PART-LEVEL SELF-BLENDING: REGION GROUND TRUTH BY CONSTRUCTION FOR DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deepfake detectors produce a single real-or-fake score and therefore cannot indicate which part of a face was manipulated. We show that the score itself misses such edits. Each of eighteen published detectors is at or near chance when a manipulation is confined to a small facial region and at chance on a single eye (worse-eye AUC 0.48–0.55). We attribute the gap to the training signal, since self-blending, the most common method of synthesizing training fakes, blends the whole face and therefore provides no information about *where* the manipulation lies. We introduce **Part-Level Self-Blending (P-SBI)**, which confines each synthetic blend to one named facial region, namely a single part, such as the left eye, or a small group of parts, such as the mouth, drawn from a fixed set of twelve that together cover every part of the face. Because the region is chosen at synthesis time, its mask is exact, and the same construction both trains a per-region prediction head and measures the region sensitivity of any detector. The verdict and the localization map need no anatomical input at test time, and only the per-region scores are read on a parser's regions. With a DINOv3-H+ backbone adapted on under one percent of its weights, our detector, PARTAWARE, acquires a region sensitivity none of the eighteen has (coverage **0.847** over three seeds, against 0.587–0.676 for all eighteen) at no cost to the properties the field already measures, reaching cross-dataset transfer **0.980** AUC and robustness to twenty-one corruption conditions **0.960** in distribution and 0.931 zero-shot, each at or above the strongest published detector. Removing P-SBI from the same backbone isolates the effect. Over three seeds, transfer falls only to 0.969 while coverage falls from 0.847 to 0.657, back into the published band. The lead also holds on diffusion-inpainted edits from two independent generators, whose region truth does not come from our construction.

1 INTRODUCTION

Deepfakes are images and videos in which a face has been synthetically altered, whether swapped for another identity, re-enacted to change its expression, or edited locally so that only the mouth or eyes change (Rössler et al., 2019; Tolosana et al., 2021). Such edits are now inexpensive to produce, difficult to identify by eye and widespread online (Chandra et al., 2026), so platforms and investigators rely on automatic *deepfake detectors*. AI-generated video of public figures now appears in fraud schemes (Federal Bureau of Investigation, Internet Crime Complaint Center, 2026; Guo, 2026), and detectors trained on a few generators are known not to generalize to new content (Hampson, 2026). A detector is typically a neural network trained on labeled real and fake faces to output one real-or-fake score per face, judged by how well that score separates real from fake faces (AUC, where 0.50 is chance and 1.00 perfect). By this measure the field has progressed rapidly, as detectors built on large pretrained vision models now reach AUC above 0.95 on datasets they were not trained on (Yan et al., 2025; Cui et al., 2025; Yermakov et al., 2026b), termed *cross-dataset transfer*. A single score, however accurate, omits *where* in the face the evidence lies, which makes a decision verifiable and matters most for partial edits that alter one feature (Liu et al., 2025). Mainstream detectors do not report it, and none has been evaluated for it under a common protocol, because such an evaluation requires region ground truth that the field lacks. The most common source of training fakes, *self-blending*, warps a face slightly and blends it back onto itself, creating a *pseudo-fake* with a

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

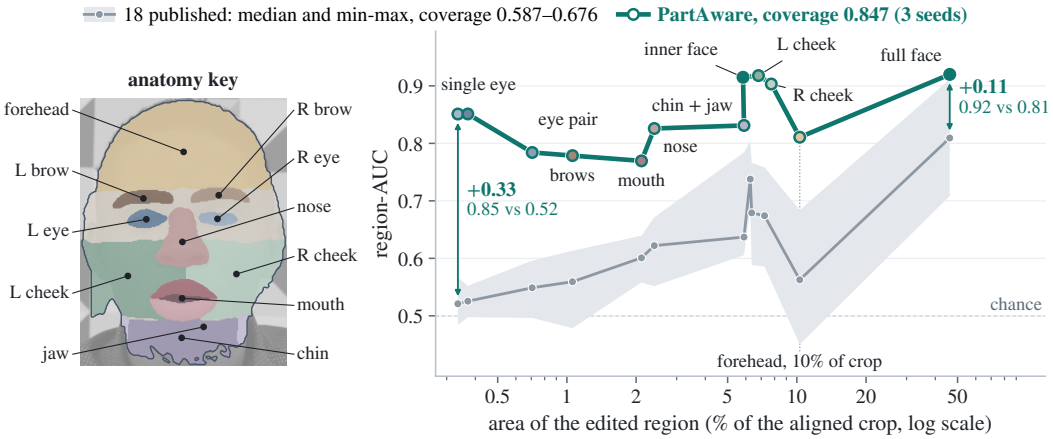

Figure 1: **Published detectors fail in proportion to how small the edited region is.** Region-AUC (0.50 is chance) against the area of the edited region, one point per P-SBI recipe. Gray marks the eighteen published detectors (median line, min–max band), teal PARTAWARE (mean of three seeds), each marker filled in its part's key color (teal for multi-part recipes). Region-AUC correlates with region area at Spearman $\rho = 0.83$ for the published median against 0.47 for PARTAWARE, and the arrows mark the gap on a single eye and on the full face. Size is not the only factor, since the forehead, at 10% of the crop, falls below regions smaller than it.

forgery boundary but no face generator (Shiohara & Yamasaki, 2022; Cheng et al., 2024). Because it blends the *whole* face, it teaches that a blending seam is suspicious but not where it lies. Methods that supply region information either need a face parser, a network that segments a face into parts, at test time (Uddin et al., 2026; Al-Zyoud & El Saddik, 2026) or rely on annotation (Sun et al., 2023), so their labels are only as reliable as the parser or annotator.

Three questions are therefore open.

RQ1 *Do today's deepfake detectors respond to a manipulation confined to one* ***region*** *of the face?*
RQ2 *Can a detector learn* ***where*** *to look without losing its ability to decide* ***whether****?*
RQ3 *What can teach a detector where to look?*

We address these questions with **Part-Level Self-Blending (P-SBI)**, which confines each synthetic blend to one named region of the face, namely a single part or a small group of parts listed by one of twelve fixed *recipes* that together cover all fourteen parsed parts, with a feathered boundary and one of three blending operators. Because the region is chosen when the blend is made, its *blend mask*, the pixels the blend alters, is *exact by construction*, so the same construction can both train a detector and test any detector. The anatomical *name* attached to that mask still comes from the parser, which Appendix I tests against independent annotations. Our contributions follow the three questions.

- **A MEASUREMENT OF REGION SENSITIVITY.** To answer RQ1, a threshold-free region-coverage protocol (Section 3.3) scores a detector by its AUC on each recipe's blends against clean frames (*region-AUC*) and their mean (*coverage*). All eighteen published detectors from 2017 to 2026, spanning five families, are at or near chance once an edit is confined to a small facial region, and at chance on a single eye (*worse-eye AUC*, the lower single-eye region-AUC, 0.48–0.55, all 18 below 0.56), and their sensitivity falls with the size of the edited region (Figure 1).
- **A DETECTOR THAT LEARNS WHERE WITHOUT LOSING WHETHER.** For RQ2, P-SBI with a DINOv3-H+ backbone (Siméoni et al., 2026) adapted on under one percent of its weights yields PARTAWARE, which predicts a verdict (*frame head*), a score for each part (*part head*) and a pixel-level localization map (*localization head*, Section 3). Read on the parser's regions, the part head identifies the edited region in 0.69 of cases and its side in 0.92. It gains region coverage none of the eighteen reaches (0.847, mean of three seeds, against 0.587–0.676) while matching the strongest of them on cross-dataset transfer (0.980 AUC) and on corrupted AUC over twenty-one conditions (0.960 *in distribution*, on the training dataset, and 0.931 *zero-shot*, on unseen datasets, Table 3), so the capability is added, not traded, and the verdict needs no anatomy at test time.

- **WHAT PRODUCES THE GAIN.** For RQ3, matched-capacity ablations attribute coverage to region confinement, not capacity or blending. A frozen backbone read without its pretraining normalization is barely above chance (0.585), adaptation alone adds little (0.636), whole-face blends stay region-insensitive (0.669), and P-SBI at the same capacity reaches 0.737. The same holds at DINOv3-L against an independent detector on that backbone (Yermakov et al., 2026b).

Section 6 examines these findings with controls and with diffusion-inpainted edits from two independent generators.

2 RELATED WORK

GENERALIZABLE DETECTION ANSWERS "WHETHER", NOT "WHERE". The common protocol in this line of work trains on FaceForensics++ (Rössler et al., 2019) and reports AUC on unseen datasets. CNN detectors (Chollet, 2017; Tan & Le, 2019; Qian et al., 2020; Liu et al., 2021; Luo et al., 2021; Ni et al., 2022; Cao et al., 2022; Yan et al., 2023a) established this transfer suite. Recent work adapts large pretrained backbones through low-rank updates (Yan et al., 2025), boundary supervision (Cui et al., 2025), LayerNorm-only tuning (Yermakov et al., 2026b), or a focus on partially manipulated faces (Yermakov et al., 2025). SFAM adds region experts and Cross-AUC (Luo et al., 2026), the AUC of real videos from one benchmark against fakes from another, which we also report (Appendix E.3). These methods decide *whether* a face was manipulated accurately, but none answers per region (RQ1). We re-run every one with a released checkpoint under one crop protocol (Appendix B) and, as SFAM has none, adopt only its metric.

PSEUDO-FAKE SYNTHESIS HAS REGION INFORMATION AND DISCARDS IT. Self-blended images (Shiohara & Yamasaki, 2022) and their descendants (Cheng et al., 2024) blend a face with a jittered copy of itself. The blend mask is exact at synthesis, but blending the whole face discards it. P-SBI is the part-level generalization, where confining the blend to one named region makes an exact region label exist (RQ2). Region-confined blending has precedents in adversarial landmark regions (Chen et al., 2022), random combinations of parsed organs (Xu et al., 2025), lip and lower-face blends (Shahzad et al., 2026), attention steered toward blending-prone regions (Nguyen et al., 2024b;a; 2026), and supervision on the exact blend boundary (Li et al., 2020a) or region (Zhao et al., 2021b). None labels a single region over a fixed anatomy and reuses that construction to evaluate detectors that never trained on it, which makes the region claim checkable against a known answer.

REGION-AWARE DETECTORS OBTAIN ANATOMY AT A COST. Segmentation-guided spatial indexing (Al-Zyoud & El Saddik, 2026) pairs DINOv3 tokens with a frozen face parser, and Face2Parts (Uddin et al., 2026) feeds region streams as network inputs. Both need anatomy at test time and take region ground truth from that same parser, so a region readout can only be checked against the parser, never against a known answer. Annotation-based supervision (VLFFD, FFTG (Sun et al., 2023; 2025)) has the same problem at coarser granularity, with an annotator in the parser's place. Labels exact by construction remove both dependencies.

LOCALIZED-MANIPULATION BENCHMARKS SHOW THE FAILURE BUT NOT ITS LOCATION. FakeParts (Liu et al., 2025) and DF40 (Yan et al., 2024) show that detectors degrade on partial manipulations, at video and manipulation-type granularity, and DiffFace-Edit (Ding et al., 2026) and in-the-wild evaluations (Chandra et al., 2026) report aggregate collapse without locating it. We add the per-region measurement. We attribute each score to facial regions with Shapley values (Shapley, 1953; Lundberg & Lee, 2017). The novelty is applying them where region truth is known, so attribution is checked for accuracy rather than plausibility. Appendix C discusses each line of work.

3 METHOD

P-SBI answers RQ2 and the region-coverage protocol built from it (Section 3.3) answers RQ1. Four principles shape both (Appendix D.1), namely region labels exact by *construction* rather than inferred, anatomy permitted in training but not required for the verdict at test time, training unsolvable by a shortcut such as a hard seam, and threshold-free evaluation, since detectors calibrate scores differently.

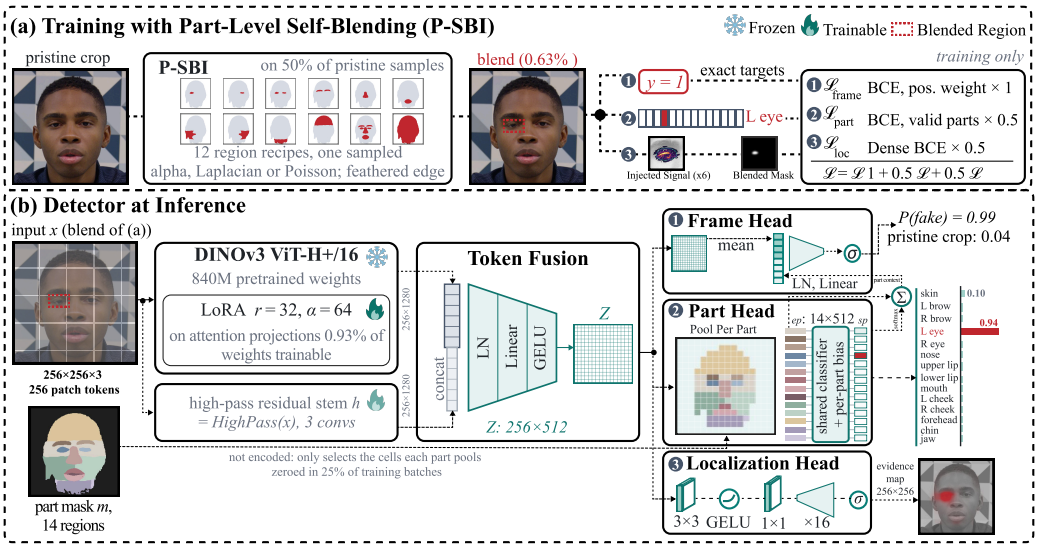

Figure 2: **Training with part-level self-blending and the detector at inference.** (a) During training, P-SBI confines a self-blend to one of twelve facial regions, here the left eye (0.63% of the crop), and records three exact targets, each supervising the head with the same number in (b). (b) At inference, three heads read the fused tokens, and the part mask only selects the cells each part pools. On this held-out DeepFakeDetection (DFD) frame the detector outputs $P(\text{fake}) = 0.99$ (0.04 on the pristine crop) and scores the left eye 0.94, with every other part at most 0.10.

3.1 Architecture

The input is an aligned 256×256 RGB face crop x, from which we take a high-frequency residual $h = \text{HighPass}(x; \text{cutoff} = 0.12)$ and a part mask $m \in \{0, \ldots, 14\}^{256\times256}$ from the face parser, labeling fourteen visible parts (Figure 2). A frozen pretrained transformer encodes the RGB view, using DINOv3-H+ (Siméoni et al., 2026) for PartAware and CLIP ViT-L/14 (Radford et al., 2021), Perception Encoders (PE) (Bolya et al., 2025) and SigLIP2-B (Tschannen et al., 2025) for the backbone comparison. Only a LoRA update (Hu et al., 2022), a trainable low-rank increment to its frozen attention projections, is trained (rank 32, $\alpha = 64$ for PartAware), so corruption tolerance comes from the frozen weights (Section 5.3). A small convolutional stem, the *frequency stem*, encodes the residual, and its tokens are concatenated with the RGB tokens and linearly fused. The stem helps smaller backbones and is neutral at the reported scale (Appendix Table F.1).

