the depth-routing analysis, providing qualitative support for differentiated query readout.

5.5 Post-processing robustness

To assess robustness to common post-processing that can weaken fine-grained forensic cues, we evaluate ForenLens under JPEG recompression and Gaussian blur across the eight evaluation groups (Figure 5). ForenLens (SigLIP2) achieves 93.2% mean AP on clean inputs and leads the compared methods under all tested post-processing settings. It achieves 91.6% AP at JPEG quality $Q = 30$ and 84.9% under Gaussian blur $\sigma = 3$ (which correspond to drops of 1.6% and 8.3% AP from clean inputs, respectively), showing stable performance under recompression and sustained effectiveness under strong blur. ForenLens (CLIP) also outperforms the compared baselines across all JPEG qualities and most blur levels, although it is more sensitive to strong blur. Together, these results support the post-processing robustness of ForenLens across both backbones.

6 Conclusion

In this paper, we introduce *LayerProbe* to analyze forensic evidence across frozen ViT layers and propose ForenLens for generalizable image forgery detection. Our controlled study shows that the best readout depth varies across backbones and that individually weaker layers can provide complementary evidence. Building on these findings, ForenLens combines query-adaptive layer routing with training-only expert guidance to retrieve evidence from a compact layer set. Expert-guided queries learn spatial roles, while free queries seek evidence beyond the prescribed forensic priors. Ablations and routing analyses support cross-layer complementarity and query specialization, while further experiments show effectiveness across ViT backbones and robustness to common post-processing. These results highlight the value of retrieving complementary evidence across layers for generalizable image forensics.

Reproducibility Statement

To support reproducibility, Section 4 specifies the ForenLens architecture and complete training objective. Appendix A reports the per-backbone *LayerProbe* results and greedy layer selections. Appendix B compares detectors on a common SigLIP2 backbone. Appendix D documents the dataset splits, backbone and selected-layer configurations, data augmentation, optimization settings, random seed, and the construction and calibration of all guidance operators. Appendices E–G provide complete evaluation metrics, post-processing settings, controlled ablations, and routing analyses. Code, configuration files, and evaluation scripts will be released upon acceptance.

AI use statement

We used large language model-based tools to assist with sentence rephrasing, grammar checking, and improving the clarity and readability of the manuscript. We also used these tools to assist with writing and reviewing parts of the code used to implement our method. We have reviewed and revised the suggested manuscript edits and take full responsibility for the final manuscript and implementation.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639, 2024a.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Align, don't blend: Orthogonal supervision disentanglement as the key to generalizable AI-generated image detection. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026. URL https://arxiv.org/abs/2512.06746.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b. URL https://arxiv.org/abs/2412.05271.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025.

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 Workshops*, pp. 4356–4366, 2024.

Bo Du, Xiaochen Ma, Xuekang Zhu, Zhe Yang, Chaoqun Niu, Chenfan Qu, Mingqi Fang, Zhenming Wang, Jingjing Liu, Jian Liu, and Ji-Zhe Zhou. Can we build a monolithic model for fake image detection? SICA: Semantic-induced constrained adaptation for unified-yet-discriminative artifact feature space reconstruction. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06676. To appear.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Hany Farid. Exposing digital forgeries from JPEG ghosts. *IEEE Transactions on Information Forensics and Security*, 4(1):154–160, 2009. doi: 10.1109/TIFS.2008.2012215.

Jianwei Fei, Yunshu Dai, Xiaoyu Zhou, Zhihua Xia, and Alessandro Piva. Enabling supervised learning of generative signatures for generalized AI-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14094, 2026.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006v6.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20606–20615, 2023.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18685–18694, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Qinghui He, Haifeng Zhang, Qiao Qin, Bo Liu, Xiuli Bi, and Bin Xiao. Diversity over uniformity: Rethinking representation in generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40407–40417, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Kimi Team. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, volume 15130 of *Lecture Notes in Computer Science*, pp. 394–411, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5074–5083, 2020.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 2405–2414, 2025a.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, and Shouhong Ding. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL `https://arxiv.org/abs/2606.15880`. To appear.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, 2026.

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, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Changshuo Wang, Jiangming Wang, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, Shunli Wang, Ran Yi, and Lizhuang Ma. Beyond [CLS] token: Query-driven token-level forgery purification for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42922–42931, 2026a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Xinyuan Wang, Yingxin Lai, Zhiming Luo, and Zhihui Liu. PPM-CLIP: Probabilistic prompt modeling for generalizable AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21316–21325, 2026b.

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*, pp. 22445–22455, 2023.

Wenbo Xu, Wei Lu, and Xiangyang Luo. MARE: Multimodal alignment and reinforcement for explainable deepfake detection via vision-language models. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026.

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, 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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288, 2025b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient AI-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. URL https://arxiv.org/abs/2311.12397.

Xiaoyu Zhou, Jianwei Fei, Peipeng Yu, Jingchang Xie, Chong Cheng, and Zhihua Xia. PGC: Peak-guided calibration for generalizable AI-generated image detection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.21207. To appear.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11022–11030, 2025.

A Additional Layer-Probe Results

A.1 Per-backbone absolute probe AUC

Figure 6 reports absolute *LayerProbe* AUC versus one-based block index for the same 14 backbones as Figure 1. Each panel is one backbone; the star annotation locates its Top-1 layer and reports its AUC. These curves complement the peak-normalized trajectories in the main text.

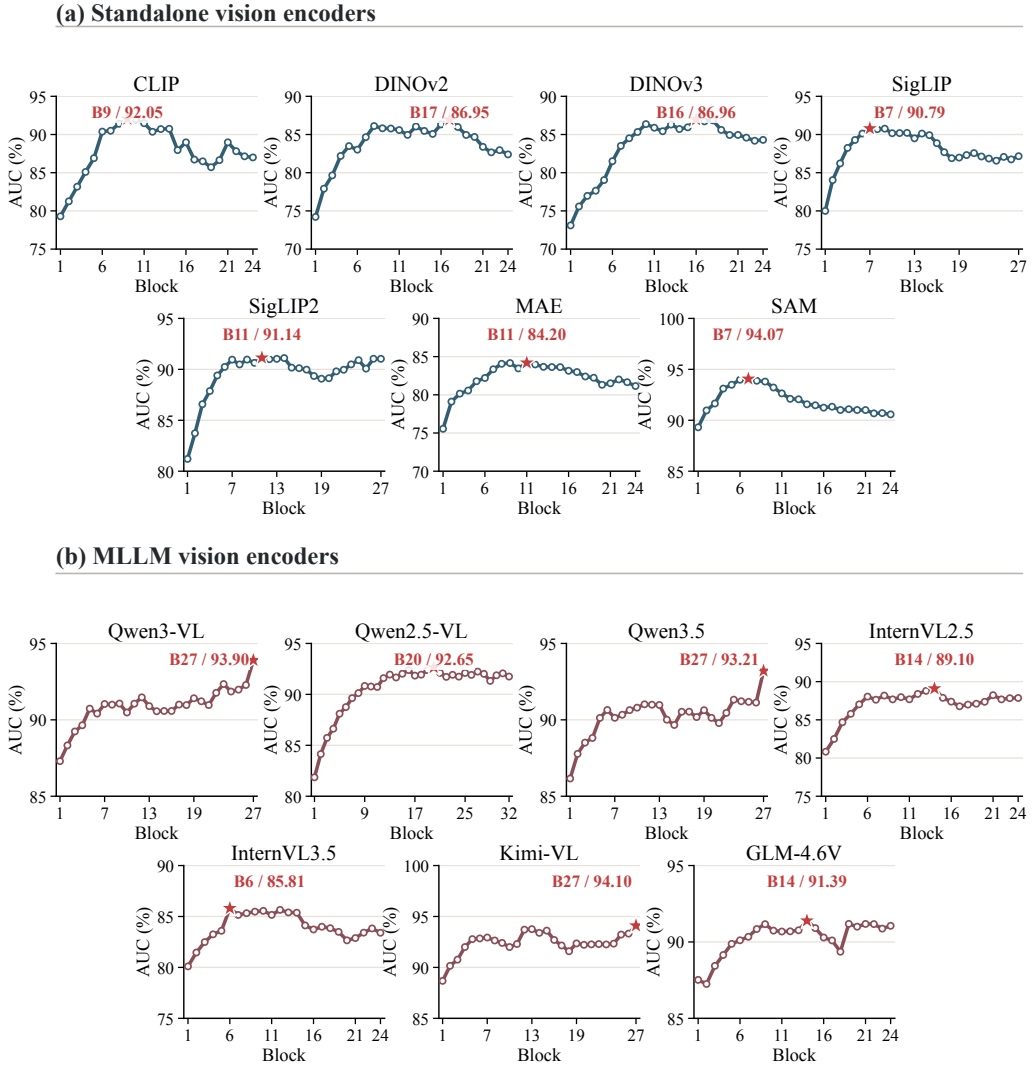

Figure 6: **Absolute single-layer probe AUC for each backbone.** Standalone and MLLM vision encoders form separate panel groups. Each panel plots ID validation AUC (%) against one-based block index and uses a panel-specific vertical range to expose within-backbone variation. The star annotation reports the Top-1 block and its AUC.

Across backbones, the absolute curves range from broad high-AUC plateaus to sharply localized maxima. Their distinct shapes reinforce that forensic readability is organized by the individual visual hierarchy rather than by a single normalized depth shared across architectures.

A.2 Selected complementary depths

Table 5 lists the greedily selected blocks ℓ_1–ℓ_4 and the corresponding cumulative validation AUCs A_1–A_4 for every backbone in Figure 2.

Table 5: **Greedy *LayerProbe* selections and cumulative ID validation AUC (%).** Indices ℓ_1–ℓ_4 are one-based blocks from ID-only greedy forward selection; A_k is the validation AUC after fusing $\{\ell_1, \ldots, \ell_k\}$, and $\Delta A_k = A_k - A_1$.

Backbone	Selected blocks				Cumulative ID AUC				Gain over A_1		
	ℓ_1	ℓ_2	ℓ_3	ℓ_4	A_1	A_2	A_3	A_4	ΔA_2	ΔA_3	ΔA_4
Standalone vision encoders											
CLIP	9	22	14	7	92.05	95.14	96.16	96.64	+3.10	+4.12	+4.59
DINOv2	17	9	6	12	86.95	90.79	91.85	92.64	+3.84	+4.89	+5.69
DINOv3	16	9	21	7	86.96	90.74	92.21	93.10	+3.78	+5.25	+6.14
SigLIP	7	15	12	8	90.79	92.79	93.42	93.77	+2.00	+2.63	+2.98
SigLIP2	11	26	7	14	91.14	95.27	96.46	96.91	+4.14	+5.33	+5.77
MAE	11	7	23	17	84.20	85.69	87.52	88.38	+1.49	+3.32	+4.18
SAM	7	10	5	13	94.07	94.80	95.27	95.49	+0.74	+1.20	+1.43
MLLM vision encoders											
Qwen3-VL	27	6	12	23	93.90	94.66	95.78	96.34	+0.75	+1.87	+2.43
Qwen2.5-VL	20	32	12	7	92.65	94.78	95.80	96.14	+2.13	+3.15	+3.49
Qwen3.5	27	5	12	23	93.21	94.19	95.30	95.79	+0.99	+2.10	+2.58
InternVL2.5	14	23	10	6	89.10	91.82	93.54	94.25	+2.72	+4.45	+5.15
InternVL3.5	6	23	9	13	85.81	88.40	90.50	91.45	+2.60	+4.69	+5.64
Kimi-VL	27	12	7	14	94.10	95.51	95.98	96.11	+1.41	+1.88	+2.01
GLM-4.6V	14	22	9	5	91.39	94.07	95.21	95.66	+2.68	+3.82	+4.27

Every backbone benefits from evidence beyond its strongest single block, and the selected additions frequently span non-adjacent representation stages. This consistent gain supports retrieving complementary evidence from multiple layers of the frozen hierarchy.