Three heads read the fused tokens. A **frame head** emits one real-or-fake logit, a **part head** scores the fourteen parts from the tokens each pools with one shared linear classifier plus a per-part bias, and a **localization head** decodes a dense localization map. The part mask is never encoded (it only selects each part's tokens) and is zeroed on a random 25% of training batches, preventing dependence on it. Scoring with it zeroed is the *mask-free* mode. A softmax over the part scores weights the part embeddings into the frame head's *part context*. Training minimizes $\mathcal{L} = \text{BCE}_{\text{frame}} + 0.5\,\mathcal{L}_{\text{part}} + 0.5\,\mathcal{L}_{\text{loc}}$. Appendix D.4 gives class weighting, localization targets and hyperparameters.

3.2 Part-level self-blending

Whole-face self-blending (Shiohara & Yamasaki, 2022) composites a face with a jittered copy of itself, whose mask spans many regions and cannot supervise *which* region was manipulated. P-SBI confines each blend to one named target. From a pristine face x and its part mask m we sample one of twelve *recipes*, namely six single regions, four groups (eye pair, brows, mouth, chin and jaw) and two large-region controls, the inner face and the full face (Appendix Figure D.2). The recipes cover every part. The part-head target marks every part of the sampled recipe. Single-side recipes are oversampled, drawn from those four with probability 0.35 and otherwise uniformly from all twelve.

The donor patch, by default a mildly warped and photometrically jittered copy of x, is composited through the recipe's blend mask by alpha, Laplacian or Poisson compositing with a Gaussian-feathered boundary, with feather widths and the rejection rule in Appendix D.4. Recording the recipe at synthesis makes the mask exact, its anatomical name inheriting the parser's segmentation of that frame (Appendix Figure D.1). Edits are small. A single-eye blend covers under one percent of the crop and injects about forty times less residual energy than a whole-face blend, almost none above 0.25 cycles per pixel (Appendix Figure D.3), so neither global image statistics nor a visible seam is an obvious shortcut. P-SBI converts a random 50% of pristine training samples into pseudo-fakes, labeled fake like the native FaceForensics++ forgeries and about 11% of the fake-labeled stream.

3.3 Region-coverage evaluation protocol

The protocol answers RQ1 for any detector mapping an image to a score. We blend each of $n = 200$ pristine frames from the held-out test split with all twelve recipes. For recipe c the **region-AUC**:

$$\mathrm{AUC}_c = \Pr\big[s^{\text{blend}} > s^{\text{clean}}\big] + \tfrac{1}{2}\Pr\big[s^{\text{blend}} = s^{\text{clean}}\big] \tag{1}$$

is the threshold-free probability that a frame with region c edited scores above a clean frame, estimated against all n clean frames pooled (the *pooled estimator*). It measures whether the score *responds* to the edit, while *naming* the region is left to the part head (Table 2). **Coverage** is the mean over the twelve recipes (grouped recipes score their parts as a group), reported with the worst region, as a high mean can conceal insensitivity to one. Post-hoc Shapley attribution (Shapley, 1953; Lundberg & Lee, 2017) credits each of the fourteen parts with its average score change, over sampled orders, when its pristine pixels are restored (Appendix L.2 gives the estimator), yielding how often the top-attributed part lies in the blended region (**top-1**) and, for left/right pairs, whether the side is right (**side accuracy**). Both are consistency checks only. Since our models train on this blend family, we add a **no-P-SBI control** (same backbone, capacity and adaptation, frame labels only), a **whole-face control** trained only on the two large-region recipes, and a **held-out-operator control** trained without Poisson compositing and evaluated on Poisson blends alone (Section 6).

3.4 Robustness evaluation protocol

Seven corruptions (lower saturation and contrast, Gaussian blur, additive Gaussian noise, pixelation, JPEG recompression, random block occlusion) are applied at three severities to the aligned transfer crops, seeded per (frame, corruption, severity) so every detector sees byte-identical corrupted pixels. Our models are scored mask-free, with the residual recomputed from the corrupted image. We report the mean AUC over the twenty-one conditions (*corrupted AUC*) and its drop from clean AUC.

4 Experimental Setup

Each choice below makes a difference between two table rows reflect the detectors, not the pipelines that produced their inputs. Appendices B and E.1 give the protocol and baseline provenance.

Datasets. We train on the official FaceForensics++ (FF++) c23 split (Rössler et al., 2019) (c23 is its lightly compressed release), as do most baselines, whereas the LAA-Net, FakeFormer and GenD releases differ in compression or training data, so those rows are not matched on it (Appendix B). Zero-shot transfer uses Celeb-DF-v2 (Li et al., 2020b), DFD (Dufour & Gully, 2019), the DFDC Preview test set (DFDC-P) (Dolhansky et al., 2020) and UADFV (Yang et al., 2019), the four benchmarks most commonly reported, which together span face swapping and reenactment, studio and in-the-wild capture, and two generations of generator. Corruption robustness uses all five (Section 3.4). Deepfake-Eval-2024 (Chandra et al., 2026) is scored once and informed no decision (Appendix J). Appendix E.1 lists split sizes and video counts.

Protocol. Face detection, alignment and crop margin can alone move cross-dataset AUC by several points, so every detector, ours and all eighteen baselines, is scored on *identical* InsightFace-aligned 256×256 crops (ArcFace (Deng et al., 2019) five-point template, margin 1.3), the convention of the DeepfakeBench release (Yan et al., 2023b) that supplies our CNN baselines. Every baseline value in this paper is thus our own measurement of that detector's released checkpoint under this protocol, and none is copied from a source paper. Appendix Table B.1 places each published value beside our measurement, including the small, disclosed penalty the shared frontend costs some

Table 1: Cross-dataset transfer, video AUC (higher is better). Every value, each baseline's included, is our own measurement of the released checkpoint on the identical aligned crops of Section 4. None is copied from a source paper (Appendix Table B.1 compares the two). Rows grouped by family, best in bold. Ref. cites each detector's paper and the dash marks this paper. Appendix Table A.1 extends the comparison to all nineteen detectors. ‡ marks training beyond FaceForensics++ c23.

Detector	Family	Ref.	FF++↑	Mean↑	Celeb-DF-v2↑	DFD↑	DFDC-P↑	UADFV↑
Xception	CNN	(Chollet, 2017)	<u>0.9881</u>	0.8784	0.8139	0.9618	0.7629	0.9751
ProDet‡	Blend	(Cheng et al., 2024)	0.9775	0.9208	0.9166	0.9450	0.8342	0.9876
FSFM	SSL	(Wang et al., 2025)	0.9616	0.8780	0.8648	0.8305	0.8744	0.9421
Effort	Adapt	(Yan et al., 2025)	0.9491	0.9570	0.9237	0.9987	0.9073	0.9984
ForensicsAdapter	Adapt	(Cui et al., 2025)	0.9675	<u>0.9717</u>	0.9494	0.9988	<u>0.9408</u>	0.9978
Unlocking-CLIP	Adapt	(Yermakov et al., 2025)	0.9767	0.9641	0.9613	0.9998	0.8952	**1.0000**
GenD CLIP-L‡	Adapt	(Yermakov et al., 2026b)	0.9777	0.9691	0.9575	0.9993	0.9195	**1.0000**
GenD DINOv3-L‡	Adapt	(Yermakov et al., 2026b)	0.9877	0.9567	0.9190	<u>0.9999</u>	0.9135	0.9946
GenD PE-L‡	Adapt	(Yermakov et al., 2026b)	0.9864	0.9618	<u>0.9628</u>	**1.0000**	0.8875	0.9968
PARTAWARE (ours)	**Adapt**	–	**0.9923**	**0.9797**	**0.9661**	<u>0.9999</u>	**0.9528**	**1.0000**

baselines. Frame scores are averaged within a face track, and a video takes the maximum over its tracks for the video-level AUC of transfer and robustness (coverage is frame-level, Appendix E.1).

BASELINES. Eighteen published detectors from 2017 to 2026, run from released weights so that each is compared as its authors published it, and grouped in every table by the cue each is designed to exploit, namely *spatial CNNs* (Xception (Chollet, 2017), EfficientNet-B4 (Tan & Le, 2019), CORE (Ni et al., 2022), RECCE (Cao et al., 2022), UCF (Yan et al., 2023a)), *frequency CNNs* with a spectral stream (F3Net (Qian et al., 2020), SPSL (Liu et al., 2021), SRM (Luo et al., 2021)), *whole-face blending*, supervised on whole-face pseudo-fakes that P-SBI generalizes (LAA-Net (Nguyen et al., 2024b), FakeFormer (Nguyen et al., 2024a), ProDet (Cheng et al., 2024)), *face SSL*, face-specific self-supervised pretraining (FSFM (Wang et al., 2025)), and *frozen foundation-model adaptation*, our own architecture class and the comparison that matters most, since it holds the strongest transfer baselines under a comparable adaptation budget (Effort (Yan et al., 2025), ForensicsAdapter (Cui et al., 2025), Unlocking-CLIP (Yermakov et al., 2025), and GenD on CLIP-L, DINOv3-L and a Perception Encoder (Yermakov et al., 2026b)). Tables abbreviate the five families as CNN, Freq, Blend, SSL and Adapt, and Appendix B records each checkpoint's source.

OUR CONFIGURATIONS. Headline comparisons use PARTAWARE, except for the matched-backbone pair and the mask-free variant, each marked. The configurations and controls trained for the ablation, with their evaluations, are in Appendix Table F.2. Coverage and robustness cover all eighteen published detectors, coverage also every configuration of ours (Appendix F.1).

5 RESULTS

PARTAWARE adapts a DINOv3-H+ backbone with rank-32 LoRA and is trained with P-SBI. Tables report a single trained model, and the text also gives the mean over three training runs. Per-detector tables are in Appendix A. In every table the best value is bold and the second best underlined.

5.1 CROSS-DATASET TRANSFER

Learning *where* does not cost transfer (RQ2, first half). In Table 1 PARTAWARE reaches 0.9797 mean transfer AUC against 0.9717 for ForensicsAdapter, the second-best detector. Over three seeds it is 0.980 ± 0.001, and the paired bootstrap puts the difference at 0.008 (95% CI 0.002–0.015). The margin is small by design. The claim is that a region-level training signal can be added without giving up cross-dataset accuracy. Pretrained weights alone do not explain the result, since our DINOv3-L configuration exceeds GenD DINOv3-L, an independent detector on the same backbone, by 1.3 points (Appendix Table F.3). In-distribution AUC does not predict the ranking. Across the eighteen published detectors it correlates negatively with transfer ($\rho = -0.49$, Appendix L). PARTAWARE is also highest under Cross-AUC (Luo et al., 2026) (Appendix E.3). The fidelity of the reproduced

Table 2: Region coverage under the protocol of Section 4, measured by us for every detector as mean region-AUC over the twelve recipes (frame level, n=200 per recipe, 95% CI half-width ≤ 0.017), the worst single region, and how often post-hoc Shapley attribution names the edited region (Top-1) and its side (Side acc.). The first row is a parameter-free floor with no anatomy, so the attribution columns are unranked (Appendix L.2 and Appendix Table L.1). Chance is ≈ 0.50. Best coverage and worst region are in bold. Appendix Table A.2 covers all detectors, Appendix Table H.1 gives per-region values, and Table 1 cites each detector. ‡ marks training beyond FaceForensics++ c23.

Detector	Family	Backbone	Coverage↑	Worst region↑	Top-1↑	Side acc.↑
Floor: parameter-free statistic (Appendix L.2)			0.528	0.493 (right eye)	0.706	0.828
ProDet‡	Blend	EffNet-B4	0.644	0.508 (left eye)	0.665	0.724
FSFM	SSL	ViT-B	0.587	0.511 (right eye)	0.684	0.798
Effort	Adapt	CLIP-L	0.646	0.552 (left eye)	0.807	0.851
ForensicsAdapter	Adapt	CLIP-L	0.637	0.529 (left eye)	0.781	0.854
Unlocking-CLIP	Adapt	CLIP-L	0.655	0.528 (left eye)	0.779	0.837
GenD CLIP-L‡	Adapt	CLIP-L	0.654	0.532 (left eye)	0.780	0.840
GenD DINOv3-L‡	Adapt	DINOv3-L	0.676	0.531 (left eye)	0.771	0.818
GenD PE-L‡	Adapt	PE-L	0.660	0.512 (each eye)	0.775	0.808
PARTAWARE (ours)	**Adapt**	DINOv3-H+ (r32)	**0.795**	**0.684** (brows)	0.857	0.925

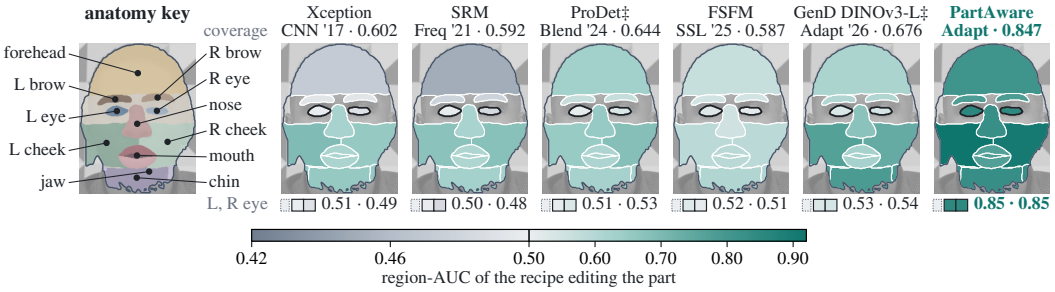

Figure 3: **Where each detector can see a manipulation.** Each part colored by its recipe's region-AUC (centered on chance), for one published detector per family and PARTAWARE (mean of three seeds). Appendix Figure H.1 shows all eighteen. ‡ marks training beyond FaceForensics++ c23.

baselines is examined in Appendix B. These four benchmarks guided our backbone selection and not the baselines'. Appendix J reports one that informed no decision of ours.

5.2 REGION COVERAGE

Table 2 answers RQ1, whether a score responds to the edited region, not only whether it separates real from fake. Its main threat is the P-SBI construction it uses, which our models train on and the published detectors do not (Section 6). All eighteen published detectors fall between 0.587 and 0.676, the eight spatial and frequency CNNs inside that band without widening it, so region insensitivity holds from 2017 to 2026 (Appendix Table A.2). Among the ten modern detectors the failure grows as the region shrinks (Figures 1 and 3 and Appendix Table H.1). Full-face blends separate at 0.75–0.91, cheeks, mouth and nose degrade, and single eyes reach chance, 10 of 10 below 0.56 on the left eye and 9 of 10 on the right. Attribution top-1 is not evidence of localization, since a parameter-free statistic reaches 0.71 top-1 and 0.83 side accuracy (Appendix L.2). Region naming rests on the part head, which, read on the parser's regions, identifies the edited region in 0.693 of cases against 0.318 for the whole-face control, at side accuracy 0.920 against 0.596 (Appendix Table L.1). PARTAWARE reaches 0.795 (0.847 $\pm$ 0.045 over three seeds), 11.9 points above GenD DINOv3-L, the second-best detector (paired gap 0.119, frame-bootstrap CI 0.102–0.135, Appendix E.2), raising the small lateral regions from $\approx$ 0.52 to 0.80 and detecting 57–58% of single-eye blends at a 10% false-positive rate, where no published detector exceeds 19% on either eye (Appendix Figure H.3). On diffusion-inpainted edits with region truth independent of our parser (Appendix I) it reaches 0.762 against 0.694, the only detector above 0.60 on either single eye, and still leads all eighteen

Table 3: Corruption robustness, video AUC over the 21 conditions of Section 3.4 (7 corruptions × 3 severities). *In distribution* is corrupted FaceForensics++, and *zero-shot* averages the four unseen benchmarks. *Drop* is the fall from clean, lower being better. Corrupted crops are byte-identical across detectors and PARTAWARE is scored mask-free. Rows match Table 1. Appendix Table A.4 covers all nineteen detectors and all five benchmarks. ‡ marks training beyond FaceForensics++ c23.

		In distribution		Zero-shot	
Detector	**Family**	**Corrupted**↑	**Drop**↓	**Corrupted**↑	**Drop**↓
Xception	CNN	0.8873	0.101	0.7807	0.098
ProDet‡	Blend	0.8476	0.130	0.8180	0.103
FSFM	SSL	0.9080	0.054	0.8214	0.057
Effort	Adapt	0.8773	0.072	0.8782	0.079
ForensicsAdapter	Adapt	0.8981	0.069	0.9079	0.064
Unlocking-CLIP	Adapt	0.9201	0.057	0.9073	0.057
GenD CLIP-L‡	Adapt	0.9178	0.060	0.9101	0.059
GenD DINOv3-L‡	Adapt	0.9469	0.041	0.9105	0.046
GenD PE-L‡	Adapt	0.9413	0.045	0.9066	0.055
PARTAWARE (ours)	**Adapt**	**0.9604**	0.032	**0.9311**	0.050

at JPEG quality 75 (0.751 against at most 0.637, Appendix Table A.3). Blending alone does not create the sensitivity. ProDet, trained on whole-face self-blends (Cheng et al., 2024), scores 0.644, our whole-face control 0.669, and GenD DINOv3-L, the same backbone without P-SBI, 0.676 (the controlled swap is in Section 5.4). Anatomy is a training scaffold, not what the protocol rewards. Mask-free scoring gives 0.791 coverage and moves transfer by +0.001 (Appendix E.1).

5.3 CORRUPTION ROBUSTNESS

RQ2 also asks whether the decision survives the compression, resizing and noise of real uploads, exactly where a model reading a high-frequency residual should fail. Because corrupted FaceForensics++ tests the *training* distribution, we also report the mean over the four unseen benchmarks, where the CNN families fall about fivefold further behind frozen foundation-model adaptation than they do in distribution (Appendix G). PARTAWARE is highest on both, at 0.9604 in distribution with the smallest drop (0.032 against ForensicsAdapter's 0.069) and 0.9311 zero-shot (Table 3). Noise separates the families most sharply. Removing the frequency stem leaves corrupted AUC at 0.9603 against 0.9604 (Appendix Table F.1), so robustness tracks the backbone, not the stem added to protect it. Per benchmark (Appendix Table A.4) PARTAWARE leads on FaceForensics++, Celeb-DF-v2 and DFDC-P, while GenD DINOv3-L and GenD CLIP-L exceed it by under 0.002 on DFD and UADFV.

5.4 ABLATION STUDY

REGION CONFINEMENT, NOT CAPACITY OR BLENDING, PRODUCES COVERAGE. RQ3 asks what produces the coverage gain. In Appendix Table F.2 a frozen backbone, read without its pretraining normalization, is barely above chance (0.585) and low-rank adaptation alone adds little (0.636). Adding P-SBI at the *same* capacity is the largest single step (0.636 → 0.737), already above every published detector with an 89 M backbone, and scaling the backbone raises both axes (0.979 transfer, 0.784 coverage at DINOv3-H+), so capacity amplifies the training signal rather than replacing it. At fixed backbone, removing P-SBI returns coverage to the published band (0.657 against 0.847 over three seeds) while transfer and in-distribution corrupted AUC stay within a point (0.969 against 0.980, 0.964 against 0.962, Appendix Table F.3). Zero-shot corrupted AUC does fall, by 0.023 on Celeb-DF-v2 (Appendix Table F.4), so the training signal helps there rather than costing. The whole-face control separates confinement from blending, with 0.669 coverage, inside the published band, at 0.924 transfer. On the backbones of published detectors, our pipeline reaches 0.762–0.821 coverage against 0.646–0.676 for those detectors (Appendix Table F.3). The frequency stem is not responsible either. At the PARTAWARE configuration it adds no transfer (0.979 against 0.980) and no robustness in distribution or zero-shot (Appendix Table F.1), and coverage without it is *higher* (0.853 against 0.795), inside the seed spread of 0.847 ± 0.045.

CAPACITY DOES NOT PREDICT TRANSFER. Across four backbone families at two scales under one configuration (Appendix Table F.2), the largest backbone we train, $2.2\times$ PARTAWARE, has the highest validation AUC of any configuration yet the weakest transfer of its scale, so in-distribution selection would have picked our worst-transferring model. No design alternative at fixed capacity outperformed fixed recipe sampling at 256 px (Appendix F.3). Single-seed coverage differences below the seed spread of 0.045 are not interpreted. The three seeds of PARTAWARE and of the rank-16 configuration, and the choice of LoRA rank, are in Appendix F.4 and Appendix Table F.4.

6 DISCUSSION AND LIMITATIONS

EXACT REGION LABELS MAKE THE CLAIM CHECKABLE. A detector that returns only a probability can be believed or doubted, but the evidence behind it cannot be examined. The field has lacked the ground truth rather than overlooked the problem. Whole-face self-blending discards the exact mask it records, boundary supervision marks a seam without naming the region, and a parser-based readout can only be compared against that parser (Section 2). Confining each blend to one named region keeps the label synthesis already produced, so *where* becomes a quantity scored for correctness, and the same construction measures detectors that never trained on it. We therefore report the model's own part head at 0.693 rather than the post-hoc 0.857, because a parameter-free statistic already reaches 0.706 under this protocol (Appendix Table L.1).

WHAT THE COVERAGE MEASUREMENT CAN BEAR. The protocol builds its test edits with the same construction as P-SBI training, so our models have seen this kind of edit and the baselines have not. Three checks address this. The no-P-SBI and whole-face controls at matched capacity show the gain is genuine. The held-out-operator control shows that at the PARTAWARE configuration coverage does not depend on the compositing operators seen in training, which rules out one shortcut, not every shortcut (Appendix D.4). On diffusion-inpainted edits from two independent generators our lead and the single-eye gap persist at a smaller margin (0.068, Appendix I). Coverage varies by ± 0.045 over three seeds (Appendix Table F.4), below the 0.119 gap to the best published detector, so single-run ablation differences under about 0.05 are not interpreted.

REGION SENSITIVITY HAS NOT YET BECOME IN-THE-WILD ACCURACY. On in-the-wild footage that informed no design decision, and on real lipsync forgeries, every detector we test remains far from deployable accuracy, ours included, and a cross-frame mouth recipe that lifts synthetic mouth coverage does not help (Appendices J and K). Knowing *where* to look differs from recognizing what an unseen generator leaves there. Appendix E.1 gives the evaluation scope and model cost.

REGION SENSITIVITY COMES FROM THE TRAINING SIGNAL, NOT THE ARCHITECTURE. Removing P-SBI at the reported scale costs 0.19 of coverage over three seeds, whereas removing the frequency stem leaves transfer and robustness unchanged, and coverage is higher without it, inside the seed spread (Appendix Table F.1). The stem's contribution decays with backbone capacity, adding coverage at DINOv3-B and DINOv3-L but not at the reported scale (Appendix F.1).

7 CONCLUSION

We asked whether deepfake detectors respond to a manipulation confined to one region of the face, whether they can learn to, and what produces that ability. They do not. Every one of the eighteen published detectors we test is at or near chance on edits confined to a small facial region (RQ1). The ability can be learned. Confining each blend to one named region makes its label exact by construction, and a detector trained on such blends reaches 0.847 mean coverage against 0.587–0.676 for those eighteen, while matching the strongest on transfer and robustness (RQ2). Confinement, not capacity, blending or the frequency stem, produces it. Removing P-SBI leaves transfer and robustness within a point while coverage falls into the published band (RQ3). Region sensitivity and generalization are therefore separable, and should be measured separately. The gain holds beyond our own blends, on diffusion-inpainted edits from two independent generators (0.762 against 0.694) and under JPEG recompression (0.751 against at most 0.637). We release the coverage protocol so that future detectors report region sensitivity alongside transfer and robustness.

Reproducibility Statement

Data and protocol. Training uses the official FaceForensics++ c23 splits with a fixed manifest, and each benchmark follows its published test protocol (Section 4). The data processing steps, namely face detection, alignment, part masks and video-level aggregation, are described in Appendix E.1, and the heads, objective and optimization settings in Appendix D.4. All detectors, ours and every baseline, are scored on byte-identical aligned crops from one preprocessing stack, and corruption randomness is seeded per (frame, perturbation, severity), so every comparison is bit-reproducible.

Numbers. Every reported value traces to a stored evaluation artifact, and before submission, every cell of the transfer, coverage, attribution, robustness, seed and ablation tables was re-derived from them by a checking script. Every bootstrap interval is computed from the stored per-video and per-frame scores. The appendix gives full per-model tables at frame and video level. Seed variability of the PartAware configuration is reported over three seeds (Appendix Table F.4). Baseline numbers are our re-evaluations, reconciled with the published ones in Appendix Table B.1.

Release. The supplementary material contains an anonymized repository with the P-SBI synthesis code (Section 3.2), including all twelve recipes and three compositing operators, and the region-coverage protocol of Section 3.3, namely the pooled estimator, its frame bootstrap and a harness that measures the coverage of any detector from its score function and aligned crops with part masks. For the same crops, masks and seed, the released P-SBI code produces the same blended images as our evaluation code. It requires only NumPy and OpenCV, and an included example runs the protocol without dataset imagery, which is not redistributed. The repository is also available at `https://anonymous.4open.science/r/Partaware-85FE/`, and the full codebase will be released upon acceptance.

Ethics Statement

This work concerns the detection of manipulated face imagery, a defensive capability. P-SBI blends a face with itself, producing low-fidelity pseudo-forgeries that add no manipulation ability beyond ordinary image editors. All datasets are established research benchmarks used under their own licenses and consent frameworks; no new data of identifiable individuals was collected.

AI Use Statement

A large language model coding assistant, Claude 5.1 Fable, was used to revise text and analyze the scripts that generate the tables.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL `https://proceedings.neurips.cc/paper/8160-sanity-checks-for-saliency-maps`.

Izaldein Al-Zyoud and Abdulmotaleb El Saddik. Segmentation-guided spatial indexing for generalizable and explainable deepfake detection, 2026.

Sarah Barrington, Maty Bohacek, and Hany Farid. The DeepSpeak dataset. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR Findings)*, 2026.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Cao_End-to-End_Reconstruction-Classification_Learning_for_Face_Forgery_Detection_CVPR_2022_paper.html.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision (ECCV)*, 2020. doi: 10.1007/978-3-030-58574-7_7.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, Sejin Paik, Aerin Kim, and Oren Etzioni. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/APAI/html/Chandra_Deepfake-Eval-2024_A_Multi-Modal_In-the-Wild_Benchmark_of_Deepfakes_Circulated_in_2024_CVPRW_2026_paper.html.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Self-Supervised_Learning_of_Adversarial_Example_Towards_Good_Generalizations_for_Deepfake_CVPR_2022_paper.html.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/2718a032d15e0b80cd164b240220df89-Abstract-Conference.html. Released implementation referred to as ProDet.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cui_Forensics_Adapter_Adapting_CLIP_for_Generalizable_Face_Forgery_Detection_CVPR_2025_paper.html.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Jingyi Deng, Chenhao Lin, Pengbin Hu, Chao Shen, Qian Wang, Qi Li, and Qiming Li. Towards benchmarking and evaluating deepfake detection. *IEEE Transactions on Dependable and Secure Computing*, 21(6):5112–5127, 2024. doi: 10.1109/TDSC.2024.3369711.

Feng Ding, Wenhui Yi, Xinan He, Mengyao Xiao, Jianfeng Xu, and Jianqiang Du. DiffFace-Edit: A diffusion-based facial dataset for forgery-semantic driven deepfake detection analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14117–14121, 2026. doi: 10.1109/ICASSP55912.2026.11462670.

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

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Dong_Implicit_Identity_Leakage_The_Stumbling_Block_to_Improving_Deepfake_Detection_CVPR_2023_paper.html.

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

Federal Bureau of Investigation, Internet Crime Complaint Center. FBI warns of scammers impersonating the IC3. Public Service Announcement I-072026-PSA, 20 July 2026, 2026. URL https://www.ic3.gov/PSA/2026/PSA260720. Accessed 19 September 2026.

Eileen Guo. Weaponized deepfakes. MIT Technology Review, 21 April 2026, 2026. URL https://www.technologyreview.com/2026/04/21/1135652/weaponized-deepfakes-ai-artificial-intelligence/. Accessed 19 September 2026.

Michelle Hampson. Deepfake detection dataset aims to keep up with generative AI. IEEE Spectrum, 3 May 2026, 2026. URL https://spectrum.ieee.org/deepfake-detector-microsoft-generative-ai. Accessed 19 September 2026.

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

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract-round2.html.

Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. SeeABLE: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Larue_SeeABLE_Soft_Discrepancies_and_Bounded_Contrastive_Learning_for_Exposing_Deepfakes_ICCV_2023_paper.html.

Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. FreqBlender: Enhancing deepfake detection by blending frequency knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4f92d2f498b88f1bd43732312272967a-Paper-Conference.pdf.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html.