A.3 Complete conditional scans along the greedy path

Figure 7 exposes every candidate layer evaluated at Steps 2–4 of the ID-only forward selection; the complete Step-1 scan is already reported in Figure 6. At Step k, the previously selected prefix is fixed, each remaining block is appended in turn, and a linear probe with the same configuration is refit. Consequently, for a given backbone and step, every candidate has the same layer budget, descriptor width, and number of trainable probe coefficients.

The same-budget candidates are not interchangeable. Across the 14 backbones, the mean max–min spread among candidate AUC gains is 2.18%, 1.15%, and 0.67% (absolute) at Steps 2, 3, and 4, respectively. Because model width and parameter count are fixed within each scan, these differences arise from the representation being added and its conditional complementarity with the current prefix, rather than from merely enlarging the probe. Together with the rank–gain mismatch in Figure 2, the complete scans identify layer content—not layer count alone—as the source of the greedy gains.

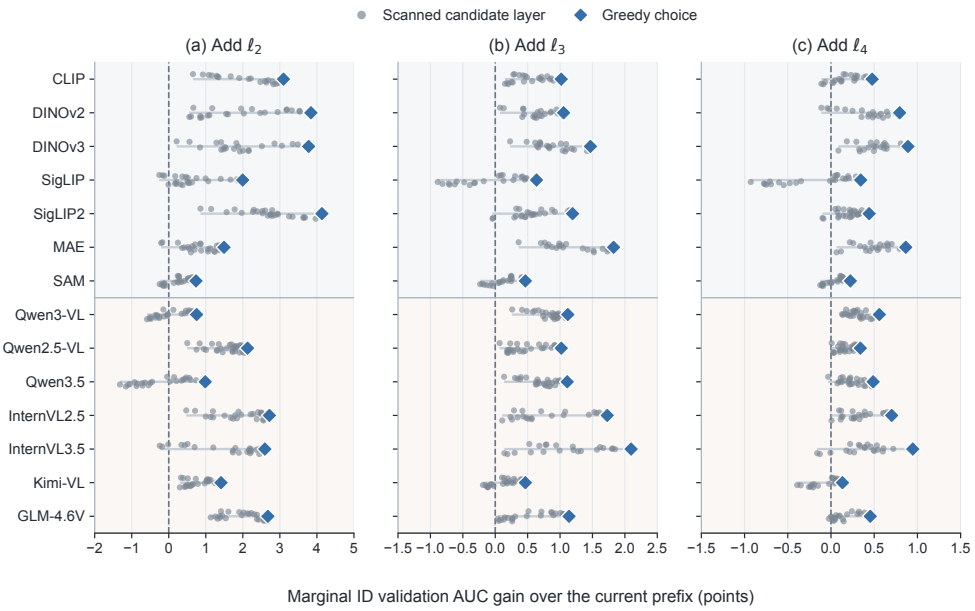

Figure 7: **Complete ID-only greedy scans at fixed layer budgets.** Panels show marginal validation-AUC gains from adding the second, third, and fourth layer to the selected prefix. Light points denote remaining candidate layers; diamonds mark the greedy choices. Upper and lower groups correspond to standalone and MLLM vision encoders. Slight vertical offsets separate overlapping points.

A.4 OOD TRANSFER AND SAME-WIDTH CONTROLS

We next fix the four-step layer sequence selected exclusively on ID data. Each prefix and control probe is fitted on the same ID training set and then evaluated directly on a fixed class-balanced subset of the OOD split, which contains 1,000 real and 1,000 fake images per primary type (28,000 in total) and is shared by all backbones. No OOD label enters layer selection, fitting, or hyperparameter tuning. Table 6 reports AUC equally averaged over the 14 OpenMMSec primary types.

The complete ID-selected set S_4 improves over S_1 for 13 of 14 backbones by 4.65% (absolute) in AUC on average, and it is the strongest of the four fixed prefixes for 13 backbones. Thus, the complementary set identified in-domain transfers broadly to unseen manipulations rather than serving only the ID selection split.

Repeat-ℓ_1 matches the descriptor width and trainable linear coefficients of S_4 without introducing information from another layer. S_4 outperforms this control for all 14 backbones by 5.32% (absolute) on average. The gain therefore cannot be explained by a wider probe alone. Rank-4 also improves over Repeat-ℓ_1 by 4.25% and over S_1 by 3.57% (absolute) on average, confirming that distinct layers carry transferable evidence. Finally, S_4 exceeds Rank-4 for 11 of 14 backbones and by 1.08% (absolute) on average, showing that conditional selection adds value beyond single-layer strength.

Table 6: **Fixed OOD transfer of ID-selected *LayerProbe* sets and same-width controls (macro AUC, %).** S_k uses the first k layers of the ID-only greedy sequence. Repeat-ℓ_1 concatenates four copies of the best ID layer, whereas Rank-4 concatenates the four layers with the highest single-layer ID AUC. Within each backbone, all $K = 4$ entries have the same descriptor width and number of linear-probe coefficients. We define $\Delta_1 = S_4 - S_1$, $\Delta_{\text{rep}} = S_4 - \text{Repeat-}\ell_1$, and $\Delta_{\text{rank}} = S_4 - \text{Rank-4}$.

Backbone	ID-selected prefix on OOD				$K = 4$ controls		Paired difference		
	S_1	S_2	S_3	S_4	Repeat-ℓ_1	Rank-4	Δ_1	Δ_{rep}	Δ_{rank}
Standalone vision encoders									
CLIP	76.71	83.33	85.78	86.07	76.40	81.19	+9.36	+9.67	+4.87
DINOv2	78.35	80.17	80.60	81.51	77.51	82.60	+3.16	+4.00	-1.09
DINOv3	78.23	80.89	83.23	83.95	77.31	83.26	+5.72	+6.64	+0.69
SigLIP	75.16	80.82	81.98	82.10	74.29	80.13	+6.94	+7.81	+1.97
SigLIP2	79.00	86.92	87.96	88.41	78.47	88.11	+9.41	+9.94	+0.30
MAE	70.37	70.93	73.59	74.09	69.80	72.37	+3.72	+4.29	+1.72
SAM	76.31	76.58	77.22	77.37	75.57	76.62	+1.06	+1.80	+0.75
MLLM vision encoders									
Qwen3-VL	82.11	80.36	81.08	83.16	81.25	83.24	+1.05	+1.92	-0.08
Qwen2.5-VL	78.37	81.83	82.62	82.82	77.90	81.04	+4.45	+4.92	+1.78
Qwen3.5	82.80	80.09	81.49	82.70	81.94	81.53	-0.10	+0.76	+1.17
InternVL2.5	74.88	78.78	80.13	80.83	74.05	79.95	+5.95	+6.78	+0.88
InternVL3.5	69.82	73.61	74.47	75.99	69.20	73.98	+6.17	+6.79	+2.01
Kimi-VL	82.00	83.89	84.44	84.70	81.37	84.41	+2.70	+3.33	+0.29
GLM-4.6V	75.91	80.39	81.30	81.42	75.53	81.57	+5.51	+5.89	-0.16
Mean	**77.14**	**79.90**	**81.13**	**81.79**	**76.47**	**80.71**	**+4.65**	**+5.32**	**+1.08**

B Comparison with SigLIP2-based baselines

To compare detection designs on the same pretrained backbone, we adapt UnivFD, RINE, Effort, SICA, and PGC to SigLIP2-SO400M/14 and train each on the OpenMMSec training split. Table 7 reports their eight-group means across four metrics.

The SigLIP2 baselines use the same pretrained SO400M/14 checkpoint and 384-pixel input as ForenLens. Baselines using global representations apply multihead attention pooling (MAP), while RINE uses per-layer patch means. UnivFD and RINE freeze the backbone; SICA and PGC train LoRA adapters, and Effort trains SVD residuals. Each baseline follows a method-specific training schedule and selects its checkpoint by in-domain validation loss.

Table 7: **SigLIP2 comparison (eight-group means, %).** Best and second-best values are bold and underlined.

Method	ROC-AUC	Accuracy	AP	F1
UnivFD	87.2	77.6	87.1	76.2
RINE	86.9	81.1	87.1	73.7
Effort	<u>89.6</u>	<u>85.5</u>	<u>89.9</u>	<u>81.8</u>
SICA	88.8	84.3	88.6	79.7
PGC	88.7	84.0	87.1	81.3
ForenLens	**93.3**	**87.1**	**93.2**	**84.4**

ForenLens ranks first across all four metrics, achieving 93.2% AP and 87.1% Accuracy. Its AP exceeds Effort, the strongest baseline in this comparison, by 3.3% (absolute). This advantage over both frozen-feature readouts and backbone-adaptation methods supports the effectiveness of learned cross-layer evidence retrieval on a common pretrained backbone.

C LIMITATIONS

ForenLens provides a modular interface for assigning forensic roles to queries. This work instantiates it with a compact portfolio of four training-time experts, whose utility and complementarity are validated through controlled ablations. The broader space of forensic expertise remains open: other artifact operators may provide additional or overlapping cues, and the appropriate expert set and query budget may vary by application. Extending ForenLens with a larger candidate bank and ID-only, budget- and redundancy-aware expert selection is a natural direction.

Expert utility also depends on the frozen backbone. Spatial role guidance benefits both CLIP and SigLIP2, while the orthogonality and reconstruction objectives contribute larger marginal gains on CLIP. We retain one shared expert portfolio across backbone-specific training runs to test the design's generality; backbone-aware expert allocation could better match the guidance to each backbone. Finally, our evaluation focuses on binary image-level detection of still images. Extending cross-layer readout to dense localization, source attribution, and video forensics remains future work.

D METHOD AND TRAINING DETAILS

D.1 TRAINING AND EVALUATION PROTOCOL

We fit and select baselines on the official OpenMMSec training and validation splits, which contain 26 training forgery types across Deepfake, AIGC, image manipulation (IMDL), and document forensics (Du et al., 2026). Evaluation uses the complete, unsubsampled OpenMMSec OOD split: 72 unseen forgery types grouped into 14 primary types over the same four domains. The 7 external benchmarks comprise AIGCDetectBenchmark (Zhong et al., 2023), the pixel- and semantic-level tracks of AIGI-Now (Chen et al., 2026), AIGIBench (Li et al., 2025b), Chameleon (Yan et al., 2025a), Holmes-Set (Zhou et al., 2025), and WildRF (Cavia et al., 2024), for 51 constituent external test sets.

For AIGIBench, we use only its two official platform-sourced subsets to assess transfer to in-the-wild online imagery: CommunityAI from AI-art communities (6,000 real/6,000 fake) and SocialRF from social networks (3,000 real/3,000 fake). The remaining 23 generator-specific subsets are not included. Both are evaluated in full and equally weighted; this source-defined scope is fixed independently of detector performance.

For Holmes, the reported group corresponds specifically to Protocol-III (P3) of AIGI-Holmes, its latest-generator generalization suite. We evaluate all ten P3 subsets—Janus, Janus-Pro-1B, Janus-Pro-7B, Show-o, LlamaGen, Infinity, VAR, PixArt-XL, SD3.5-Large, and FLUX—and weight them equally. Protocol-I reuses AIGCDetectBenchmark, which is reported separately, while Protocol-II follows the AntiFakePrompt setting; neither is included in the Holmes group.

Accuracy and F1 use a fixed fake-probability threshold of 0.5; ROC-AUC and AP use continuous probabilities. For OpenMMSec, metrics are averaged over primary types and then equally over the four domains. External test sets are equally weighted within each benchmark, and the reported eight-group average equally weights OpenMMSec and the 7 external benchmarks. *LayerProbe* selects candidate layers exclusively on the official in-domain validation split. Neither OpenMMSec OOD labels nor labels from the 7 external benchmarks enter layer selection or training. The backbone study reports all 14 frozen backbones under this same protocol rather than retuning a configuration on external results.

For CNNDetection, p denotes the data-augmentation probability during training. We retrain MARE's forgery-disentanglement module (FDM) on OpenMMSec. DoU and GenSign retain their official preprocessing, with the ProGAN20 and GenImage-SD1.4 GenSign checkpoints reported separately.

The SigLIP2 instantiation freezes the backbone at its standard 384-pixel input, reads LayerProbe-selected Layers 7, 11, 14, and 26, and uses a 1152-dimensional reader. The CLIP instantiation freezes OpenAI CLIP ViT-L/14@336 with its official 336-pixel crop, reads Layers 7, 9, 14, and 22, and uses a 1024-dimensional reader. Both use an effective global batch of 64.

Training augmentation includes horizontal flips, JPEG recompression, Gaussian blur, down/up rescaling, and additive Gaussian noise, with the augmentation probability set to 0.5. We optimize for 10 epochs with AdamW, learning rate and weight decay 10^{-4}, BF16 arithmetic, and gradient clipping

at 1.0. ForenLens results use the final model after 10 epochs of training with random seed 42. Loss weights are $\lambda_{\text{guide}} = 0.2$, $\lambda_{\text{orth}} = 0.05$, and $\lambda_{\text{rec}} = 0.05$.

D.2 Forensic guidance construction

The four operators instantiate distinct forensic mechanisms while sharing a common interface: a non-negative spatial map and a calibrated confidence.

Noiseprint++ consistency. We use the pretrained Noiseprint++ extractor distributed with TruFor (Guillaro et al., 2023). A 17-pixel reflection pad is applied before the extractor so that its DnCNN zero-padding frame lies outside the returned map. Given the signed residual R, we first subtract a Gaussian local center, then compare local and regional residual energies:

$$E_{\text{loc}} = G_{2.0} * \big(R - G_{1.5} * R\big)^2, \qquad E_{\text{reg}} = G_{12.0} * E_{\text{loc}}, \qquad g_N = \left|\log \frac{E_{\text{loc}} + \epsilon}{E_{\text{reg}} + \epsilon}\right|, \tag{8}$$

where G_σ denotes Gaussian filtering and $\epsilon = 10^{-6}$. The resulting map emphasizes spatial departures from locally stationary camera-model residual statistics.

JPEG Ghost consistency. Following the recompression principle of JPEG ghosts (Farid, 2009), we encode and decode the input at qualities $\mathcal{Q}_{\text{JPEG}} = \{55, 65, 75, 85, 95\}$. For each quality q, the residual map is the channel-mean absolute difference D_q between the input and its recompressed image. We aggregate disagreement across qualities as

$$g_J = \text{Std}_{q \in \mathcal{Q}_{\text{JPEG}}}(D_q) + 0.25\,\text{Mean}_{q \in \mathcal{Q}_{\text{JPEG}}}(D_q), \tag{9}$$

which responds to regions whose compression history is inconsistent with the surrounding image.

Boundary transition. This deterministic operator combines color and texture transitions without an external model. Let L_c denote the three CIELAB channels and Y the grayscale image. We compute

$$g_B = \frac{1}{3}\sum_{c=1}^{3} \|\text{Sobel}(L_c)\|_2 + \|\text{Sobel}(G_{1.0} * |Y - G_{2.0} * Y|)\|_2\,, \tag{10}$$

so a strong response requires an abrupt color or local-texture transition of the kind commonly introduced by compositing and regional editing.

SD-VAE reconstruction. Motivated by reconstruction-based diffusion detection such as AEROBLADE (Ricker et al., 2024), we use a frozen `stabilityai/sd-vae-ft-mse` autoencoder. The input is resized to dimensions divisible by eight and normalized to $[-1, 1]$; the VAE posterior mean, rather than a random sample, is decoded for deterministic reconstruction. The map g_V is the channel-mean absolute reconstruction error, resized back to the input geometry. It exposes regions that are poorly represented by the frozen generative autoencoder.

Common map processing. Each non-negative map is area-pooled to the ViT patch grid and flattened. For expert e, we summarize its strength by the mean of its top-5% patch responses and linearly calibrate that scalar to $[0, 1]$ using the expert's training-set 5th and 95th percentiles. A target is valid only when its calibrated strength is at least 0.05 and the map has a nonzero response. Noiseprint++, JPEG Ghost, and SD-VAE maps are cached offline, while the boundary map is recomputed online from the augmented image. Horizontal flips are applied to cached maps; rescaling, blur, additive noise, or JPEG augmentation invalidates a cached target because it no longer corresponds to the augmented input. The spatial mean of g_V additionally supplies the reconstruction auxiliary target. We set $\mathcal{L}_{\text{guide}}$ to zero if no guidance target in the minibatch is valid, and $\mathcal{L}_{\text{rec}}$ to zero if no SD-VAE target is valid. These maps supervise attention during training only and are absent at inference.

Query configuration. Both reported configurations use an eight-head shared reader, four guided queries (Noiseprint++, JPEG Ghost, boundary, and SD-VAE), and one free query. Each query reads all four *LayerProbe*-selected blocks through query-adaptive layer routing; the resulting five evidence tokens are combined by the attention-based decision readout in Section 4. Query dropout is zero and the reader uses no query residual connection.

D.3 COMPUTATIONAL EFFICIENCY AND GUIDANCE PREPROCESSING

All efficiency measurements are conducted on the same device using BF16 execution, with methods grouped by backbone and input resolution. The five SigLIP2 adapted baselines share the SO400M/14 backbone and 384-pixel input used by ForenLens. For detectors built on pretrained backbones, we count batch-1 eager-execution logical FLOPs with one MAC equal to two FLOPs. F_B denotes the pure visual-backbone forward matched in implementation, pretrained weights, visual input, and output boundary, excluding detector-specific additions; F_{inf} denotes the complete detector forward. We report $\Delta F_{\text{inf}} = F_{\text{inf}} - F_B$ and $\rho_{\text{inf}} = \Delta F_{\text{inf}}/F_B$. Training compute $F_{\text{train}} = F_{\text{inf}} + F_{\text{bwd}}$, where F_{bwd} denotes backward-pass FLOPs, excludes optimizer updates and offline guidance preprocessing, with $\rho_{\text{train}} = (F_{\text{train}} - F_B)/F_B$. GenSign takes 256-pixel inputs: F_B covers its 224-pixel CLIP branch, whereas F_{inf} includes all branches. End-to-end CNNs serve as absolute-compute references.

Table 8: **Computation and backbone-gradient paths.** FLOPs use BF16 eager execution, except `fvcore` counts for CNNs (SAFE: partial graph). DoU and GenSign (ProGAN20) use author-released weights. Dashes denote inapplicable or unavailable values; ✓ indicates backbone backpropagation.

Method	Backbone	Input (px)	F_B (GFLOPs)	F_{inf} (GFLOPs)	ΔF_{inf} (GFLOPs)	ρ_{inf} (%)	F_{bwd} (GFLOPs)	ρ_{train} (%)	Backbone Backward
UnivFD	CLIP ViT-L/14	224	162.03	162.03	< 0.01	< 0.01	< 0.01	< 0.01	×
RINE	CLIP ViT-L/14	224	162.03	162.13	0.11	0.07	0.17	0.17	×
SICA	CLIP ViT-L/14	224	162.03	163.64	1.61	0.99	224.80	139.74	✓
Effort	CLIP ViT-L/14	224	162.02	162.23	0.20	0.12	221.98	137.13	✓
DoU	CLIP ViT-L/14	224	162.03	188.73	26.70	16.48	–	–	–
ForenLens	CLIP ViT-L/14	336	381.92	391.74	9.82	2.57	9.98	5.18	×
UnivFD	SigLIP2 SO400M	384	670.35	670.35	< 0.01	< 0.01	< 0.01	< 0.01	×
RINE	SigLIP2 SO400M	384	670.35	670.48	0.13	0.02	0.19	0.05	×
SICA	SigLIP2 SO400M	384	670.35	672.64	2.29	0.34	976.29	145.98	✓
Effort	SigLIP2 SO400M	384	670.35	670.63	0.29	0.04	972.28	145.08	✓
PGC	SigLIP2 SO400M	384	670.35	679.53	9.19	1.37	780.94	117.87	✓
ForenLens	SigLIP2 SO400M	384	670.35	686.03	15.68	2.34	15.89	4.71	×
Other backbones and hybrid architectures									
PGC	DINOv2-L	224	162.02	164.00	1.98	1.22	170.38	106.38	✓
GenSign	CLIP + CNN	256	162.02	240.79	78.77	48.61	–	–	–
CNNDet.	ResNet-50	224	–	8.22	–	–	–	–	✓
NPR	CNN	224	–	3.47	–	–	–	–	✓
SAFE	CNN	256	–	4.49	–	–	–	–	✓

The cross-layer readout measured in Table 8 obtains all selected layers from a single backbone forward. It adds only 2.57% and 2.34% to the CLIP and SigLIP2 backbone FLOPs, respectively. The matched SigLIP2 comparison further shows the training benefit of keeping the backbone outside the backward graph: ForenLens requires 701.92 GFLOPs for a training forward and backward, compared with 1460.47–1648.93 GFLOPs for PGC, Effort, and SICA. UnivFD and RINE retain lower compute, while ForenLens achieves stronger detection in the same-backbone comparison (Table 7). Together, these results show that ForenLens improves detection with a modest readout increment while avoiding costly backpropagation through the backbone.

Device-side runtime excludes loading and preprocessing: batch-1 latency averages 600 forwards after 50 warmups, batch-16 throughput averages 150 calls, and peak memory is batch-1 allocated memory. For the SigLIP2 comparison, ForenLens stores frozen weights in BF16, whereas the adapted baselines retain FP32 weights under BF16 autocast; memory measurements reflect these implementations.

Table 9: **BF16 runtime of measured image detectors.** Updated parameters follow each training recipe. DoU and GenSign (ProGAN20) use author-released checkpoints.

Method	Backbone	Input (px)	Updated params (M)	B1 latency (ms)	B16 throughput (images/s)	Peak memory (GiB)
UnivFD	CLIP ViT-L/14	224	< 0.01	7.52	505.6	1.65
RINE	CLIP ViT-L/14	224	6.32	7.96	485.6	0.97
SICA	CLIP ViT-L/14	224	1.18	10.18	479.9	1.66
Effort	CLIP ViT-L/14	224	0.20	14.69	489.5	1.65
DoU	CLIP ViT-L/14	224	–	15.31	438.7	1.13
ForenLens	CLIP ViT-L/14	336	8.42	9.19	240.5	0.67
UnivFD	SigLIP2 SO400M	384	< 0.01	11.95	137.8	1.66
RINE	SigLIP2 SO400M	384	6.46	12.26	137.2	1.77
SICA	SigLIP2 SO400M	384	1.99	15.79	135.0	1.67
Effort	SigLIP2 SO400M	384	0.25	18.55	133.3	3.26
PGC	SigLIP2 SO400M	384	4.72	29.36	115.3	1.69
ForenLens	SigLIP2 SO400M	384	10.65	13.15	129.7	0.93
Other backbones and hybrid architectures						
PGC	DINOv2-L	224	2.34	15.45	449.8	1.19
GenSign	CLIP + CNN	256	–	17.27	198.5	1.26
CNNDet.	ResNet-50	224	23.51	4.34	3460.3	0.13
NPR	CNN	224	1.44	2.01	7576.5	0.04
SAFE	CNN	256	1.44	2.50	5073.4	0.04

Table 9 reports execution speed. With the same SigLIP2 backbone, ForenLens runs at 13.15 ms per image, below SICA, Effort, and PGC, while UnivFD and RINE are slightly faster. Its batch-16 throughput is 129.7 images/s, compared with 115.3–137.8 images/s across the matched baselines. At its 336-pixel input, ForenLens (CLIP) reaches 9.19 ms per image and 240.5 images/s at batch 16.