Xinghan Li, Junhao Xu, and Jingjing Chen. VIGIL: Part-grounded structured reasoning for generalizable deepfake detection, 2026.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Liu_Spatial-Phase_Shallow_Learning_Rethinking_Face_Forgery_Detection_in_Frequency_Domain_CVPR_2021_paper.html.

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, Ziyou Liang, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a5a5b0ff87c59172a13342d428b1e033-Abstract-Conference.html.

Ziyi Liu, Firas Gabetni, Awais Hussain Sani, Xi Wang, Soobash Daiboo, Gaetan Brison, Gianni Franchi, and Vicky Kalogeiton. FakeParts: a new family of AI-generated deepfakes, 2025.

Zijie Lou, Gang Cao, Kun Guo, Lifang Yu, and Shaowei Weng. Exploring multi-view pixel contrast for general and robust image forgery localization. *IEEE Transactions on Information Forensics and Security*, 20:2329–2341, 2025. doi: 10.1109/TIFS.2025.3541957.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Generalizing_Face_Forgery_Detection_With_High-Frequency_Features_CVPR_2021_paper.html.

Yuhan Luo, Tao Chen, and Decheng Liu. Rethinking cross-domain evaluation for face forgery detection with semantic fine-grained alignment and mixture-of-experts, 2026. Method referred to as SFAM; defines the Cross-AUC metric.

Dat Nguyen, Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. FakeFormer: Efficient vulnerability-driven transformers for generalisable deepfake detection, 2024a.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Nguyen_LAA-Net_Localized_Artifact_Attention_Network_for_Quality-Agnostic_and_Generalizable_Deepfake_CVPR_2024_paper.pdf.

Dat Nguyen, Enjie Ghorbel, Anis Kacem, Marcella Astrid, and Djamila Aouada. LAA-X: Unified localized artifact attention for quality-agnostic and generalizable face forgery detection, 2026.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: Consistent representation learning for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 12–21, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/WMF/html/Ni_CORE_COnsistent_REpresentation_Learning_for_Face_Forgery_Detection_CVPRW_2022_paper.html.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision (ECCV)*, 2020. doi: 10.1007/978-3-030-58610-2_6.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.

Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html.

Steven Schwarcz and Rama Chellappa. Finding facial forgery artifacts with parts-based detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/WMF/papers/Schwarcz_Finding_Facial_Forgery_Artifacts_With_Parts-Based_Detectors_CVPRW_2021_paper.pdf.

Sahibzada Adil Shahzad, Ammarah Hashmi, Junichi Yamagishi, Yusuke Yasuda, Yu Tsao, Chia-Wen Lin, Yan-Tsung Peng, and Hsin-Min Wang. SAVe: Self-supervised audio-visual deepfake detection exploiting visual artifacts and audio-visual misalignment, 2026.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *European Conference on Computer Vision (ECCV)*, pp. 712–728, 2022. doi: 10.1007/978-3-031-19778-9_41.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Shiohara_Detecting_Deepfakes_With_Self-Blended_Images_CVPR_2022_paper.html.

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.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. doi: 10.1609/aaai.v36i2.20130.

Ke Sun, Shen Chen, Taiping Yao, Haozhe Yang, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Towards general visual-linguistic face forgery detection, 2023. Preprint known as VLFFD; distinct from the CVPR 2025 paper of the same title.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Method known as FFTG.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance. In *Pattern Recognition. ICPR International Workshops and Challenges (ICPR 2020 Workshops)*, Lecture Notes in Computer Science, pp. 442–456. Springer, 2021. doi: 10.1007/978-3-030-68821-9_38.

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, 2025.

Kutub Uddin, Nusrat Tasnim, and Byung Tae Oh. Face2Parts: Exploring coarse-to-fine inter-regional facial dependencies for generalized deepfake detection. *IEEE Access*, 14:55111–55125, 2026. doi: 10.1109/ACCESS.2026.3679459.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. FSFM: A generalizable face security foundation model via self-supervised facial representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jian Wang, Yunlian Sun, and Jinhui Tang. LiSiam: Localization invariance siamese network for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2022. doi: 10.1109/TIFS.2022.3186803.

Junfeng Xu, Xintao Liu, Weiguo Lin, Wenqing Shang, and Yuefeng Wang. Localization and detection of deepfake videos based on self-blending method. *Scientific Reports*, 15(3927), 2025. doi: 10.1038/s41598-025-88523-1.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yan_UCF_Uncovering_Common_Features_for_Generalizable_Deepfake_Detection_ICCV_2023_paper.html.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward next-generation deepfake detection. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Abstract-Datasets_and_Benchmarks_Track.html.

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

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. doi: 10.1109/ICASSP.2019.8683164.

Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of CLIP in generalizable deepfake detection, 2025. arXiv preprint; distinct from the same group's WACV 2026 paper, cited separately.

Andrii Yermakov, Jan Cech, Mario Fritz, and Jiri Matas. The alpha blending hypothesis: Compositing shortcut in deepfake detection, 2026a.

Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026b. doi: 10.1109/WACV61042.2026.00082. Method name GenD.

Jiangling Zhang, Weijie Zhu, Jirui Huang, and Yaxiong Chen. LAMM-ViT: AI face detection via layer-aware modulation of region-guided attention. In *European Conference on Artificial Intelligence (ECAI)*, 2025.

Yaning Zhang, Linlin Shen, Zitong Yu, Chunjie Ma, and Zan Gao. GazeCLIP: Gaze-guided CLIP with adaptive-enhanced fine-grained language prompt for deepfake attribution and detection, 2026.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021b. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Learning_Self-Consistency_for_Deepfake_Detection_ICCV_2021_paper.html.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision (ECCV)*, 2022. doi: 10.1007/978-3-031-20065-6_23.

APPENDIX

A FULL RESULTS FOR ALL PUBLISHED DETECTORS

Every published detector on every sweep, so that any ranking claim in Section 5 can be checked against the full field, whereas the main text compares PARTAWARE only with the strongest. Our own further configurations are in Appendix Table F.2.

A.1 CROSS-DATASET TRANSFER

Table A.1: Cross-dataset transfer for all nineteen detectors, our re-evaluations of released checkpoints on identical aligned crops (Section 4). Means are frame (fr.) and video (vid.) level over the same frames, and per-benchmark columns are video level. Family codes abbreviate the five families of Section 4. Rows are sorted by family then year, best in bold, second best underlined, and the dash marks this paper. ‡ marks training beyond FaceForensics++ c23 (Appendix B).

			FF++ (in-dist.) (AUC ↑)		**Cross-dataset mean** (AUC ↑)		**Cross-dataset** (video AUC ↑)			
Det.	**Fam.**	**Year**	fr.	vid.	fr.	vid.	**Celeb-DF-v2**	**DFD**	**DFDC-P**	**UADFV**
Xception	CNN	CVPR'17	0.9745	0.9881	0.8286	0.8784	0.8139	0.9618	0.7629	0.9751
EfficientNet-B4	CNN	ICML'19	0.9664	0.9839	0.7940	0.8442	0.8257	0.9317	0.6638	0.9557
CORE	CNN	CVPRW'22	0.9693	0.9843	0.8312	0.8752	0.8167	0.9608	0.7352	0.9881
RECCE	CNN	CVPR'22	0.9734	0.9898	0.8037	0.8599	0.8083	0.9532	0.7289	0.9492
UCF	CNN	ICCV'23	<u>0.9790</u>	0.9905	0.8229	0.8564	0.8285	0.9328	0.6961	0.9681
F3Net	Freq	ECCV'20	<u>0.9790</u>	<u>0.9907</u>	0.8096	0.8465	0.8070	0.9216	0.7190	0.9383
SPSL	Freq	CVPR'21	0.9474	0.9752	0.8227	0.8811	0.7992	0.9620	0.7821	0.9811
SRM	Freq	CVPR'21	0.9628	0.9860	0.8030	0.8478	0.8307	0.9656	0.6684	0.9264
LAA-Net‡	Blend	CVPR'24	0.8245	0.8673	0.8708	0.9344	0.9301	0.9524	0.8652	0.9897
FakeFormer‡	Blend	arXiv'24	0.7674	0.7892	0.8869	0.9296	0.9277	0.9048	0.8961	0.9897
ProDet‡	Blend	NeurIPS'24	0.9433	0.9775	0.8415	0.9208	0.9166	0.9450	0.8342	0.9876
FSFM	SSL	CVPR'25	0.9390	0.9616	0.8405	0.8780	0.8648	0.8305	0.8744	0.9421
Effort	Adapt	ICML'25	0.9246	0.9491	0.9237	0.9570	0.9237	0.9987	0.9073	0.9984
ForensicsAdapter	Adapt	CVPR'25	0.9410	0.9675	<u>0.9318</u>	<u>0.9717</u>	0.9494	0.9988	<u>0.9408</u>	0.9978
Unlocking-CLIP	Adapt	arXiv'25	0.9585	0.9767	0.9269	0.9641	0.9613	0.9998	0.8952	**1.0000**
GenD CLIP-L‡	Adapt	WACV'26	0.9587	0.9777	0.9305	0.9691	0.9575	0.9993	0.9195	**1.0000**
GenD DINOv3-L‡	Adapt	WACV'26	0.9700	0.9877	0.8960	0.9567	0.9190	<u>0.9999</u>	0.9135	0.9946
GenD PE-L‡	Adapt	WACV'26	0.9761	0.9864	0.9170	0.9618	<u>0.9628</u>	**1.0000**	0.8875	0.9968
PARTAWARE	**Adapt**	–	**0.9849**	**0.9923**	**0.9449**	**0.9797**	**0.9661**	<u>0.9999</u>	**0.9528**	**1.0000**

Table A.2: Region coverage for all nineteen detectors (frame level, chance ≈ 0.50, 95% CI half-widths 0.010–0.017 for the nine detectors of Table 2). Columns as in Table 2, with best coverage and worst region in bold, second best underlined, and attribution columns unranked. ‡ marks training beyond FaceForensics++ c23.

Detector	Family	Ref.	Coverage ↑	Worst region ↑	Top-1 ↑	Side acc. ↑
Xception	CNN	(Chollet, 2017)	0.602	0.472 (forehead)	0.686	0.771
EfficientNet-B4	CNN	(Tan & Le, 2019)	0.595	0.507 (forehead)	0.681	0.737
CORE	CNN	(Ni et al., 2022)	0.606	0.494 (forehead)	0.707	0.824
RECCE	CNN	(Cao et al., 2022)	0.608	0.491 (forehead)	0.725	0.804
UCF	CNN	(Yan et al., 2023a)	0.614	0.506 (right eye)	0.669	0.743
F3Net	Freq	(Qian et al., 2020)	0.599	0.505 (right eye)	0.698	0.766
SPSL	Freq	(Liu et al., 2021)	0.593	0.467 (forehead)	0.639	0.755
SRM	Freq	(Luo et al., 2021)	0.592	0.450 (forehead)	0.671	0.743
LAA-Net‡	Blend	(Nguyen et al., 2024b)	0.672	0.542 (left eye)	0.763	0.911
FakeFormer‡	Blend	(Nguyen et al., 2024a)	0.623	0.541 (right eye)	0.671	0.706
ProDet‡	Blend	(Cheng et al., 2024)	0.644	0.508 (left eye)	0.665	0.724
FSFM	SSL	(Wang et al., 2025)	0.587	0.511 (right eye)	0.684	0.798
Effort	Adapt	(Yan et al., 2025)	0.646	<u>0.552</u> (left eye)	0.807	0.851
ForensicsAdapter	Adapt	(Cui et al., 2025)	0.637	0.529 (left eye)	0.781	0.854
Unlocking-CLIP	Adapt	(Yermakov et al., 2025)	0.655	0.528 (left eye)	0.779	0.837
GenD CLIP-L‡	Adapt	(Yermakov et al., 2026b)	0.654	0.532 (left eye)	0.780	0.840
GenD DINOv3-L‡	Adapt	(Yermakov et al., 2026b)	<u>0.676</u>	0.531 (left eye)	0.771	0.818
GenD PE-L‡	Adapt	(Yermakov et al., 2026b)	0.660	0.512 (each eye)	0.775	0.808
PARTAWARE	**Adapt**	–	**0.795**	**0.684** (brows)	0.857	0.925

A.2 REGION COVERAGE

Two observations Appendix Table A.2 supports and the main text omits. A spectral stream does not help coverage, since SRM and SPSL, two of the three detectors with one, are the lowest of their era (0.592, 0.593) and F3Net is fourth (0.599), behind EfficientNet-B4 (0.595). The CNN-era detectors (spatial and frequency CNNs) also fail differently from the modern ones. For six of the eight the worst region is the *forehead* (0.450–0.507), five of them below chance, whereas every modern detector fails first on a single eye. Region size therefore predicts failure within a family, not across families.

RECOMPRESSED BLENDS. Blends are composited onto c23 frames without recompression, so a region could in principle be found by the missing second compression pass rather than by the edit. Re-encoding the clean frame and every blend once as JPEG before scoring removes that cue (Appendix Table A.3). PARTAWARE keeps its lead at every quality, with 0.776 against at most 0.654 at quality 90 and 0.751 against at most 0.637 at quality 75, margins of 0.122 and 0.114 against 0.119 uncompressed. Every detector loses coverage, PARTAWARE 0.043 at quality 75 and the published detectors 0.007 to 0.053, so the lead does not rest on a compression cue.

Table A.3: Region coverage with the clean frame and every blend JPEG-recompressed once before scoring (200 frames, twelve recipes, pooled estimator of Section 3.3). Columns give uncompressed, quality 90, quality 75. Best in bold, second best underlined.

Detector	clean	q90	q75	Detector	clean	q90	q75
GenD DINOv3-L‡	<u>0.676</u>	<u>0.654</u>	<u>0.637</u>	UCF	0.614	0.597	0.585
LAA-Net‡	0.672	0.638	0.619	RECCE	0.608	0.593	0.574
GenD PE-L‡	0.660	0.633	0.614	CORE	0.606	0.594	0.580
Unlocking-CLIP	0.655	0.639	0.631	Xception	0.602	0.590	0.573
GenD CLIP-L‡	0.654	0.638	0.631	F3Net	0.599	0.590	0.576
Effort	0.646	0.631	0.619	EfficientNet-B4	0.595	0.585	0.576
ProDet‡	0.644	0.631	0.612	SPSL	0.593	0.586	0.569
ForensicsAdapter	0.637	0.621	0.611	SRM	0.592	0.578	0.563
FakeFormer‡	0.623	0.617	0.610	FSFM	0.587	0.582	0.580
PARTAWARE	**0.795**	**0.776**	**0.751**				

A.3 CORRUPTION ROBUSTNESS

Results are per detector and corruption family, while Appendix G reads the same sweep by family and Appendix Figure G.1 by severity. Drops are ranked only where clean AUC exceeds 0.85, since a detector near chance cannot fall far.

Table A.4: Corruption robustness on all five benchmarks, giving clean video AUC (Cln), the mean over the 21 corruption conditions of Section 3.4 (Cor), and their drop (Drp, lower is better). *Mean* averages the five Cor columns. Best in bold and second best underlined, with drops ranked only where clean AUC exceeds 0.85. Clean AUC comes from the corruption sweep and can differ from Appendix Table A.1 in the fourth decimal. ‡ marks training beyond FaceForensics++ c23.

		FaceForensics++			Celeb-DF-v2			DFD			DFDC-P			UADFV			
Detector	**Fam.**	Cln↑	Cor↑	Drp↓	Cln↑	Cor↑	Drp↓	Cln↑	Cor↑	Drp↓	Cln↑	Cor↑	Drp↓	Cln↑	Cor↑	Drp↓	**Mean↑**
Xception	CNN	0.9881	0.8873	0.101	0.8140	0.7256	0.088	0.9618	0.8259	0.136	0.7629	0.6894	0.073	0.9751	0.8818	0.093	0.8020
EfficientNet-B4	CNN	0.9839	0.8827	0.101	0.8258	0.7261	0.100	0.9317	0.7789	0.153	0.6638	0.6310	0.033	0.9557	0.8765	0.079	0.7790
CORE	CNN	0.9843	0.8768	0.107	0.8167	0.7340	0.083	0.9608	0.8090	0.152	0.7352	0.6871	0.048	0.9881	0.8766	0.111	0.7967
RECCE	CNN	0.9898	0.8970	0.093	0.8083	0.7280	0.080	0.9532	0.8184	0.135	0.7289	0.6727	0.056	0.9492	0.8782	0.071	0.7989
UCF	CNN	0.9905	0.9070	0.083	0.8285	0.7517	0.077	0.9328	0.7900	0.143	0.6961	0.6580	0.038	0.9681	0.8882	0.080	0.7990
F3Net	Freq	<u>0.9907</u>	0.9076	0.083	0.8070	0.7478	0.059	0.9216	0.7905	0.131	0.7190	0.6811	0.038	0.9383	0.8650	0.073	0.7984
SPSL	Freq	0.9752	0.8839	0.091	0.7992	0.7230	0.076	0.9620	0.8354	0.127	0.7821	0.7060	0.076	0.9811	0.9195	0.062	0.8136
SRM	Freq	0.9860	0.8968	0.089	0.8307	0.7562	0.075	0.9656	0.8143	0.151	0.6684	0.6631	0.005	0.9264	0.8571	0.069	0.7975
LAA-Net‡	Blend	0.8673	0.7544	0.113	0.9301	0.8033	0.127	0.9524	0.8014	0.151	0.8652	0.7696	0.096	0.9897	0.8734	0.116	0.8004
FakeFormer‡	Blend	0.7892	0.7468	0.042	0.9277	0.8584	0.069	0.9048	0.8043	0.100	0.8961	0.8513	**0.045**	0.9897	0.9642	0.026	0.8450
ProDet‡	Blend	0.9775	0.8476	0.130	0.9166	0.8051	0.111	0.9450	0.7918	0.153	0.8342	0.7554	0.079	0.9876	0.9197	0.068	0.8239
FSFM	SSL	0.9616	0.9080	0.054	0.8648	0.8029	<u>0.062</u>	0.8305	0.7397	0.091	0.8744	0.8245	0.050	0.9421	0.9184	<u>0.024</u>	0.8387
Effort	Adapt	0.9491	0.8773	0.072	0.9237	0.8283	0.095	0.9987	0.9097	0.089	0.9073	0.8160	0.091	0.9984	0.9586	0.040	0.8780
ForensicsAdapter	Adapt	0.9675	0.8981	0.069	0.9493	0.8723	0.077	0.9988	0.9175	0.081	<u>0.9410</u>	<u>0.8682</u>	0.073	0.9978	0.9734	<u>0.024</u>	0.9059
Unlocking-CLIP	Adapt	0.9767	0.9201	0.057	0.9613	0.8897	0.072	0.9998	0.9353	0.065	0.8951	0.8332	0.062	**1.0000**	0.9709	0.029	0.9098
GenD CLIP-L‡	Adapt	0.9777	0.9178	0.060	0.9575	0.8746	0.083	0.9993	0.9303	0.069	0.9195	0.8577	0.062	**1.0000**	**0.9777**	**0.022**	0.9116
GenD DINOv3-L‡	Adapt	0.9877	<u>0.9469</u>	<u>0.041</u>	0.9190	0.8660	**0.053**	<u>0.9999</u>	**0.9421**	**0.058**	0.9135	0.8647	0.049	0.9946	0.9691	0.026	<u>0.9178</u>
GenD PE-L‡	Adapt	0.9864	0.9413	0.045	<u>0.9628</u>	<u>0.8982</u>	0.065	**1.0000**	0.9315	0.069	0.8875	0.8408	<u>0.047</u>	0.9968	0.9560	0.041	0.9135
PARTAWARE	**Adapt**	**0.9925**	**0.9604**	**0.032**	**0.9706**	**0.8991**	0.072	<u>0.9999</u>	<u>0.9416</u>	**0.058**	**0.9550**	**0.9078**	<u>0.047</u>	**1.0000**	<u>0.9759</u>	<u>0.024</u>	**0.9370**

B REPORTED VERSUS REPRODUCED BASELINE ACCURACY

The eight CNN baselines come from official FaceForensics++ c23 checkpoints via DeepfakeBench (Yan et al., 2023b), whose preprocessing convention our shared frontend follows, so the largest group runs on its home preprocessing. The remaining ten come from each work's own release, and no baseline is retrained or reimplemented by us. Their training data is not uniform, which the comparison cannot remove and we therefore record. The LAA-Net and FakeFormer releases are trained on raw-compression self-blends rather than FaceForensics++ c23 forgeries, ProDet on c23 with additional blended pseudo-fakes, and the three GenD releases across several benchmarks rather

than FaceForensics++ alone. Every other release is FaceForensics++ c23. Every baseline is re-run rather than quoted, and Appendix Table B.1 places the two side by side. Three factors account for the differences.

Table B.1: Reported AUC (Rep.) against our re-evaluation (Ours) at the source's aggregation level, frame level above and video level (†) below. Sources are DeepfakeBench (Yan et al., 2023b) for Xception through SRM and otherwise each method's own paper. A dash marks a number the source did not report. ‡ marks training beyond FaceForensics++ c23.

		FF++ (in-dist.)		Celeb-DF-v2		DFD		DFDC-P		UADFV	
Detector	**Fam.**	Rep.	Ours	Rep.	Ours	Rep.	Ours	Rep.	Ours	Rep.	Ours
Sources reporting frame-level AUC											
Xception	CNN	0.9637	0.9745	0.7365	0.7398	0.8163	0.8974	0.7374	0.7400	0.9379	0.9373
EfficientNet-B4	CNN	0.9567	0.9664	0.7487	0.7621	0.8148	0.8588	0.7283	0.6460	0.9472	0.9093
CORE	CNN	0.9638	0.9693	0.7428	0.7487	0.8018	0.9052	0.7341	0.7168	0.9412	0.9541
RECCE	CNN	0.9621	0.9734	0.7319	0.7341	0.8119	0.9064	0.7419	0.7060	0.9446	0.8683
UCF	CNN	0.9705	0.9790	0.7527	0.7700	0.8074	0.8942	0.7594	0.7000	0.9528	0.9275
F3Net	Freq	0.9635	0.9790	0.7352	0.7449	0.7975	0.8870	0.7354	0.7090	0.9347	0.8974
SPSL	Freq	0.9610	0.9474	0.7650	0.7348	0.8122	0.8651	0.7408	0.7497	0.9424	0.9411
SRM	Freq	0.9576	0.9628	0.7552	0.7554	0.8120	0.9075	0.7408	0.6870	0.9427	0.8623
ProDet‡	Blend	–	0.9433	0.8448	0.8207	–	0.8160	0.8116	0.7825	–	0.9467
FSFM	SSL	–	0.9390	0.8505	0.8200	–	0.7970	0.8550	0.8515	–	0.8934
ForensicsAdapter	Adapt	–	0.9410	0.9000	0.8927	0.9330	0.9519	0.8900	0.9080	–	0.9747
PARTAWARE (ours)	**Adapt**	–	0.9849	–	0.9024	–	0.9652	–	0.9330	–	0.9786
Sources reporting video-level AUC											
FSFM†	SSL	–	0.9616	0.9144	0.8648	–	0.8305	0.8971	0.8744	–	0.9421
ForensicsAdapter†	Adapt	–	0.9675	0.9570	0.9494	0.9720	0.9988	0.9290	0.9408	–	0.9978
Effort†	Adapt	–	0.9491	0.9560	0.9237	0.9650	0.9987	0.9090	0.9073	–	0.9984
PARTAWARE (ours)†	**Adapt**	–	0.9923	–	0.9661	–	0.9999	–	0.9528	–	1.0000

FRAME-LEVEL AND VIDEO-LEVEL AUC ARE NOT INTERCHANGEABLE. Most prior work reports frame-level AUC, whereas we report video level, since a moderator judges a video and the strongest recent baselines report it too. Scored at both levels on the same frames, all nineteen detectors are lower at frame level by 0.033 to 0.079 (mean 0.047), so the upper block of Appendix Table B.1 pairs frame-level sources with frame-level reproductions. Aggregation flatters the baselines more than PARTAWARE, whose gap of 0.035 is the third smallest, and the ranking survives the choice, since at frame level PARTAWARE exceeds ForensicsAdapter by 0.013 (0.945 against 0.932), above its 0.008 video-level margin. Mixing levels across papers, a common practice, therefore yields rankings that partly reflect aggregation.

WHERE BOTH PROTOCOLS SCORE THE SAME VIDEOS, THE REPRODUCTION IS UNBIASED. Across the twenty-eight Celeb-DF-v2 and DFDC-P pairs our number averages 1.5 points below the published one, unevenly. Over all fifty-five pairs in Appendix Table B.1 the mean difference is $+0.2$ points, with UADFV the one benchmark where we sit further below (-3.1). On Celeb-DF-v2 the difference is unbiased (mean -0.009, median -0.004, with seven of the fourteen pairs *above* the published value). Almost the whole deficit sits on DFDC-P (mean -0.021, ten of fourteen below), where the shared frontend finds no usable face in 235 videos, so the two protocols score different subsets and the residual reflects the evaluation set, not any model. Where the subsets agree, so do the numbers, as Effort on DFDC-P is within 0.2 points of the published value (0.9073 against 0.9090), ForensicsAdapter within 0.8 points on Celeb-DF-v2, and on DFD our reproductions exceed the published values (0.9988 against 0.972 and 0.9987 against 0.965). The largest Celeb-DF-v2 residuals are FSFM (-5.0) and Effort (-3.2).

THE SHARED CROP EXPLAINS PART OF THE REMAINDER. Each method was tuned with its own detection, alignment and margin. Rebuilding crops at margins 1.1 and 1.6 on fixed subsets of DeepSpeak-v2 (Appendix K) and Celeb-DF-v2 rescores FSFM, which carries the largest residual and is the only face-pretrained backbone, against ForensicsAdapter as a control that already reproduces its published number. On videos accepted at all three margins FSFM improves as the crop tightens from 1.3 to 1.1 ($0.767 \rightarrow 0.833$ on DeepSpeak-v2, $0.899 \rightarrow 0.917$ on Celeb-DF-v2), while the control *loses* 0.047 and 0.004 at that margin and is flat across the three on Celeb-DF-v2 (spread 0.004). The effect is therefore specific to FSFM rather than a protocol-wide penalty, and partial, covering about half its DeepSpeak gap and a third of its Celeb-DF-v2 gap. Native per-method preprocessing would

recover more but forfeit the cross-method comparability our claims rest on, so the penalty is disclosed and each method's own number appears in the reported column.

C Extended Related Work

C.1 Generalizable and Region-Aware Detectors

Generalizable detection. The dominant line trains on FaceForensics++ (Rössler et al., 2019) and reports AUC on held-out manipulations and datasets. A line of CNN detectors fixed the cross-dataset suite (Celeb-DF-v2, DFD, DFDC-P, UADFV) and the habit of one transfer number per model, namely Xception (Chollet, 2017; Rössler et al., 2019), EfficientNet-B4 (Tan & Le, 2019), F3Net (Qian et al., 2020), SPSL (Liu et al., 2021), SRM (Luo et al., 2021), CORE (Ni et al., 2022), RECCE (Cao et al., 2022) and UCF (Yan et al., 2023a). Large pretrained backbones followed. Effort is a low-rank update of a CLIP-scale backbone (Yan et al., 2025). ForensicsAdapter adds forgery-boundary supervision on the same backbone family and gives the strongest transfer among the detectors reproduced here (Cui et al., 2025). GenD tunes *only* LayerNorm (0.03% of the backbone) of CLIP, DINOv3 and Perception Encoder backbones in an L2-normalized hyperspherical space (Yermakov et al., 2026b), and Unlocking-CLIP targets partially manipulated faces, where single features change and context is kept (Yermakov et al., 2025), which is our motivation too, tested for sufficiency in Section 5.2. ProDet combines whole-face blending with real deepfakes on EfficientNet-B4, asking whether pseudo-fakes can replace real forgery data (Cheng et al., 2024), and FSFM pretrains a face-specific self-supervised backbone before fine-tuning (Wang et al., 2025). Neither localizes regions (Section 5), so for ProDet self-blend training alone does not suffice. Foveation and attention-zoom methods (MAT (Zhao et al., 2021a), CADDM (Dong et al., 2023), LAA-Net (Nguyen et al., 2024b), LAMM-ViT (Zhang et al., 2025)) sharpen fine spatial detail without validated region attribution. They are complementary, and foveation is unused because the median FaceForensics++ face holds no recoverable detail beyond our crop. Their differing cropping conventions motivate one shared protocol (Section 4).

Region-aware detectors. Segmentation-guided spatial indexing (Al-Zyoud & El Saddik, 2026) is the closest prior architecture, combining DINOv3 patch tokens with a frozen face parser, pooled region tokens and region-indexed AUCs. We share its backbone family and region pooling, and differ in three ways, namely exact region truth from P-SBI construction rather than a parser, an audit of the whole published detector zoo rather than one model, and validation by Shapley attribution against known truth (Section 3). Face2Parts (Uddin et al., 2026) feeds frame, face and key-region streams (lips, eyes, nose) as network *inputs*, whereas here anatomy is training and evaluation signal only, unused at test time (mask-free scoring changes AUC by under 0.001, Section 5.4). Concurrent preprints explore part-grounded reasoning (VIGIL (Li et al., 2026)) and gaze-guided prompting (GazeCLIP (Zhang et al., 2026)) over facial regions, promising for a later reasoning-over-attribution stage. VLFFD and FFTG (Sun et al., 2023; 2025) annotate forgery masks with text over four coarse regions and five forgery types for CLIP co-training, but that vocabulary cannot be scored for correctness, whereas P-SBI labels are checkable by construction at a finer fourteen-part, left/right-aware granularity.