The three expensive training guidance maps—Noiseprint++, JPEG Ghost, and SD-VAE—are generated in one offline pass and reused across all epochs; their joint pipeline processes 11.1 images/s at the 336-pixel CLIP geometry. Boundary transition is recomputed online at low cost because it depends on the augmented input, and all guidance operators are removed completely at inference.

E Complete Evaluation Metrics

E.1 OpenMMSec

We report ROC-AUC, Accuracy, AP, and F1 across all 14 OpenMMSec primary types and four domain aggregates, including additional baselines reported by SICA. The complete AP comparison is given in Table 12. Accuracy and F1 use a fixed threshold of 0.5; AP uses continuous scores. Across Tables 10–13, bold and underlined entries denote the best and second-best displayed values, with ties sharing the same style. † marks values transcribed from SICA, while italicized ‡ methods use author checkpoints without OpenMMSec retraining.

Table 10: **Complete OpenMMSec ROC-AUC (%).**

Method	Deepfake					AIGC							IMDL			Document			Avg
	EFS	FE	FR	FS	Avg	GAN	Lat.	Pix.	AR	Comm.	Other	Avg	Gen.	Non-Gen.	Avg	AIGC	Non-AIGC	Avg	
ResNet†	78.0	58.6	75.3	68.4	70.1	86.0	70.6	71.0	97.7	70.8	79.3	79.2	88.9	75.4	82.2	53.6	86.5	70.1	75.4
EfficientNet†	58.7	52.8	59.2	59.2	57.5	73.7	63.2	63.1	84.3	61.8	67.7	69.0	52.5	50.6	51.2	46.7	72.5	59.6	59.4
CapsuleNet†	67.7	52.4	72.9	62.2	63.8	82.2	65.3	65.4	90.1	61.9	78.8	74.0	85.7	71.3	78.5	39.6	86.1	62.9	69.8
SegFormer†	89.5	80.5	89.8	81.9	85.4	91.8	88.0	84.6	99.8	91.8	77.1	88.9	89.7	78.5	84.1	59.5	87.9	73.7	83.0
Swin†	91.6	85.2	86.2	77.3	85.1	93.6	91.5	85.7	99.9	93.6	80.1	90.7	92.3	80.7	86.5	61.4	85.6	73.5	84.0
ConvNeXt†	84.7	84.4	90.0	80.3	84.9	92.5	87.0	84.3	99.9	92.2	75.0	88.5	89.5	75.3	82.4	59.5	84.3	71.9	81.9
Recce†	68.9	54.1	51.0	55.4	57.4	83.2	71.2	70.2	94.2	67.9	74.4	76.9	59.6	55.2	57.4	62.1	66.5	64.3	64.0
Sia†	67.4	53.7	81.0	68.1	67.6	86.3	74.4	74.9	96.7	77.7	81.9	82.0	79.3	55.9	67.6	58.8	89.1	74.0	72.8
IML-ViT†	76.0	78.8	85.0	71.1	77.7	87.7	77.3	72.2	99.9	82.3	81.8	83.5	88.8	75.6	82.2	52.7	87.7	70.2	78.4
TruFor†	78.6	65.1	83.1	69.8	74.2	88.0	84.8	77.1	99.2	88.8	81.0	86.5	90.7	78.2	84.5	51.8	**90.5**	71.2	79.1
FFDN†	76.2	51.5	87.2	76.3	72.8	95.9	92.9	94.9	**100.0**	93.3	74.8	92.0	93.3	80.2	86.8	56.1	84.7	70.4	80.5
CLIP-FFT†	81.8	66.9	85.7	74.9	77.3	95.4	90.4	88.8	**100.0**	93.1	76.8	90.8	87.8	70.8	79.3	56.9	89.2	73.1	80.1
UnivFD†	88.5	94.2	69.8	73.7	81.6	91.5	69.7	78.3	98.8	66.7	87.1	82.0	68.2	84.6	76.4	33.0	79.2	56.1	74.0
Effort†	94.7	**96.4**	87.3	91.3	**92.4**	95.9	79.3	88.2	99.3	76.8	88.7	88.0	89.0	**91.0**	90.0	48.3	83.6	66.0	84.1
Mesorch†	83.0	72.7	87.1	78.8	80.4	88.2	80.1	78.3	99.6	87.5	82.5	86.0	87.6	76.0	81.8	58.9	88.5	73.7	80.5
CO-SPY†	80.9	62.3	85.1	71.4	74.9	92.4	88.8	85.5	98.9	94.1	70.2	88.3	84.4	68.0	76.2	52.2	85.7	69.0	77.1
SICA†	**95.9**	87.1	**91.5**	**92.5**	91.8	99.2	98.6	95.5	**100.0**	97.1	77.3	94.6	90.7	86.4	88.6	62.9	87.6	75.3	87.5
CNNDet. (p=0.1)	70.0	62.7	80.6	71.7	71.3	83.9	76.8	70.9	99.5	82.3	82.0	82.6	86.2	72.5	79.3	55.6	84.0	69.8	75.7
CNNDet. (p=0.5)	69.0	64.9	79.3	69.1	70.6	82.6	74.3	66.4	97.6	77.9	82.0	80.1	83.6	67.0	75.3	59.4	86.3	72.9	74.7
NPR	84.1	43.3	75.0	61.5	66.0	89.4	90.0	81.5	89.9	82.6	71.5	84.1	92.8	75.3	84.0	59.1	84.3	71.7	76.5
AIDE	72.7	64.2	50.8	52.0	59.9	84.2	72.6	80.5	83.0	54.3	52.7	71.2	77.9	66.5	72.2	29.0	78.3	53.6	64.2
PGC	94.3	92.1	88.6	84.9	90.0	96.6	97.9	98.3	99.8	96.0	65.5	92.4	95.2	83.0	89.1	61.3	82.9	72.1	85.9
RINE	91.5	81.4	78.7	81.1	83.2	96.7	96.0	89.4	**100.0**	95.0	70.2	91.2	**96.0**	89.7	**92.8**	42.5	81.5	62.0	82.3
SAFE	91.5	79.5	55.4	49.6	69.0	75.4	73.6	74.6	93.2	71.0	77.7	77.6	89.0	80.3	84.7	56.7	53.9	55.3	71.6
MARE	77.4	58.4	85.0	70.6	72.9	89.4	74.4	81.7	99.4	82.5	82.7	85.0	81.0	62.1	71.6	58.0	89.3	73.6	75.8
DRCT‡	56.9	41.9	61.1	52.7	53.2	42.4	96.5	70.8	89.3	93.2	81.5	79.0	80.3	79.9	80.1	38.7	57.7	48.2	65.1
DoU‡	70.3	13.6	42.4	65.4	47.9	98.4	98.1	95.1	**100.0**	96.5	57.2	90.9	75.4	72.9	74.1	36.5	54.9	45.7	64.7
GenSign (ProGAN20)‡	74.4	16.2	50.0	66.3	51.7	**99.9**	97.8	96.0	99.8	93.7	28.7	86.0	48.4	48.3	48.4	35.6	55.8	45.7	57.9
GenSign (SD1.4)‡	91.8	60.2	43.1	60.1	63.8	94.7	96.1	78.8	96.8	93.7	51.8	85.3	90.2	74.2	82.2	46.7	41.4	44.1	68.8
ForenLens (CLIP)	94.0	89.2	90.8	91.1	91.3	99.3	98.5	95.8	99.8	**97.8**	74.2	94.2	88.7	82.5	85.6	67.5	90.4	79.0	87.5
ForenLens (SigLIP2)	91.9	85.9	89.1	87.7	88.7	99.4	**99.1**	**98.5**	95.7	96.9	**94.7**	**97.4**	92.7	85.5	89.1	**72.5**	89.1	**80.8**	**89.0**

Table 11: **Complete OpenMMSec Accuracy (%).**

Method	Deepfake					AIGC							IMDL			Document			Avg
	EFS	FE	FR	FS	Avg	GAN	Lat.	Pix.	AR	Comm.	Other	Avg	Gen.	Non-Gen.	Avg	AIGC	Non-AIGC	Avg	
ResNet†	71.3	65.4	70.0	64.7	67.8	77.4	65.2	75.7	78.5	72.7	75.5	74.2	80.3	74.2	77.3	61.9	81.5	71.7	72.7
EfficientNet†	45.0	29.0	50.8	48.5	43.3	65.2	58.9	63.1	65.1	62.1	63.7	63.0	51.7	51.1	51.4	47.0	76.8	61.9	54.9
CapsuleNet†	63.4	58.7	66.5	59.6	62.0	79.2	63.1	77.5	81.3	73.2	77.8	75.4	79.2	75.3	77.2	59.8	84.8	72.3	71.7
SegFormer†	83.8	80.5	83.5	75.1	80.7	87.4	80.9	87.1	89.7	87.8	82.3	85.9	83.2	80.3	81.7	66.8	80.1	73.4	80.4
Swin†	83.6	80.5	80.6	71.5	79.0	86.9	83.8	85.8	87.6	86.8	81.3	85.4	85.0	80.8	82.9	66.9	77.1	72.0	79.8
ConvNeXt†	80.5	84.0	83.9	72.4	80.2	89.0	79.8	87.9	92.0	89.6	83.4	87.0	83.6	80.2	81.9	65.0	75.3	70.1	79.8
Recce†	59.0	50.8	50.7	53.7	53.5	90.4	66.1	90.8	97.1	85.3	87.8	86.3	63.1	77.1	70.1	78.6	52.8	65.7	68.9
Sia†	64.0	63.5	73.1	63.7	66.1	81.6	68.7	80.2	84.1	79.2	80.4	79.0	71.9	63.8	67.9	64.4	83.4	73.9	71.7
IML-ViT†	66.6	64.9	74.6	66.0	68.0	83.3	70.6	81.8	86.1	81.5	80.9	80.7	81.5	78.9	80.2	68.9	78.5	73.7	75.7
TruFor†	73.7	72.7	77.1	65.4	72.2	83.7	77.9	83.0	86.0	84.0	80.3	82.5	82.7	78.3	80.5	58.1	**86.6**	72.3	76.9
FFDN†	73.7	75.5	81.2	69.5	75.0	93.6	89.2	93.9	96.1	94.5	86.8	92.4	84.7	80.6	82.6	64.8	75.3	70.0	80.0
CLIP-FFT†	77.4	78.2	80.4	68.0	76.0	92.2	82.0	91.0	95.1	91.7	85.4	89.6	83.5	79.3	81.4	66.6	81.2	73.9	80.2
UnivFD†	57.0	44.8	57.6	57.4	54.2	84.1	68.3	81.7	84.9	76.9	82.9	79.8	65.7	76.2	70.9	39.0	86.3	62.7	66.9
Effort†	87.1	87.2	80.7	85.0	85.0	86.6	71.3	84.6	86.3	78.8	83.9	81.9	81.7	85.7	83.7	61.1	78.1	69.6	80.1
Mesorch†	76.7	72.6	79.9	73.4	75.7	83.0	73.8	82.1	85.5	83.3	81.0	81.4	81.2	78.7	79.9	63.6	81.8	72.7	77.4
CO-SPY†	75.1	71.3	77.0	65.5	72.2	84.7	80.8	84.2	86.2	85.7	78.5	83.3	77.8	74.2	76.0	61.9	78.0	70.0	75.4
SICA†	**91.3**	89.4	**86.5**	**86.6**	**88.4**	96.3	94.5	94.7	96.7	94.8	86.9	94.0	85.4	85.1	85.3	69.2	78.4	73.8	**85.4**
CNNDet. (p=0.1)	67.1	71.8	74.9	66.7	70.1	80.0	71.7	79.0	82.8	79.6	79.1	78.7	80.2	76.4	78.3	68.5	73.1	70.8	74.5
CNNDet. (p=0.5)	67.2	70.7	73.5	64.8	69.1	79.3	69.4	77.9	82.4	77.8	78.9	77.6	77.8	73.4	75.6	71.1	73.2	72.1	73.6
NPR	76.2	64.7	69.5	60.3	67.7	80.3	81.7	79.5	79.9	78.8	75.1	79.2	85.5	77.7	81.6	64.8	77.3	71.1	74.9
AIDE	59.2	48.8	50.1	50.5	52.1	59.8	62.5	57.4	56.0	54.0	53.4	57.2	66.4	60.0	63.2	28.6	85.0	56.8	57.3
PGC	87.2	**90.5**	82.1	79.6	84.9	92.6	92.4	93.8	94.2	91.9	84.1	91.5	**88.4**	82.9	85.7	71.4	71.8	71.6	83.4
RINE	88.5	89.8	74.7	75.5	82.1	90.6	89.6	88.1	91.3	90.0	81.8	88.6	87.1	84.8	86.0	51.9	83.7	67.8	81.1
SAFE	61.1	49.0	53.9	49.1	53.3	56.5	62.7	54.9	54.3	56.3	57.6	57.1	79.2	72.8	76.0	33.6	71.1	52.3	59.7
MARE	72.9	73.4	79.0	64.5	72.5	85.9	69.7	85.1	88.5	83.8	83.4	82.7	76.3	72.3	74.3	71.6	76.6	74.1	75.9
DRCT‡	58.6	59.6	57.3	52.1	56.9	76.0	88.8	82.5	86.2	86.4	83.0	83.8	66.8	79.0	72.9	61.9	42.2	52.0	66.4
DoU‡	66.1	56.1	48.2	62.2	58.2	94.7	92.3	94.9	**98.7**	94.3	86.9	93.6	68.6	78.1	73.3	**81.2**	15.8	48.5	68.4
GenSign (ProGAN20)‡	53.7	26.3	48.4	55.9	46.1	93.1	93.4	91.8	92.3	90.8	81.3	90.5	60.4	65.9	63.2	79.9	18.4	49.1	62.2
GenSign (SD1.4)‡	77.4	64.2	49.4	58.1	62.3	82.0	86.6	78.6	80.3	81.9	71.7	80.2	75.9	78.3	77.1	59.6	30.0	44.8	66.1
ForenLens (CLIP)	86.0	83.5	84.1	83.5	84.3	**96.8**	94.2	95.2	97.2	**95.6**	87.2	94.4	83.9	86.0	84.9	74.6	78.9	**76.7**	85.1
ForenLens (SigLIP2)	83.5	81.0	81.4	80.0	81.5	96.7	**95.9**	**96.4**	96.6	94.7	**93.9**	**95.7**	86.1	**87.3**	**86.7**	76.2	73.7	74.9	84.7