C.2 Pseudo-Fake Synthesis and Consistency Objectives

Self-blended images (SBI) (Shiohara & Yamasaki, 2022) blend a face with a geometrically and color-jittered copy of itself, a strong frame-level signal without real forgeries, and P-SBI generalizes this to parts. SLADD (Chen et al., 2022) is the nearest precedent for region-configurable blending, with pseudo-fakes drawn from a landmark pool, the configuration hardest for the current model selected adversarially, and region and type prediction heads. We differ by a fixed fourteen-part left/right schema instead of a landmark-pool space, by reusing the construction for exact-truth *evaluation*, and by showing that hardest-of-k selection on our configuration *reduces* coverage (Section 5), to our knowledge unreported for this family. Xu et al. (2025) blend random facial-organ combinations to supervise pixel-level localization, and SAVe (Shahzad et al., 2026) adds lip- and lower-face-confined blends to a whole-face self-blend branch for audio-visual detection. Neither yields a single-region label over a named schema, a per-part head, or an evaluation. Earlier parts-based detection trained one classifier per coarse region on existing fakes (Schwarcz & Chellappa, 2021), without synthesis and hence without exact region truth. PCL+I2G (Zhao et al., 2021b), the earliest exact-mask precedent

we know, supervises patch-pair consistency with exact blend masks, and our region-contrastive term (Section 5.4) is its region-pooled analog, ablated at part granularity. SeeABLE (Larue et al., 2023) localizes soft discrepancies on a coarse 4×4 grid by bounded contrastive regression to location-by-type prototypes, which a prototype head over our part embeddings would extend. FreqBlender (Li et al., 2024) and DCL (Sun et al., 2022) blend or contrast in the frequency domain, orthogonal to our question. A separate line enforces consistency instead of synthesizing fakes. CORE and LiSiam enforce global or localization-map invariance across augmented views (Ni et al., 2022; Wang et al., 2022), and our region-contrastive term is the part-level analog, applied across a pristine/blended pair rather than two augmented views. UIA-ViT uses attention-derived pseudo-masks for patch consistency (Zhuang et al., 2022), where we use exact masks. Multi-view pixel contrast and CFL-Net apply pixel-level contrastive objectives with tamper masks in general, non-face image forensics (Lou et al., 2025).

C.3 Localized-Manipulation Benchmarks and Attribution Methods

FakeParts / FakePartsBench (Liu et al., 2025) is closest in spirit, with $44{,}000$ videos with *partial* manipulations (inpainting, region swaps, temporal edits), on which detectors degrade sharply relative to fully synthesized content. It is video level, without face-part granularity or a per-region AUC, and its public release contains whole-face swaps but no face-part subsets or region masks, so it cannot serve as a per-region test. Our protocol is the face-part-level complement, measuring which regions detectors cannot localize and why (Section 6). The same distinction applies to DF40's manipulation-type finding (Yan et al., 2024), to DiffFace-Edit's region-edited dataset evaluated only in aggregate (Ding et al., 2026), and to in-the-wild evaluations reporting score collapse without locating it (Chandra et al., 2026). Segmentation-guided spatial indexing (Al-Zyoud & El Saddik, 2026) reports per-region AUCs, but for its *own* region probes on fully manipulated data, not for an arbitrary detector's response to a single-region edit. Per-region detectors trained on fully manipulated faces (Tolosana et al., 2021) ask the inverse question, which regions carry the most detectable signal. Seq-DeepFake and SEED (Shao et al., 2022) recover component-level edit sequences rather than detect and attribute, and Chai et al. (2020) study which *patches* make a fully generated face detectable, coarser than anatomical attribution. Our attribution uses Shapley values (Shapley, 1953; Lundberg & Lee, 2017) with antithetic permutation sampling, removing a region by restoring its pixels from the pristine frame. Shapley attribution is not new, but applying it where region truth is known by construction is, since it lets attribution accuracy rather than plausibility be scored (Section 3).

C.4 Positioning Relative to Prior Work

The delta against each neighboring line of work is stated below in one sentence each.

- **Self-blending** (Shiohara & Yamasaki, 2022; Cheng et al., 2024). A whole-face blend carries a frame label but no region label, and confining it to one region enables region supervision and evaluation. ProDet, trained on whole-face blends, covers at 0.644, inside the region-insensitive band (Section 5.2).
- **SLADD** (Chen et al., 2022) judges region-configurable synthesis only by detection AUC, and we add left/right granularity, use the construction for evaluation as well as training, and show hardest-of-k selection lowering coverage (Section 5.4).
- **LAA-Net and FakeFormer** (Nguyen et al., 2024b;a) supervise a blend-derived heatmap rather than a categorical region label, so localization is inspected, not scored. Their checkpoints transfer at 0.934 and 0.930 and cover at 0.672 and 0.623, inside the same band.
- **SFAM** (Luo et al., 2026) uses mask-guided localized augmentation with facial-region experts, evaluated by Cross-AUC. No checkpoint is released, so we adopt its metric rather than score it (Appendix E.3).
- **Region-conditioned architectures** (Uddin et al., 2026; Al-Zyoud & El Saddik, 2026) take anatomy as an inference-time input carrying a parser dependency, whereas ours is training signal and evaluation coordinates only, and removing it at inference moves AUC by $+0.0002$.
- **Annotation-based region supervision** (Sun et al., 2023; 2025) relies on annotations over coarse regions, which are estimates that cannot be scored for correctness, whereas our labels are recorded at synthesis.

- **Localized-manipulation benchmarks and partial-manipulation detectors** (Liu et al., 2025; Yermakov et al., 2025) establish *that* detectors degrade on partial manipulations, but not which regions fail, how that failure differs from calibration, or how to repair it.
- **Joint transfer and robustness evaluation** has a precedent (Deng et al., 2024), but none adds region localization, and measuring all three under one protocol is what separates their sources (Section 6).

D Design Rationale, Blend Signal and Training Details

D.1 Requirements That Determine the Design

Exact region labels. Region truth from a parser, boundary annotator or saliency estimator carries that estimator's error into every reported number, confounding poor localization with noisy ground truth, and only *constructing* the manipulation, recording its region at synthesis, removes the confound. Hence P-SBI (Section 3.2), unlike annotation-based supervision (VLFFD (Sun et al., 2023)) and parser-conditioned architectures (Face2Parts (Uddin et al., 2026), segmentation-guided spatial indexing (Al-Zyoud & El Saddik, 2026)).

No anatomy at test time. A parser at inference adds its failures and cost and breaks where landmarks are unreliable, the in-the-wild conditions detectors must survive. The part mask is therefore a training stream and an *evaluation coordinate system*, never a required input for the verdict or the localization map (the part head's per-region scores are defined on its regions). Dropped for a random 25% of training batches, its removal at evaluation shifts in-distribution AUC by under 0.001 for every adapted configuration in Appendix Table F.2 ($+0.0002$ for PartAware, $+0.0014$ on transfer, Appendix E.1, Section 5.4). Part-conditioned architectures, fed region crops or parser outputs, choose the opposite.

No shortcut in the training signal. A hard-boundary paste leaves a sharp edge that the frequency stem would learn, yielding strong training metrics and no anatomical knowledge. Boundaries are therefore feathered with a random-width Gaussian alpha ramp and the compositing operator drawn from three alternatives (alpha, Laplacian pyramid, Poisson), so no operator's signature substitutes for the region. Withholding one operator, then scoring coverage on it alone (Section 3.3), tests whether the model learned "where" or one blend's appearance.

Comparable scores across detectors. Median clean-frame scores differ by up to 0.32 across the nine detectors of Table 2, so any fixed-threshold metric measures calibration as much as capability. Coverage is therefore a threshold-free, ranking-based mean of region-AUCs (Section 3.3), computed for all detectors on byte-identical crops from one preprocessing stack, so differences reflect models, not crop pipelines (Section 4).

D.2 Signal Injected by a Part-Level Blend

Appendix Figure D.1 shows the construction on one held-out frame and Appendix Figure D.2 the twelve recipes, which cover all fourteen parts and span 0.49% of the crop for a single eye to 46% for the full-face recipe. Only the full-face recipe includes face skin, so no confined recipe scores it. Because the region is chosen before the donor is composited, the mask in the last column of Appendix Figure D.1 is the label itself rather than an estimate of one, which is the property the coverage protocol of Section 3.3 depends on.

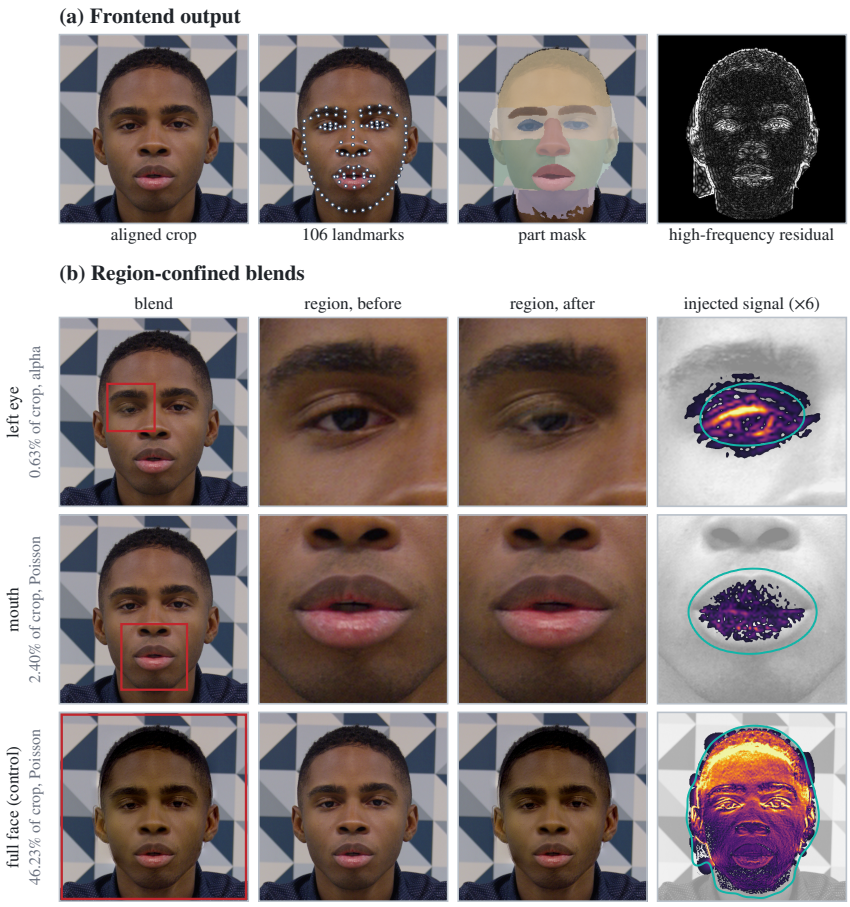

Figure D.1: **What a part-level self-blend does to a face.** **(a)** Shared-frontend output for one held-out DFD frame, with aligned crop, 106 landmarks, part mask over fourteen parts of the face and high-frequency residual. Each row of (b) names its compositing operator. **(b)** Three recipes spanning edited area 0.63%, 2.40% and the 46% full-face recipe. Columns give the blended crop with the region boxed, the region magnified before and after, and the absolute difference (×6) over the desaturated face with the recorded blend mask outlined in teal. The frame is the held-out DFD frame of Figure 2 (yaw proxy 0.02). Appendix Figure D.2 names the recipes and Appendix Figure D.3 gives the crop-scale view.

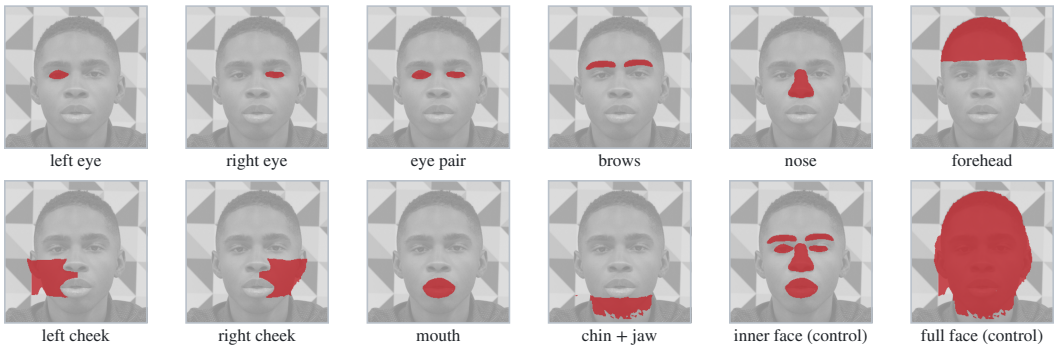

Figure D.2: **The twelve P-SBI recipes** on the frame of Appendix Figure D.1, namely ten region-confined recipes and two large-region controls covering all fourteen parts. Red marks the region each blend is confined to, and the regions span 0.49% of the crop for a single eye to 46% for the full-face recipe. Only `full_face` includes face skin, so no confined recipe tests it and it is gray in Appendix Figure H.1.

D.3 Residual Injected by a Confined Blend

Appendix Figure D.3 measures what a confined blend injects, over 60 pristine frontal test faces synthesized with the training code and rejection rule. Two properties carry into the results. Injected energy scales with edited area, a single eye receiving $41\times$ less than a full-face blend, which is why a detector trained on whole-face blends is insensitive to small regions (Section 5.2). And over 99% of that energy lies below 0.25 cycles per pixel, so a region-confined blend is not a high-frequency artifact and cannot be found by a spectral shortcut. Appendix Figure D.4 shows the three supervision targets one blend yields.

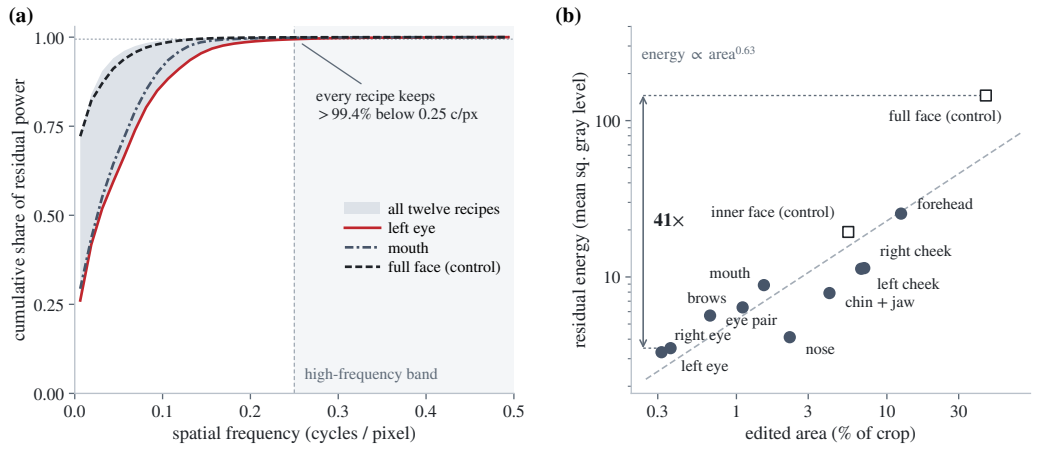

Figure D.3: **Signal injected by each recipe**, on 60 pristine frontal FaceForensics++ test faces with the training synthesis code and rejection rule (a recipe is skipped where its region is too small, 0–11 faces per recipe). **(a)** Cumulative share of residual power (blend − clean, luminance) against frequency, as a gray min–max band over all twelve recipes, with the smallest region (left eye), a mid-sized one (mouth) and the whole-face control drawn individually. Over 99% of the power lies below 0.25 cycles per pixel for every recipe. **(b)** Residual energy against edited area, where a single eye (0.3% of the crop) injects $41\times$ less energy than a full-face blend (Appendix Figure D.2).

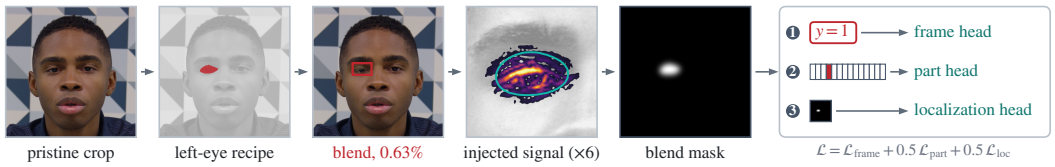

Figure D.4: **Where the region label comes from**, on the frame of Figure 2, showing a pristine frame, one of twelve recipes naming a region, the blend confined to it (feathered boundary, randomly chosen compositing operator), the injected residual magnified against the blend mask, and the region label itself. The right-hand panel shows the three targets one blend yields, where the blend mask supervises the localization head, the frame label the frame head, and the fourteen-dimensional region label the part head. Appendix Figure D.1 shows recipes from 0.49% to 46% of the crop.

D.4 Heads, Objective, Optimization and Overlap Controls

Heads and objective. The **frame head** concatenates mean-pooled fused tokens with a part-context vector, namely part embeddings weighted by a softmax over the per-part logits, and emits one logit linearly. Zeroing that vector gives the mask-free mode scored by the part-stream ablation and the robustness sweep. It is trained with binary cross-entropy at $\text{pos_weight} = N_{\text{real}}/N_{\text{fake}} \approx 0.251$, taken from the manifest labels before P-SBI converts about half the pristine samples, each into one blend drawn once per sample and fixed across epochs, and after conversion the weighted fake-to-real ratio

is about two to one. The **part head** pools tokens per region via the part mask on the patch grid (a cell joins a region at 5% coverage) and scores each with one shared linear classifier plus a learned per-part bias, giving fourteen logits. Geometric rather than learned pooling anchors the region signal to the image coordinates Section 3.3 scores against. The **localization head** decodes a dense localization map, supervised by the exact P-SBI blend mask for pseudo-fakes and, for native FaceForensics++ forgeries, by an artifact mask from ECC registration to the pristine source and thresholding of a fused intensity, chroma and gradient difference. The objective is

$$\mathcal{L} = \mathrm{BCE}_{\text{frame}}(\hat{y}, y;\ \text{pos_weight}) \ +\ 0.5\,\mathcal{L}_{\text{part}} \ +\ 0.5\,\mathcal{L}_{\text{loc}}. \tag{2}$$

Optimization and augmentation. AdamW, cosine-annealed learning rate 10^{-4}, weight decay 0.05, gradient clipping 1.0, batch size 8–64 by backbone (16 for PartAware), up to 10 epochs with early stopping on validation AUC (patience 4 for PartAware and the other rank-32 configurations, 3 otherwise), seed 123. Augmentation is the photometric jitter of the P-SBI donor plus, with probability 0.8 after blending, a downscale by a factor in $[0.5, 1]$ and JPEG recompression at quality 40–100. Noise, blur, occlusion, saturation and contrast are never applied, so five of the seven corruptions of Section 3.4 are unseen in training and the other two seen only below the sweep's settings.

Controls for train/evaluation overlap. Our models train on the blend family the protocol evaluates, so three controls are trained. The **no-P-SBI control** matches backbone, capacity and adaptation but uses frame labels only, so it would score comparably were coverage a product of the protocol rather than the training signal. The **whole-face control** restricts P-SBI to its two large-region recipes, isolating confinement from self-blending. The **held-out-operator control** withholds Poisson compositing and is evaluated on Poisson blends only. Run at frozen DINOv3-B and at the PartAware configuration (Appendix Table F.2), it shows operator dependence to be a small-capacity effect. At frozen DINOv3-B, withholding Poisson costs coverage on Poisson blends, 0.602 with the operator seen in training against 0.546 without it. At the PartAware configuration it costs nothing, since the Poisson-withheld model reaches 0.741 on Poisson blends against 0.720 for the reported model, at 0.978 transfer against 0.980. The frozen-B pair is scored on 40 frames per recipe rather than the 200 used everywhere else. The first two appear in Section 5.4. The walkthrough of Figure 2 is not particular to its frame. On a second held-out DFD subject with a left-eye blend covering 0.56% of the crop, the model returns 0.99 fake against 0.05 for the pristine frame, 0.96 on the left eye and below 0.08 on every other part. Appendix I adds a benchmark with region truth independent of our training and our parser.

E Evaluation Protocol and Estimator Diagnostics

E.1 Face Frontend, Part Masks and Video-Level Aggregation

All detectors must see the same pixels, at the cost of the exclusions below, so every detector reads aligned 256 px crops of the same videos, from FaceForensics++ c23 and from the four zero-shot benchmarks. The splits hold $111{,}173$ train, $22{,}346$ validation and $21{,}840$ test frames from 689 FaceForensics++ test videos, covering Deepfakes, Face2Face, FaceSwap, NeuralTextures and pristine frames. The transfer benchmarks are Celeb-DF-v2 (518 test videos), DeepFakeDetection (363 real and a fixed-stride sample of 400 of its $3{,}068$ fakes), the DFDC Preview test set (777 videos) and UADFV ($49{+}49$).

Frontend. Crops come from InsightFace detection with the 106-point landmark model, ArcFace five-point alignment at margin 1.3, a SegFormer parser trained on the nineteen CelebAMask-HQ labels, and an amodal completion network recovering occluded parts' visible extent, and the fourteen parts are that output restricted to the visible face. We reject faces below 64 pixels or under 90% visible and frames without a mask. All $21{,}840$ FaceForensics++ test frames receive one. In 600 sampled masks, nose, lips, cheeks, forehead and chin always appear, brows and eyes in 98.7–100%, jawline 97.7% and inner mouth 92.0% (closed mouths), and mean areas span 0.2% (inner mouth) to 12% (forehead) of the crop. On transfer, 412 of $2{,}156$ videos yield no accepted face (Celeb-DF-v2 1, DFD 164, DFDC-P 235, UADFV 12), mostly multi-person or profile footage from DFD and DFDC-P, excluded for every detector alike. The exclusion is not class-balanced, as DFD loses 29% of real and 15% of fake videos ($105/363$, $59/400$), DFDC-P 24% and 34% ($66/276$, $169/501$),

UADFV six of each. Our per-video manifests therefore carry acceptance flags. Scores shift across domains. At threshold 0.5, PARTAWARE calls 25% of Celeb-DF-v2 real videos fake, against 8% (DFDC-P), 7% (UADFV) and 3% (DFD), with median real scores 0.32, 0.17, 0.18 and 0.11. The shift does not make a fixed threshold unusable, since video accuracy at 0.5 is 0.899, 0.869, 0.965 and 0.988 on those four benchmarks, and a threshold fitted on the FaceForensics++ test split alone (0.465) lands within 0.02 of each benchmark's own optimum except on UADFV (0.05), whose 86 accepted videos leave that optimum unstable. Comparisons remain threshold-free because published detectors calibrate differently, so a fixed-threshold metric would partly measure calibration rather than detection, although deployment should still calibrate per domain.

MASK ERRORS BARELY CHANGE THE FRAME SCORE. In training a blend's part target comes from the exact blend mask, not the parser (Section 3.2), so mis-segmentation shifts which pixels a recipe blends and which cells the part head pools, never the correctness of the target it trains against. At evaluation the score barely depends on the mask (Appendix Table E.1). Against 0.795 coverage with the true mask, dilating PARTAWARE's scoring mask by 4 or 10 pixels gives 0.790 and 0.788, eroding by the same amounts 0.791 and 0.788, shifting by 8 or 16 pixels 0.794 and 0.792, permuting the fourteen part identities 0.795, and removing it entirely 0.791. Transfer needs no mask either, since scored mask-free PARTAWARE reaches a four-benchmark mean of 0.9814 (Celeb-DF-v2 0.9705, DFD 0.9999, DFDC-P 0.9549, UADFV 1.000) against 0.9797 with it, so every detection and coverage number is reproducible without a test-time parser, and only the part head's region names read the mask. Published detectors' coverage uses no mask beyond blend synthesis, itself the same parser for every detector.

Table E.1: Coverage of PARTAWARE under a degraded scoring mask, where blends are always synthesized from the true mask.

Scoring mask	Coverage	Worst region
true mask	0.795	0.684 (brows)
dilated 4 px / 10 px	0.790 / 0.788	0.682 / 0.676 (brows)
eroded 4 px / 10 px	0.791 / 0.788	0.673 / 0.666 (brows)
shifted 8 px / 16 px	0.794 / 0.792	0.683 / 0.681 (brows)
part identities permuted	0.795	0.684 (brows)
no mask	0.791	0.677 (brows)

VIDEO AGGREGATION. Frame scores are averaged within a face track, and a video takes the maximum over its tracks, since a video is fake if any face in it is and the common largest-face rule scores the wrong face when the manipulated face is not the largest. A secondary column using the largest track by face-pixel count agrees within 0.2 points on Celeb-DF-v2 and 1.6 on DFDC-P and UADFV, but on DFD, where multi-person scenes are common, diverges by 2.3 points for PARTAWARE and up to 5.6 for some published detectors. Max-over-tracks is used throughout, and both per-video columns are stored.

SCOPE AND COST OF THE SHARED PROTOCOL. We evaluate neither fully synthesized faces nor full-frame video and use no temporal or audio signal. One shared crop stack keeps every comparison fair but rejects videos in which it finds no usable face (235 on DFDC-P) and penalizes detectors tuned to their own preprocessing, most visibly FSFM. Both effects are measured and disclosed (Appendix B). The part head names the edited region in 0.69 of cases, so localization is a readout to be audited rather than trusted (Appendix L.3). The reported model runs at 18.3 images per second, a third of the DINOv3-B throughput, which is what its transfer and robustness cost (Appendix Table F.5).

E.2 SAMPLE-SIZE STABILITY OF THE COVERAGE ESTIMATOR

The coverage ranking is insensitive to the number of frames scored. Appendix Table E.2 subsamples the $n = 200$ frames of every coverage run (2,000 bootstrap draws per size) and adds an independent $n = 400$ run on a fresh sample for the nine detectors of Table 2. The interval half-width shrinks as $n^{-1/2}$, from 0.045 at $n = 25$ to 0.016 at $n = 200$ for PARTAWARE. The ranking of the nine holds over bootstrap subsamples from $n = 50$ upward, and the $n = 400$ run agrees with $n = 200$ within

the interval, swapping only adjacent published detectors within 0.01 (Effort/ProDet, GenD CLIP-L/Unlocking-CLIP). Appendix H gives per-region ROC curves and operating points.

Table E.2: Coverage against sample size, with 95% bootstrap intervals from subsamples of the n=200 run, its point estimate, and an independent n=400 run on fresh frames. ‡ marks training beyond FaceForensics++ c23.

Detector	**Fam.**	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 200$	$n = 400$
ProDet‡	Blend	[0.612, 0.693]	[0.620, 0.678]	[0.627, 0.665]	[0.631, 0.658]	0.644	0.649
FSFM	SSL	[0.571, 0.631]	[0.574, 0.616]	[0.576, 0.607]	[0.579, 0.600]	0.587	0.584
Effort	Adapt	[0.614, 0.702]	[0.620, 0.685]	[0.626, 0.672]	[0.631, 0.662]	0.646	0.641
ForensicsAdapter	Adapt	[0.610, 0.687]	[0.615, 0.668]	[0.621, 0.658]	[0.625, 0.651]	0.637	0.635
Unlocking-CLIP	Adapt	[0.622, 0.709]	[0.630, 0.693]	[0.635, 0.680]	[0.642, 0.671]	0.655	0.656
GenD CLIP-L‡	Adapt	[0.624, 0.709]	[0.629, 0.689]	[0.637, 0.677]	[0.641, 0.670]	0.654	0.657
GenD DINOv3-L‡	Adapt	[0.641, 0.733]	[0.649, 0.715]	[0.655, 0.701]	[0.662, 0.694]	0.676	0.672
GenD PE-L‡	Adapt	[0.631, 0.711]	[0.638, 0.692]	[0.643, 0.682]	[0.648, 0.675]	0.660	0.657
PARTAWARE	**Adapt**	[0.754, 0.843]	[0.767, 0.829]	[0.773, 0.817]	[0.779, 0.810]	0.795	0.796

E.3 CROSS-DOMAIN SCORE COMPARABILITY (CROSS-AUC)

Per-benchmark AUC cannot show whether a single decision scale separates real from fake across domains. Cross-AUC (Luo et al., 2026) averages $\mathrm{AUC}(R_i, F_j)$ over all ordered pairs $i \neq j$ of the real videos of benchmark i against the fake videos of benchmark j, so scores that drift between domains lose Cross-AUC even at high per-benchmark AUC. Appendix Table E.3 computes it over the four transfer benchmarks from the per-video scores of Appendix Table A.1. PARTAWARE attains the highest Cross-AUC with no gap to its within-benchmark mean, while LAA-Net, ProDet and SPSL show the largest gaps, hence the least comparable scores across domains.