Table 12: **Complete OpenMMSec Average Precision (%).**

Method	Deepfake					AIGC							IMDL			Document			Avg
	EFS	FE	FR	FS	Avg	GAN	Lat.	Pix.	AR	Comm.	Other	Avg	Gen.	Non-Gen.	Avg	AIGC	Non-AIGC	Avg	
ResNet†	72.4	41.5	67.7	56.9	59.6	65.6	52.5	15.1	84.1	25.8	29.6	45.5	83.6	41.8	62.7	18.3	97.3	57.8	56.4
EfficientNet†	46.0	21.2	53.6	52.1	43.2	24.2	43.3	10.4	8.6	17.5	17.2	20.2	39.1	22.9	31.0	15.8	92.8	54.3	37.2
CapsuleNet†	61.1	41.0	68.2	52.6	55.7	55.8	48.8	12.2	28.9	19.6	31.7	32.8	81.0	39.0	60.0	14.2	96.6	55.4	51.0
SegFormer†	87.0	60.4	88.7	76.5	78.2	75.1	83.2	40.3	95.4	76.8	33.4	67.4	87.4	52.9	70.1	22.4	97.6	60.0	68.9
Swin†	89.7	63.8	86.0	70.7	77.5	80.9	87.4	46.1	97.0	79.1	33.4	70.6	90.1	56.9	73.5	23.3	97.1	60.2	70.5
ConvNeXt†	83.2	66.9	89.5	74.6	78.5	75.8	82.0	33.1	96.9	79.4	29.7	66.2	87.7	52.4	70.0	23.1	96.8	60.0	68.7
Recce†	63.6	37.2	50.5	50.1	50.3	63.5	60.1	17.8	69.6	31.9	35.1	46.3	47.4	27.1	37.3	22.6	93.3	57.9	48.0
Sia†	59.7	32.5	77.3	58.7	57.0	63.2	60.6	15.4	79.1	39.6	36.6	49.1	72.0	26.9	49.5	21.8	97.9	59.8	53.9
IML-ViT†	68.3	57.4	78.4	59.7	65.9	70.4	63.9	16.9	98.0	50.1	37.5	56.1	84.9	47.9	66.4	17.8	97.6	57.7	61.5
TruFor†	72.9	36.3	78.3	58.8	61.6	66.8	77.2	25.9	70.7	69.4	32.8	57.1	86.7	47.9	67.3	17.2	**98.0**	57.6	60.9
FFDN†	74.0	43.9	85.7	69.2	68.2	86.8	94.1	69.8	**99.9**	91.4	42.4	80.7	91.2	57.1	74.2	19.8	97.0	58.4	70.4
CLIP-FFT†	77.9	49.8	82.1	65.8	68.9	84.6	86.6	46.2	**99.9**	81.0	30.8	71.5	86.6	46.7	66.6	20.7	97.9	59.3	66.6
UnivFD†	87.7	88.2	68.9	72.1	79.2	75.3	60.3	25.9	83.8	24.3	58.7	54.7	54.8	68.5	61.6	12.9	94.5	53.7	62.3
Effort†	94.1	**93.3**	87.9	91.6	**91.7**	87.1	67.3	48.4	91.0	30.5	48.0	62.0	85.8	**78.7**	82.2	16.8	96.2	56.5	73.1
Mesorch†	76.1	53.2	85.0	69.3	70.9	68.5	72.3	21.1	92.8	69.6	42.2	61.1	84.1	47.8	65.9	21.6	97.8	59.7	64.4
CO-SPY†	79.4	45.8	86.1	65.3	69.2	78.5	85.9	47.1	92.4	81.9	25.6	68.6	81.7	40.6	61.1	19.2	97.2	58.2	64.3
SICA†	**95.5**	82.1	**92.4**	**92.2**	90.5	97.0	97.9	77.3	99.7	91.2	34.9	83.0	89.4	70.8	80.1	27.1	97.5	62.3	79.0
CNNDet. (p=0.1)	57.5	35.4	71.3	59.1	55.9	63.8	68.8	15.3	94.2	58.0	36.7	56.1	82.4	40.7	61.5	20.4	96.8	58.6	58.0
CNNDet. (p=0.5)	59.2	40.7	72.9	58.2	57.7	62.3	63.5	12.1	82.8	41.3	34.9	49.5	79.0	35.4	57.2	21.4	97.2	59.3	55.9
NPR	86.5	38.6	69.8	53.0	62.0	79.0	89.1	68.1	48.9	68.8	25.9	63.3	90.3	44.8	67.5	22.4	96.7	59.5	63.1
AIDE	66.7	35.2	47.6	46.2	48.9	52.5	62.7	34.2	16.6	19.9	17.0	33.8	72.3	44.0	58.1	12.3	94.8	53.6	48.6
PGC	91.7	77.1	88.8	84.3	85.4	84.4	96.5	86.2	85.8	84.0	22.6	76.6	**93.5**	63.0	78.2	25.8	96.4	61.1	75.3
RINE	91.8	80.4	79.9	81.1	83.3	87.8	93.9	34.6	99.7	84.3	24.2	70.7	**93.5**	74.1	**83.8**	15.1	95.6	55.4	73.3
SAFE	90.3	48.3	47.8	42.2	57.2	47.4	63.2	38.1	47.1	32.4	26.1	42.4	82.2	47.1	64.7	19.8	86.9	53.4	54.4
MARE	68.1	36.1	78.3	58.3	60.2	71.4	63.8	31.3	89.6	59.2	43.6	59.8	76.5	32.0	54.3	21.4	97.8	59.6	58.5
DRCT‡	52.7	17.8	54.2	44.9	42.4	11.6	95.5	15.9	21.1	82.0	45.9	45.3	74.3	57.0	65.6	16.2	87.2	51.7	51.3
DoU‡	65.2	11.2	42.1	59.2	44.4	90.8	97.4	69.9	98.4	88.1	16.8	76.9	71.3	50.8	61.1	13.7	86.4	50.0	58.1
GenSign (ProGAN20)‡	72.6	11.4	48.1	58.2	47.6	**99.3**	97.1	83.5	90.2	83.0	11.2	77.4	50.8	25.9	38.4	15.1	87.1	51.1	53.6
GenSign (SD1.4)‡	90.2	24.0	43.0	53.2	52.6	62.0	86.3	23.8	28.3	60.1	12.4	45.5	82.6	45.0	63.8	16.1	82.1	49.1	52.8
ForenLens (CLIP)	92.5	83.7	90.8	90.2	89.3	96.3	97.5	73.2	95.0	**91.9**	34.4	81.4	87.0	71.5	79.2	32.4	**98.0**	65.2	78.8
ForenLens (SigLIP2)	91.1	80.7	89.9	86.8	87.1	97.0	**98.7**	**90.2**	82.0	90.4	**84.1**	**90.4**	91.1	75.9	83.5	**34.6**	97.7	**66.2**	**81.8**

Table 13: **Complete OpenMMSec F1 (%).**

Method	Deepfake					AIGC							IMDL			Document			Avg
	EFS	FE	FR	FS	Avg	GAN	Lat.	Pix.	AR	Comm.	Other	Avg	Gen.	Non-Gen.	Avg	AIGC	Non-AIGC	Avg	
ResNet†	66.5	33.2	68.4	59.1	56.8	47.4	48.2	24.5	20.1	31.5	37.5	34.9	75.0	46.6	60.8	20.9	88.5	54.7	51.8
EfficientNet†	56.4	31.6	63.4	60.8	53.1	35.7	47.3	17.7	12.8	27.7	29.2	28.4	44.8	31.5	38.2	24.9	85.7	55.3	43.7
CapsuleNet†	57.5	26.5	66.1	54.5	51.1	46.7	39.5	19.7	19.8	22.5	38.1	31.0	69.9	36.6	53.2	12.6	90.7	51.7	46.8
SegFormer†	79.4	52.9	81.9	68.4	70.7	61.5	72.0	42.1	35.2	65.2	31.7	51.3	75.0	46.3	60.7	26.8	87.3	57.1	59.9
Swin†	80.1	58.1	78.9	63.9	70.2	63.6	78.1	43.5	31.1	65.1	34.6	52.7	78.8	51.0	64.9	27.9	85.2	56.5	61.1
ConvNeXt†	72.8	59.9	81.9	62.1	69.1	64.0	68.7	36.5	41.1	68.5	27.7	51.1	75.8	46.0	60.9	27.7	83.9	55.8	59.2
Recce†	57.9	29.0	49.1	52.5	47.1	55.8	21.4	14.9	59.5	20.9	27.2	33.3	6.4	3.5	4.9	7.3	61.8	34.5	30.0
Sia†	55.4	31.4	73.1	58.6	54.6	49.9	50.7	23.5	24.8	42.5	42.0	38.9	66.0	31.7	48.8	25.7	89.8	57.7	50.0
IML-ViT†	63.3	45.7	76.6	64.8	62.6	52.6	53.3	24.0	28.6	47.8	38.6	40.8	72.1	41.9	57.0	11.7	86.0	48.9	52.3
TruFor†	66.0	39.3	75.2	55.1	58.9	54.5	68.6	33.0	28.5	58.1	35.6	46.4	74.7	40.8	57.7	18.7	91.9	55.3	54.6
FFDN†	62.5	35.5	79.2	59.0	59.1	77.2	84.3	64.4	58.7	82.3	30.9	66.3	75.4	32.7	54.0	24.0	83.8	53.9	58.3
CLIP-FFT†	67.4	37.9	77.1	53.8	59.1	72.5	71.3	44.4	53.2	71.9	24.5	56.3	74.3	34.1	54.2	25.2	88.1	56.7	56.6
UnivFD†	63.5	40.7	65.6	64.9	58.7	58.2	48.8	29.8	26.8	29.6	51.1	40.7	51.5	59.5	55.5	18.9	**92.2**	55.5	52.6
Effort†	84.4	73.6	78.3	83.2	79.9	65.0	54.7	41.7	28.8	33.9	52.0	46.0	74.5	**71.0**	**72.8**	18.8	86.1	52.4	62.8
Mesorch†	72.4	45.5	79.7	70.0	66.9	53.1	60.9	29.5	27.7	56.2	41.4	44.8	72.4	44.7	58.6	26.6	88.6	57.6	57.0
CO-SPY†	68.6	34.8	75.4	55.8	58.6	58.6	73.2	38.1	28.2	62.9	23.3	47.4	67.3	32.5	49.9	25.7	86.0	55.9	53.0
SICA†	**88.9**	73.5	**84.6**	**83.9**	**82.7**	87.6	92.4	68.7	62.7	82.8	27.6	70.3	77.7	59.3	68.5	28.8	86.0	57.4	69.7
CNNDet. (p=0.1)	52.8	34.4	72.0	57.1	54.1	46.9	58.4	23.1	24.4	48.0	40.2	40.2	71.0	38.1	54.6	17.2	81.9	49.6	49.6
CNNDet. (p=0.5)	56.0	38.8	71.2	56.0	55.5	45.0	53.8	17.6	23.3	41.4	40.7	37.0	68.8	35.5	52.2	15.9	81.8	48.8	48.4
NPR	73.0	26.9	67.0	50.3	54.3	53.5	77.6	37.1	18.7	50.4	29.9	44.5	79.6	37.8	58.7	26.0	85.4	55.7	53.3
AIDE	62.2	36.9	52.6	52.2	51.0	37.9	59.8	24.1	10.1	23.5	18.4	29.0	63.1	40.1	51.6	21.5	91.5	56.5	47.0
PGC	82.7	**76.5**	78.3	73.1	77.7	75.7	89.8	70.2	48.9	74.4	19.4	63.1	**82.9**	48.1	65.5	27.6	80.8	54.2	65.1
RINE	84.6	73.9	66.2	65.6	72.6	71.7	86.2	42.3	39.1	71.2	19.2	55.0	81.3	60.8	71.1	15.7	90.2	52.9	62.9
SAFE	65.8	40.6	58.9	51.7	54.2	32.9	61.2	21.2	10.7	33.7	34.7	32.4	76.1	53.7	64.9	32.7	82.4	57.6	52.3
MARE	62.6	34.4	76.9	51.1	56.2	58.0	48.5	34.3	32.4	51.7	43.3	44.7	66.3	32.3	49.3	21.1	84.5	52.8	50.8
DRCT‡	47.1	23.5	50.7	38.9	40.1	9.0	86.0	24.1	22.0	64.6	46.2	42.0	22.7	18.1	20.4	15.8	51.9	33.8	34.1
DoU‡	58.2	0.1	30.0	55.2	35.9	77.9	88.7	60.8	**81.2**	77.9	8.6	65.8	27.9	5.6	16.7	0.2	1.2	0.7	29.8
GenSign (ProGAN20)‡	60.5	5.3	55.6	64.0	46.3	79.9	91.5	63.5	42.0	72.7	7.7	59.6	36.0	22.4	29.2	5.8	7.9	6.9	35.5
GenSign (SD1.4)‡	75.4	29.8	32.1	47.6	46.2	59.5	84.5	32.5	22.1	60.4	12.2	45.2	56.7	26.7	41.7	19.1	36.1	27.6	40.2
ForenLens (CLIP)	83.3	65.8	83.2	81.7	78.5	**89.1**	91.9	70.3	66.2	**85.4**	26.2	71.5	75.3	64.3	69.8	32.7	86.2	**59.5**	69.8
ForenLens (SigLIP2)	80.4	61.6	80.3	77.6	75.0	88.7	**94.4**	**79.9**	57.8	81.8	**75.0**	**79.6**	78.5	65.5	72.0	**35.4**	82.2	58.8	**71.3**

Across the complete taxonomy, ForenLens (SigLIP2) attains the strongest aggregate ROC-AUC, AP, and F1 while remaining competitive in Accuracy. The CLIP variant also stays near the leading aggregate results, showing that the benefit of cross-layer evidence readout transfers across both reported backbones and is not tied to a single evaluation metric.

E.2 Cross-benchmark Evaluation

The 7 external benchmarks use the same 51 constituent test sets as Table 1. Daggered OpenMMSec entries use values reported by SICA (Du et al., 2026); all external entries use the unified evaluation. Average Precision is reported in the main comparison, Table 1; the tables below provide the three complementary metrics under the same group aggregation.

Table 14: **Cross-benchmark ROC-AUC (%).**

Method	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
CNNDet. (p=0.1)	75.7	68.7	84.1	51.4	65.5	58.9	86.9	78.8	71.2
CNNDet. (p=0.5)	74.7	61.9	74.6	46.9	61.9	59.7	79.6	74.4	66.7
UnivFD	74.0†	85.7	68.2	80.5	81.2	73.9	85.3	80.4	78.6
NPR	76.5	80.9	93.5	66.3	49.5	41.4	90.9	74.0	71.6
RINE	82.3	98.7	**98.9**	48.0	47.5	31.5	99.3	68.0	71.8
Mesorch	80.5†	79.4	86.4	55.4	58.2	44.5	92.2	78.2	71.8
AIDE	64.2	83.5	62.2	46.1	49.5	49.4	72.4	49.7	59.6
Effort	84.1†	88.5	36.2	29.5	24.6	16.6	60.8	45.3	48.2
SAFE	71.6	84.1	96.4	46.4	48.9	62.0	93.4	51.8	69.3
SICA	87.5†	98.2	93.8	66.0	62.7	47.9	98.6	93.7	81.0
PGC	85.9	98.7	96.0	42.8	55.0	40.9	98.9	73.5	73.9
MARE	75.8	71.4	73.3	38.2	71.2	61.7	79.7	83.9	69.4
DRCT‡	65.1	81.6	94.5	69.4	78.3	74.6	97.6	80.6	80.2
DoU‡	64.7	**99.2**	91.4	40.4	49.6	45.8	94.5	72.3	69.7
GenSign (ProGAN20)‡	57.9	98.4	65.4	53.8	54.5	46.9	76.6	67.4	65.1
GenSign (SD1.4)‡	68.8	98.0	97.9	50.3	27.2	36.6	**99.5**	69.3	68.5
ForenLens (CLIP)	87.5	97.1	97.7	61.9	69.8	68.7	95.8	92.0	83.8
ForenLens (SigLIP2)	**89.0**	98.1	94.6	**85.7**	**92.3**	**88.1**	99.3	**99.1**	**93.3**

Table 15: **Cross-benchmark Accuracy (%).**

Method	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
CNNDet. (p=0.1)	74.5	60.5	70.3	51.3	61.4	55.1	75.8	72.3	65.1
CNNDet. (p=0.5)	73.6	57.8	62.8	48.4	58.2	55.8	71.2	68.6	62.1
UnivFD	66.9†	75.0	57.4	72.9	74.1	69.3	74.2	74.3	70.5
NPR	74.9	78.1	84.2	59.6	50.1	43.7	79.7	68.6	67.4
RINE	81.1	90.7	83.8	50.4	50.3	49.9	88.1	60.4	69.3
Mesorch	77.4†	70.5	78.4	54.6	57.5	50.1	81.3	71.7	67.7
AIDE	57.3	72.2	57.6	47.2	47.4	47.4	64.6	48.0	55.2
Effort	80.1†	81.6	45.6	39.6	48.8	53.3	59.8	51.7	57.5
SAFE	59.7	76.0	89.9	49.9	49.4	51.3	84.9	50.2	63.9
SICA	**85.4**†	91.1	79.0	60.9	64.1	58.7	91.4	85.7	77.1
PGC	83.4	93.3	73.2	45.7	56.1	54.8	91.2	66.9	70.6
MARE	75.9	63.0	66.5	41.3	65.8	59.0	70.9	76.7	64.9
DRCT‡	66.4	72.2	70.4	59.3	72.1	70.6	80.5	73.8	70.7
DoU‡	68.4	91.4	66.2	50.0	55.7	59.1	79.8	59.0	66.2
GenSign (ProGAN20)‡	62.2	85.5	63.3	51.9	58.3	59.3	71.6	63.7	64.5
GenSign (SD1.4)‡	66.1	**95.3**	83.4	49.8	47.1	49.8	88.2	62.8	67.8
ForenLens (CLIP)	85.1	89.0	**91.7**	56.2	66.0	63.4	87.8	85.1	78.0
ForenLens (SigLIP2)	84.7	92.0	84.6	**78.3**	**84.3**	**81.1**	**97.1**	**94.4**	**87.1**

Table 16: **Cross-benchmark F1 (%).**

Method	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
CNNDet. (p=0.1)	49.6	44.4	59.6	58.6	60.0	51.5	68.6	69.8	57.8
CNNDet. (p=0.5)	48.4	38.4	46.7	57.6	56.2	52.6	63.2	66.6	53.7
UnivFD	52.6†	67.4	32.6	70.5	73.1	61.0	66.7	73.0	62.1
NPR	53.3	72.1	79.2	66.5	45.4	38.7	70.0	64.7	61.2
RINE	62.9	88.4	75.9	49.2	44.3	22.7	83.3	61.9	61.1
Mesorch	57.0†	67.4	72.1	49.4	51.9	33.4	77.0	71.0	59.9
AIDE	47.0	75.2	54.6	58.0	44.0	41.0	64.4	45.3	53.7
Effort	62.8†	78.0	17.3	52.0	5.9	1.1	49.0	11.9	34.7
SAFE	52.3	74.3	89.8	66.2	61.1	61.2	83.7	63.6	69.0
SICA	69.7†	91.1	70.6	57.0	55.1	36.0	89.7	84.4	69.2
PGC	65.1	93.1	61.2	41.8	47.0	22.3	89.8	66.1	60.8
MARE	50.8	54.6	54.7	47.6	64.1	51.2	63.6	74.9	57.7
DRCT‡	34.1	57.0	55.1	42.8	69.0	62.9	74.1	70.4	58.2
DoU‡	29.8	89.5	40.4	0.2	20.8	8.8	63.4	26.9	35.0
GenSign (ProGAN20)‡	35.5	87.8	53.2	17.2	35.0	11.0	64.0	46.6	43.8
GenSign (SD1.4)‡	40.2	**95.0**	71.9	49.5	22.3	21.2	81.4	59.3	55.1
ForenLens (CLIP)	69.8	89.2	**90.9**	26.0	52.4	29.5	83.1	84.7	65.7
ForenLens (SigLIP2)	**71.3**	91.6	81.6	**77.9**	**84.7**	**76.8**	**96.9**	**94.6**	**84.4**

Among the configurations in these cross-benchmark tables, ForenLens (SigLIP2) achieves the highest eight-group mean ROC-AUC, Accuracy, and F1, consistent with the main AP comparison. The advantages under both ranking and thresholded metrics support its generalization across test distributions.

F Complete Post-processing Robustness Results

We report the AP values underlying Figure 5 at every JPEG and Gaussian-blur setting. Each cell equally averages OpenMMSec and the 7 external groups.

Table 17: **Eight-group AP (%) under JPEG recompression with quality Q.**

Method	Clean	100	90	80	70	60	50	40	30
CNNDet. (p=0.5)	64.2	63.5	61.7	59.4	58.2	56.9	56.5	56.7	56.7
NPR	71.8	65.4	55.3	54.3	52.0	51.6	51.0	49.7	49.3
SAFE	66.5	54.1	50.9	50.3	49.6	49.4	49.7	49.2	50.2
UnivFD	77.1	76.9	76.6	75.1	75.5	72.1	73.2	73.3	73.4
RINE	73.9	73.2	47.9	45.4	46.4	45.6	47.3	66.3	69.0
SICA	80.8	78.2	73.7	67.8	67.4	66.3	66.8	67.4	68.3
Effort	52.5	46.7	45.0	43.7	42.9	41.8	41.0	40.6	41.1
ForenLens (CLIP)	83.5	81.4	82.9	85.4	86.3	85.5	85.9	86.3	86.2
ForenLens (SigLIP2)	93.2	92.2	92.2	92.1	92.0	91.6	91.5	92.3	91.6

Table 18: **Eight-group AP (%) under Gaussian blur with standard deviation σ.**

Method	Clean	0.5	1.0	1.5	2.0	2.5	3.0
CNNDet. (p=0.5)	64.2	61.2	58.0	62.5	65.1	65.1	63.8
NPR	71.8	70.8	61.7	57.0	56.5	56.5	56.4
SAFE	66.5	66.9	54.7	52.1	50.5	48.8	47.7
UnivFD	77.1	76.1	72.6	68.8	68.8	69.1	68.6
RINE	73.9	72.5	67.2	65.7	65.6	65.1	64.5
SICA	80.8	76.4	70.8	68.4	69.2	67.1	63.5
Effort	52.5	50.4	46.6	42.8	41.2	41.2	42.0
ForenLens (CLIP)	83.5	81.5	76.3	75.0	74.1	71.2	67.4
ForenLens (SigLIP2)	93.2	92.0	90.4	89.8	88.9	87.0	84.9

ForenLens (SigLIP2) remains the strongest method at every JPEG quality and blur strength in AP. Its clear margins under the strongest recompression and blur show that cross-layer evidence readout remains effective after substantial post-processing. The CLIP configuration also leads the compared baselines under most conditions, showing that the trend is not tied to a single backbone.

G Supplementary Experiments and Analyses

This section consolidates all four metrics for the main ablations and examines backbone-dependent behavior, additional backbone metrics, and decision-readout analyses.

G.1 Compact controlled ablations across backbones

Tables 2 and 3 highlight AP. The tables below provide all four metrics under the same evaluation splits and equal eight-group aggregation. Query-composition controls remove individual expert or free queries, while component controls vary the readout, layer budget, routing, and auxiliary losses. Removing $\mathcal{L}_{\text{guide}}$ tests spatial guidance at fixed query count and capacity. Top-K follows the ID-only *LayerProbe* orders $\{9, 22, 14, 7\}$ for CLIP and $\{11, 26, 7, 14\}$ for SigLIP2 (one-based); shaded rows identify the fixed complete model.

Table 19: **Query-composition ablation: eight-group means across four metrics (%).** N/J/B/V/F denote Noiseprint++, JPEG Ghost, boundary transition, SD-VAE, and the free query.

N	J	B	V	F	ROC-AUC		Accuracy		AP		F1	
					CLIP	SigLIP2	CLIP	SigLIP2	CLIP	SigLIP2	CLIP	SigLIP2
×	✓	✓	✓	✓	80.5	93.2	76.3	86.9	80.9	92.6	64.5	84.3
✓	×	✓	✓	✓	79.7	92.2	77.1	85.4	80.6	91.9	64.3	83.1
✓	✓	×	✓	✓	79.4	92.6	76.6	85.3	80.5	92.0	63.6	83.1
✓	✓	✓	×	✓	82.8	93.1	76.5	86.9	82.3	92.7	62.6	84.5
✓	✓	✓	✓	×	80.8	92.0	76.9	86.0	81.4	92.0	63.8	83.1
✓	✓	✓	✓	✓	83.8	93.3	78.0	87.1	83.5	93.2	65.7	84.4

Table 20: **Reader, selected-layer, routing, and loss controls: eight-group means across four metrics (%).**

Reader	Top-K	Routing	$\mathcal{L}_{\text{guide}}$	$\mathcal{L}_{\text{orth}}$	$\mathcal{L}_{\text{rec}}$	ROC-AUC		Accuracy		AP		F1	
						CLIP	SigLIP2	CLIP	SigLIP2	CLIP	SigLIP2	CLIP	SigLIP2
Pooled	4	–	–	–	–	81.5	89.6	73.4	81.0	79.5	88.9	64.0	78.7
Query	1	Adaptive	✓	✓	✓	80.9	82.6	74.6	76.2	79.5	81.8	59.6	69.3
Query	2	Adaptive	✓	✓	✓	81.5	94.0	76.4	87.1	80.6	93.6	64.6	84.8
Query	3	Adaptive	✓	✓	✓	81.0	93.6	77.5	87.5	81.0	93.4	65.8	84.6
Query	4	Uniform	✓	✓	✓	82.9	92.6	77.6	85.9	82.8	92.3	66.8	83.8
Query	4	Static	✓	✓	✓	83.0	92.3	77.5	85.8	83.2	91.9	65.1	83.1
Query	4	Adaptive	×	✓	✓	79.4	91.1	77.0	84.6	80.5	90.5	63.8	82.6
Query	4	Adaptive	✓	×	✓	81.6	92.3	75.5	86.5	81.4	92.2	63.3	84.1
Query	4	Adaptive	✓	✓	×	78.9	92.1	74.0	85.3	79.1	91.9	57.2	83.5
Query	4	Adaptive	✓	✓	✓	83.8	93.3	78.0	87.1	83.5	93.2	65.7	84.4

Across all four metrics, the tables show a consistent overall pattern. The complete query composition remains strong on both backbones, and removing individual roles generally lowers the aggregate. The component controls likewise show that query-based readout, complementary selected layers, and training supervision jointly yield stable gains.

G.2 Backbone-dependent component behavior

Across both backbones, the principal effect of Top-K is to expose complementary selected-layer evidence. Moving from Top-1 to Top-2 raises AP by 1.1% and 11.8% (absolute) for CLIP and SigLIP2, respectively. Additional layers continue to help CLIP, whereas SigLIP2 saturates after Top-2, showing that the amount of useful complementary evidence is backbone-dependent. Our default configuration uses four layers selected using only in-domain data. The full query-based readout framework also improves over pooling on both backbones, from 79.5% to 83.5% AP for CLIP and from 88.9% to 93.2% for SigLIP2.

The uniform and static controls retain access to all four layer-specific responses but use image-independent layer weights. Relative to these controls, adaptive routing improves AP by 0.3–0.7% and 0.9–1.3% (absolute) on CLIP and SigLIP2, respectively. The trainable reader and decision readout may partly compensate for fixed layer aggregation, helping explain the moderate performance

gaps. Adaptive routing learns query- and image-dependent layer preferences, making the cross-layer division of labor explicit.

Role guidance remains effective across both backbones. Removing $\mathcal{L}_{\text{guide}}$ lowers AP by 3.0% and 2.7% (absolute) on CLIP and SigLIP2, respectively, and individual expert-query removals also tend to have smaller margins on SigLIP2. These comparable gains show that explicit role guidance remains useful with stronger frozen features. The larger gains from orthogonality and reconstruction on CLIP further indicate that the benefits of representation-level supervision depend on the backbone. The results show that, when trained separately with each frozen hierarchy, the same method learns a backbone-specific division of labor.

G.3 Additional backbone metrics

Table 4 reports AP in the main paper. Here we report ROC-AUC, Accuracy, and F1 under the same fixed query composition, candidate-layer protocol, and full released splits. Protocol and eight-group averaging follow the main AP table for all three supplementary metrics.

Table 21: **Backbone ablation: ROC-AUC (%).**

Backbone	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
Standalone vision encoders									
CLIP	87.5	97.1	97.7	61.9	69.8	68.7	95.8	92.0	83.8
DINOv2	84.5	93.9	76.3	83.2	79.1	73.0	94.1	89.0	84.1
DINOv3	86.5	94.5	74.7	82.5	85.2	76.2	94.7	95.7	86.2
SigLIP	83.0	93.1	82.4	90.3	72.4	65.0	95.5	94.7	84.6
SigLIP2	89.0	98.1	94.6	85.7	92.3	88.1	99.3	99.1	93.3
MAE	76.8	78.0	64.3	70.7	59.7	52.5	80.5	77.1	69.9
SAM	80.0	86.9	98.1	79.0	51.9	49.9	97.6	81.9	78.2
MLLM vision encoders									
Qwen3-VL	86.1	93.1	99.8	88.2	75.6	70.1	99.6	97.0	88.7
Qwen2.5-VL	84.3	93.5	99.9	94.9	66.4	69.2	97.6	88.9	86.8
Qwen3.5	84.9	94.1	99.9	91.4	68.2	61.0	98.8	94.5	86.6
InternVL2.5	83.1	88.6	94.0	81.4	83.7	77.3	96.9	91.2	87.0
InternVL3.5	81.2	83.8	86.5	86.7	78.5	66.3	91.8	91.7	83.3
Kimi-VL	81.8	93.2	99.9	99.5	57.9	46.1	97.6	86.6	82.8
GLM-4.6V	85.7	92.6	99.4	88.3	56.2	41.1	98.4	86.8	81.1

Table 22: **Backbone ablation: Accuracy (%).**

Backbone	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
Standalone vision encoders									
CLIP	85.1	89.0	91.7	56.2	66.0	63.4	87.8	85.1	78.0
DINOv2	80.7	86.6	67.4	74.8	72.3	67.4	89.1	82.1	77.6
DINOv3	79.1	86.0	69.8	74.6	75.3	67.7	88.6	86.5	78.4
SigLIP	79.5	83.0	61.6	65.8	65.8	62.8	86.4	86.5	73.9
SigLIP2	84.7	92.0	84.6	78.3	84.3	81.1	97.1	94.4	87.1
MAE	72.3	70.1	59.7	64.3	57.3	51.4	73.5	68.6	64.6
SAM	78.3	79.2	91.5	72.7	52.9	52.4	88.5	76.1	74.0
MLLM vision encoders									
Qwen3-VL	82.0	83.6	98.0	74.8	69.7	64.1	96.7	90.1	82.4
Qwen2.5-VL	82.7	81.3	99.0	80.6	65.5	63.4	87.3	84.0	80.5
Qwen3.5	83.3	83.9	97.9	80.5	67.1	60.0	87.4	90.8	81.4
InternVL2.5	77.9	77.6	82.2	75.7	75.9	71.5	88.8	82.6	79.0
InternVL3.5	76.2	74.7	77.2	79.6	69.2	60.0	83.6	83.5	75.5
Kimi-VL	82.6	86.7	98.9	92.9	62.6	60.1	86.3	80.4	81.3
GLM-4.6V	83.2	83.7	96.3	64.6	57.3	49.2	92.5	81.4	76.0

Table 23: **Backbone ablation: F1 (%).**

Backbone	OpenMMSec	AIGC-DB	AIGI-P	AIGI-S	AIGIBench	Chameleon	Holmes	WildRF	Avg
Standalone vision encoders									
CLIP	69.8	89.2	90.9	26.0	52.4	29.5	83.1	84.7	65.7
DINOv2	62.5	85.9	53.8	68.1	70.4	59.0	87.5	81.2	71.1
DINOv3	63.0	86.6	61.6	74.4	77.4	65.9	88.1	87.6	75.6
SigLIP	60.5	80.1	38.6	48.5	60.1	48.1	83.2	85.7	63.1
SigLIP2	71.3	91.6	81.6	77.9	84.7	76.8	96.9	94.6	84.4
MAE	52.0	68.3	49.5	65.4	60.4	50.9	72.4	71.3	61.3
SAM	55.1	80.2	90.4	69.4	39.1	26.1	85.8	73.4	64.9
MLLM vision encoders									
Qwen3-VL	65.8	83.3	98.0	77.4	72.3	58.6	96.5	90.6	80.3
Qwen2.5-VL	63.3	77.7	98.9	83.2	52.0	33.6	82.1	82.9	71.7
Qwen3.5	63.4	82.2	97.8	81.7	61.9	39.7	82.7	90.4	75.0
InternVL2.5	60.4	77.6	77.8	72.1	75.6	61.0	87.4	83.5	74.4
InternVL3.5	57.3	73.3	71.9	78.8	73.1	58.7	81.8	84.5	72.4
Kimi-VL	63.7	87.4	98.8	93.3	41.2	19.7	78.9	77.3	70.1
GLM-4.6V	64.5	83.5	96.1	72.7	49.4	24.7	90.7	80.2	70.2

Across ROC-AUC, Accuracy, AP, and F1, ForenLens is instantiated and its readout is trained separately with all 14 standalone and MLLM vision encoders. SigLIP2 achieves the strongest eight-group average on every metric, while several other backbones specialize on individual benchmark groups. These results support the applicability of a common ForenLens design across backbones and highlight their distinct forensic strengths.

G.4 Additional CLIP query analysis

We analyze the same trained CLIP model used for the main-paper mechanism visualizations. Class-balanced sampling over the 14 OpenMMSec primary types and 51 external subsets yields 126,502 samples.

G.4.1 Four-layer query-adaptive routing

For query q, Eq. (2) obtains the routing weights $\alpha_{q\ell}$ by applying a softmax over the selected layers $\ell \in \mathcal{K}$ to the sum of the learned, image-independent layer prior $b_{q\ell}$ and the image-conditioned score $\langle \mathrm{LN}(\boldsymbol{t}_{q\ell}), \boldsymbol{u}_q \rangle / \sqrt{d}$. We examine the four selected layers (7, 9, 14, and 22).

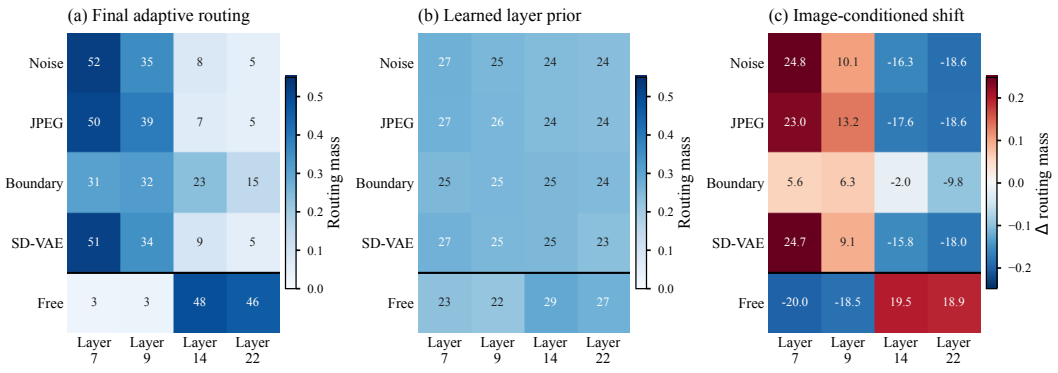

Figure 8: **Layer-routing decomposition of the CLIP model.** (a) Mean adaptive routing weights; (b) learned layer priors $\mathrm{softmax}_{\ell \in \mathcal{K}}(b_{q\ell})$; (c) the difference (a) minus (b). Cell annotations use a percentage scale; color bars show weights and differences as decimals.

The learned layer prior remains close to uniform over the four selected layers. The image-conditioned score creates the separation: the Noiseprint++, JPEG Ghost, and SD-VAE queries shift mass toward Layers 7 and 9, the free query shifts it toward Layers 14 and 22, and the boundary-transition query remains distributed across layers. The query-specific layer preferences therefore arise mainly from the image-conditioned score rather than the learned prior.

G.4.2 DOMAIN-CONDITIONED EVIDENCE COMPOSITION

Decision-readout weights are averaged separately for real and fake samples within each subset, then equally over subsets within each of eleven groups: the four OpenMMSec domains and seven external benchmark groups. Table 24 reports these real/fake means (left) and the cross-group ranges after averaging the two class means equally (right).

Table 24: **CLIP decision-readout composition across groups.** Left: real/fake mean weights (%) after equal-set aggregation. Right: the range over eleven label-balanced group means and the groups attaining its endpoints. N/J/B/V/F denote Noiseprint++/ JPEG Ghost/boundary-transition/SD-VAE/free queries; $\Delta = \mathrm{Max} - \mathrm{Min}$ is the absolute difference (%).

Real/fake readout weights

Group	N	J	B	V	F
OM-AIGC	6.5/1.2	31.1/11.2	24.7/58.3	10.3/4.1	27.4/25.2
OM-Deepfake	3.1/0.3	25.5/7.2	44.9/67.2	0.9/0.1	25.6/25.2
OM-IMDL	8.3/2.0	26.9/12.7	29.1/56.1	9.2/4.0	26.5/25.2
OM-Doc	3.2/1.0	20.3/13.8	44.2/55.6	7.0/4.5	25.3/25.1
AIGC-DB	5.8/0.8	27.5/5.7	31.3/66.3	7.8/2.1	27.6/25.0
AIGI-P	9.8/3.2	26.2/7.8	28.8/63.1	8.5/1.0	26.7/25.0
AIGI-S	4.7/0.1	26.8/31.9	27.8/26.4	9.7/15.6	30.9/26.0
AIGIBench	4.9/1.1	29.3/22.5	29.4/41.8	9.7/9.2	26.7/25.4
Chameleon	3.6/1.0	29.6/30.0	21.9/29.2	13.6/13.8	31.4/25.9
Holmes	8.8/1.1	26.7/8.7	28.5/60.1	9.1/3.8	27.0/26.3
WildRF	7.2/1.3	27.6/9.1	31.2/61.3	8.7/3.3	25.3/25.0

Cross-group dispersion

Q	Min	Max	Δ	Min $\rightarrow$ max group
N	1.7	6.5	4.8	OM-Deepfake $\rightarrow$ AIGI-P
J	16.3	29.8	13.5	OM-Deepfake $\rightarrow$ Chameleon
B	25.6	56.1	30.5	Chameleon $\rightarrow$ OM-Deepfake
V	0.5	13.7	13.2	OM-Deepfake $\rightarrow$ Chameleon
F	25.2	28.6	3.4	OM-Doc $\rightarrow$ Chameleon

The decision-readout weights of guided queries vary with domain and label. The boundary-transition query receives greater weight on fake samples in most groups, whereas JPEG Ghost is often weighted more strongly on real samples. Across label-balanced group means, boundary-transition weights have the largest range, while free-query weights have the smallest. These patterns show that the decision readout adjusts the mixture of retrieved evidence across test distributions.

H QUERY-ATTENTION VISUALIZATIONS

Using the same trained CLIP model as the right panel of Figure 4, we visualize one real and eight fake images. The overlays locate where each query reads and reveal distinct spatial coverage across queries.

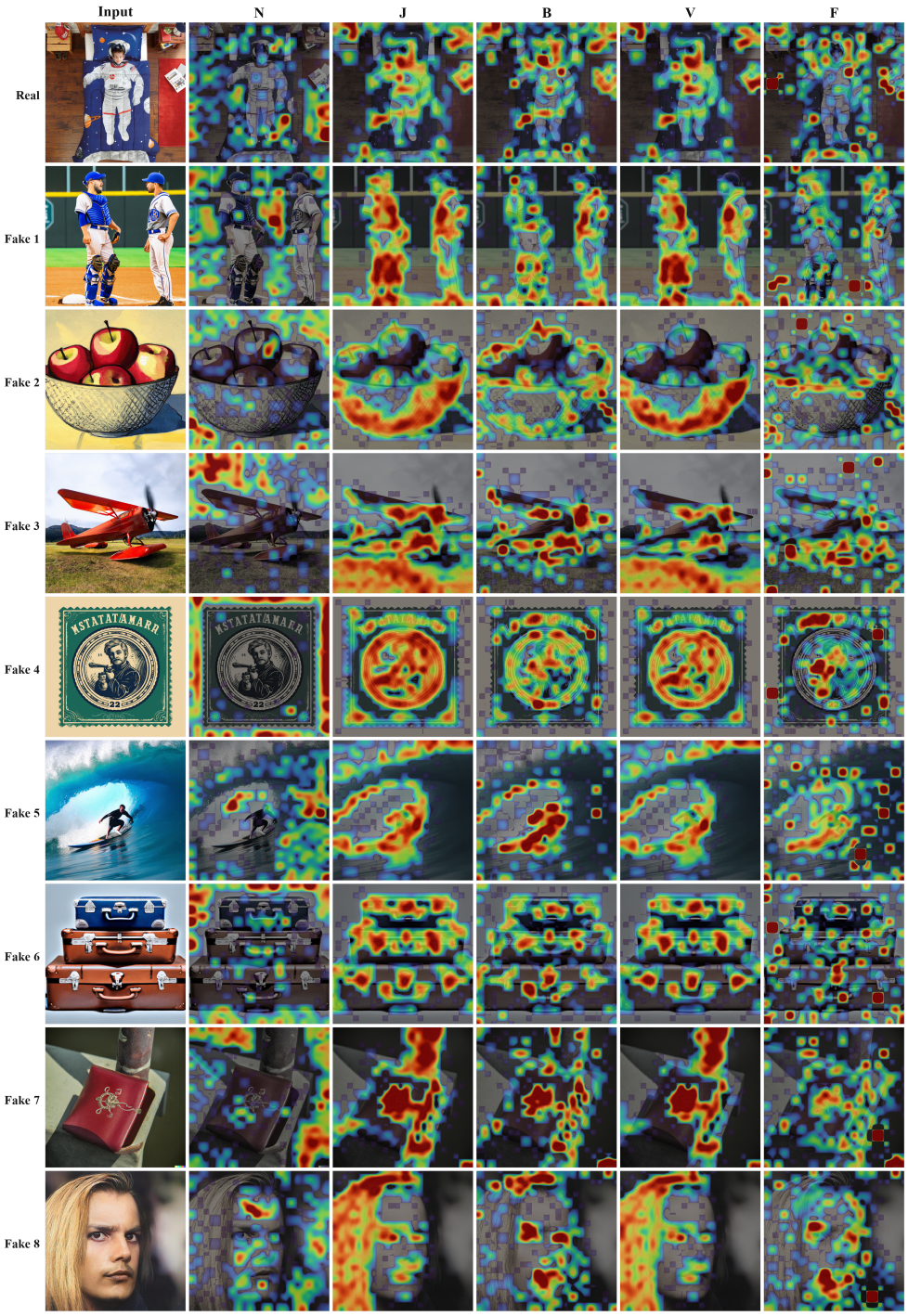

Figure 9: **Additional query-attention visualizations on one real and eight fake images.**