Table E.3: Cross-AUC (Luo et al., 2026) over the four transfer benchmarks (mean of $\mathrm{AUC}(R_i, F_j)$ over the twelve ordered pairs), the within-benchmark mean of Appendix Table A.1, and their gap. Best per column in bold, second best underlined. ‡ marks training beyond FaceForensics++ c23.

Detector	**Family**	**Ref.**	**Within-benchmark mean ↑**	**Cross-AUC ↑**	**Gap**
Xception	CNN	(Chollet, 2017)	0.8784	0.8690	+0.0095
EfficientNet-B4	CNN	(Tan & Le, 2019)	0.8442	0.8178	+0.0264
CORE	CNN	(Ni et al., 2022)	0.8752	0.8738	+0.0014
RECCE	CNN	(Cao et al., 2022)	0.8599	0.8413	+0.0185
UCF	CNN	(Yan et al., 2023a)	0.8564	0.8452	+0.0112
F3Net	Freq	(Qian et al., 2020)	0.8465	0.8449	+0.0016
SPSL	Freq	(Liu et al., 2021)	0.8811	0.8516	+0.0295
SRM	Freq	(Luo et al., 2021)	0.8478	0.8323	+0.0155
LAA-Net‡	Blend	(Nguyen et al., 2024b)	0.9344	0.9031	+0.0313
FakeFormer‡	Blend	(Nguyen et al., 2024a)	0.9296	0.9329	-0.0033
ProDet‡	Blend	(Cheng et al., 2024)	0.9208	0.8912	+0.0297
FSFM	SSL	(Wang et al., 2025)	0.8780	0.8789	-0.0009
Effort	Adapt	(Yan et al., 2025)	0.9570	0.9488	+0.0082
ForensicsAdapter	Adapt	(Cui et al., 2025)	<u>0.9717</u>	0.9613	+0.0104
Unlocking-CLIP	Adapt	(Yermakov et al., 2025)	0.9641	<u>0.9731</u>	-0.0090
GenD CLIP-L‡	Adapt	(Yermakov et al., 2026b)	0.9691	0.9729	-0.0038
GenD DINOv3-L‡	Adapt	(Yermakov et al., 2026b)	0.9567	0.9654	-0.0086
GenD PE-L‡	Adapt	(Yermakov et al., 2026b)	0.9618	0.9439	+0.0179
PARTAWARE	**Adapt**	–	**0.9797**	**0.9813**	-0.0017

F Ablations of Training Signal, Backbone and Capacity

F.1 All Trained Configurations and Controls

Each row of Appendix Table F.2 changes one factor from the base, sharing the objective of Equation (2), the optimizer of Appendix D.4 and seed 123, as a construction record, not a candidate set (RQ3). Every configuration feeds its backbone crops scaled to $[0, 1]$ without the pretraining mean and standard deviation, in training and evaluation alike, which LoRA adapts to but the frozen row reads unadapted, so that row's coverage is a lower bound for a frozen encoder, while published detectors use their released normalization.

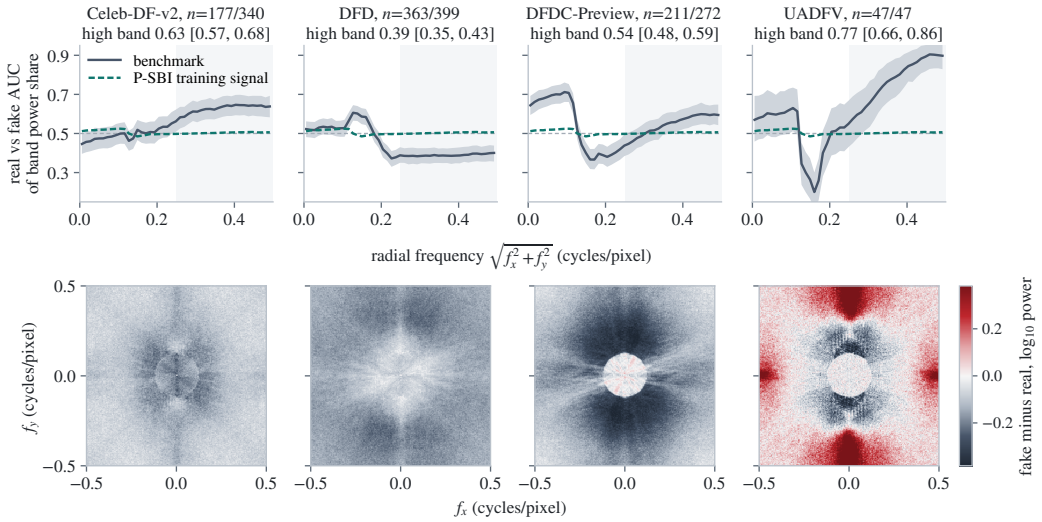

Figure F.1: Generator traces in the residual the frequency stem reads, with no detector involved. The top row gives AUC of real against fake at each radial frequency, from each video's share of residual power in that bin, with a bootstrap 95% interval. The dashed line is chance and the light gray region the high band whose AUC the title gives, with n real/fake videos, and the dashed teal curve is the same measurement for P-SBI blends of training reals, which never leaves 0.49–0.55. The bottom row gives fake minus real class-mean log power over the two-dimensional spectrum, where the flat disc is the band removed by the residual's cutoff. Each generator leaves a distinct trace, yet the stem adds no transfer at the reported scale (Appendix Table F.1).

Training signal. Removing P-SBI costs 2.9 transfer and 10.1 coverage points at DINOv3-B, and 1.1 against 19.0 at the reported scale over three seeds with corrupted AUC unchanged (Appendix Table F.4), so P-SBI supplies region sensitivity and costs almost nothing else. Dropping the part loss, the localization loss or both costs 1.6 to 5.4 coverage points at DINOv3-B, and a whole-face self-blend (Cheng et al., 2024) recovers only part of the gain, so confinement rather than blending produces region sensitivity, and ProDet's own coverage stays inside the published band despite blend training (Appendix Table A.2).

Frequency stem. The stem adds coverage at DINOv3-B (0.029) and DINOv3-L (0.026) but not at DINOv3-H+, where coverage is higher without it by 0.058, a difference comparable to the three-seed spread of PartAware (Appendix Table F.4), and transfer ($+0.0011$), in-distribution corrupted AUC ($+0.0001$) and zero-shot corrupted AUC ($+0.0002$) do not move, so its contribution decays with backbone capacity (Appendix Table F.1). The trained model barely reads its stem at inference either, since a zero residual in place of the real one changes transfer by -0.0004 and coverage by -0.001. Robustness otherwise tracks the backbone, each capacity step gaining one to three points of in-distribution corrupted AUC (Section 5.3). High-frequency residuals have clear precedent (Qian et al., 2020; Liu et al., 2021; Luo et al., 2021), so we claim no novelty for the stem, and the reported model retains it as trained.

Table F.1: The frequency stem at three backbone scales, rows within a pair differing only in the stem. Transfer is given at the reported scale only. *FF++ corr.* and *ZS corr.* are mean corrupted AUC over the 21 conditions of Section 3.4, in distribution and on the four zero-shot benchmarks. Best in bold, second best underlined.

Backbone	Stem	Transfer$^{\uparrow}$	Coverage$^{\uparrow}$	FF++ corr.$^{\uparrow}$	ZS corr.$^{\uparrow}$
DINOv3-B	with	–	0.737	0.9246	0.8498
DINOv3-B	without	–	0.708	0.9218	0.8517
DINOv3-L	with	–	<u>0.821</u>	0.9505	0.9101
DINOv3-L	without	–	0.795	0.9512	0.9029
DINOv3-H+ (PARTAWARE)	with	**0.9797**	0.795	**0.9604**	**0.9311**
DINOv3-H+	without	<u>0.9786</u>	**0.853**	<u>0.9603</u>	<u>0.9309</u>

RESIDUAL TRACES WITHOUT A RESIDUAL BENEFIT. Scoring the residual alone, with no detector involved, by each video's share of residual power in a frequency band, every benchmark's forgeries leave a distinct trace (Appendix Figure F.1), yet the stem adds no transfer at the reported scale, so a separable residual does not imply that a frequency stem helps a strong backbone. The same measurement on the training side settles a different question, since P-SBI blends of FaceForensics++ training reals against those reals, 400 videos per class, sit at chance in every band below 0.5 cycles per pixel (high-band AUC 0.503, 95% CI 0.463–0.544), so region sensitivity cannot be learned as a global spectral shortcut.

FREQUENCY STEM WITHOUT CORRUPTION. On the three evaluations that leave the training distribution without corrupting it, removing the stem changes in-the-wild AUC by -0.028 (809 videos), inpainting coverage by $+0.016$ and $+0.031$, and lipsync AUC by -0.008 on FakeAVCeleb and $+0.027$, $+0.031$ and $+0.016$ on AVLips and DeepSpeak v1 and v2 (Appendices I to K). Only the in-the-wild and FakeAVCeleb changes favor the stem, and the first is inside that benchmark's bootstrap interval.

NO BASELINE COMBINES BOTH PARTS. Every detector in Appendix Table A.1 ends in one whole-face logit. The closest relative, ForensicsAdapter (Cui et al., 2025), is supervised by a Face X-ray boundary target (Li et al., 2020a), so it outputs a boundary map rather than region labels, and the CLIP and Perception Encoder adaptations match our transfer by tuning LayerNorms on a frozen backbone, as we tune LoRA, but cannot answer per region. Region-level behavior therefore requires region-confined synthesis paired with a region-level readout, not scale, backbone family or adaptation budget, all of which baselines match.

F.2 TRAINING SIGNAL ISOLATED AT FIXED CAPACITY

Appendix Table F.3 isolates the training signal twice over, holding the backbone fixed in each pair, comparing an independently trained DINOv3-L detector against our DINOv3-L configuration, and the reported configuration with and without P-SBI. The second pair differs only in P-SBI and the part and localization targets it supplies, whereas the first also differs in tuning, so it bounds the effect rather than isolating it. Appendix Figure F.2 isolates the training signal from the other factors of Appendix Table F.2, and fixed capacity rules out localization by scale alone. DINOv3-B coverage rises from 0.585 (frozen) to 0.636 (LoRA on frame labels), 0.669 (whole-face blends) and 0.737 (region-confined blends), the only step whose interval clears the published band. Transfer rises in that order at video level (0.851, 0.914, 0.924, 0.943) and frame level (0.809, 0.856, 0.881, 0.898), so region confinement costs nothing on the axis the field optimizes. Frame level sits 0.041–0.058 below video level throughout, as expected when a video takes the maximum over its tracks.

F.3 DESIGN ALTERNATIVES, BACKBONE FAMILIES AND VALIDATION-AUC SELECTION

FIXED RECIPE SAMPLING OUTPERFORMS ADAPTIVE ALTERNATIVES. A difficulty curriculum and adversarial hardest-of-three selection both lower coverage (0.666 and 0.657), since each retains the blends that most confuse the current model, near-invisible edits on which the "fake" label then rests. A region-contrastive term between pristine and blended part embeddings lowers both axes at the weight tested (0.708 / 0.931). A 512 px input lowers coverage to 0.696, because the median

Table F.2: All configurations of our detector trained for this paper, each changing the factor named by its group heading from the DINOv3-B base (LoRA rank 16 unless stated, one seed, repeats in Appendix Table F.4). Transfer is the four-benchmark mean video AUC, coverage the twelve-recipe mean region-AUC, and FF++ corr. the mean AUC over the 21 corrupted FaceForensics++ conditions (dash where not swept). Best in bold, second best underlined. Transfer uses the original evaluation path (Reproducibility Statement) and is not compared across the stem. The frozen row reads unnormalized input, so its coverage is a lower bound.

Configuration	Transfer	Coverage	FF++ corr.
Adaptation and training signal, DINOv3-B			
frozen backbone, no P-SBI	0.8508	0.585	–
LoRA, no P-SBI	0.9140	0.636	–
LoRA, whole-face blends only	0.9241	0.669	–
LoRA + P-SBI (base configuration)	0.9434	0.737	0.9246
base configuration, frame loss only (no part or localization loss)	0.9330	0.713	0.9143
base configuration, no localization loss	0.9310	0.683	0.9160
base configuration, no part loss	0.9338	0.721	0.9134
Design alternatives, DINOv3-B with part-level blends			
no frequency stem	0.9379	0.708	0.9218
512 px input	0.9124	0.696	–
curriculum blend sampling	0.9336	0.666	–
hardest-of-3 blend sampling	0.9222	0.657	–
region-contrastive auxiliary loss	0.9308	0.708	–
Backbone family, matched rank			
SigLIP2-B	0.8726	0.627	–
CLIP-L	0.9399	0.762	–
PE-L	0.9534	0.817	–
DINOv3-L	0.9700	0.821	0.9505
PE-G (1.89 B)	0.9241	0.819	–
Scale, adaptation rank and donor recipe			
DINOv3-L, rank 32	0.9555	0.836	0.9472
DINOv3-L, cross-frame mouth donor	0.9695	0.798	0.9508
DINOv3-L, no frequency stem	0.9630	0.795	0.9512
DINOv3-H+, rank 16	0.9790	0.784	0.9585
DINOv3-H+, rank 32, no frequency stem	0.9786	0.853	<u>0.9603</u>
DINOv3-H+, rank 32, no P-SBI (three seeds, Appendix Table F.4)	**0.9800**	0.682	0.9599
DINOv3-H+, rank 32, Poisson blends withheld	0.9781	0.868	–
PARTAWARE (DINOv3-H+, rank 32)	<u>0.9797</u>	0.795	**0.9604**

Table F.3: Same backbone, different training. Published detectors, run from their released checkpoints, against our pipeline on the same backbone. Our CLIP-L and PE-L rows use the development configuration (LoRA rank 16, one seed), and a dash marks a sweep not run. The DINOv3-H+ pair differs only in P-SBI and reports mean ± s.d. over three seeds (Appendix Table F.4). Bold marks the better value within each backbone group.

Backbone	Training	Transfer$^{\uparrow}$	Coverage$^{\uparrow}$	Corrupted FF++$^{\uparrow}$
CLIP-L	Effort (Yan et al., 2025)	0.9570	0.646	0.8773
CLIP-L	Unlocking-CLIP (Yermakov et al., 2025)	0.9641	0.655	**0.9201**
CLIP-L	GenD (Yermakov et al., 2026b)	**0.9691**	0.654	0.9178
CLIP-L	our pipeline	0.9399	**0.762**	–
PE-L	GenD (Yermakov et al., 2026b)	**0.9618**	0.660	**0.9413**
PE-L	our pipeline	0.9534	**0.817**	–
DINOv3-L	GenD (Yermakov et al., 2026b)	0.9567	0.676	0.9469
DINOv3-L	our pipeline	**0.9700**	**0.821**	**0.9505**
DINOv3-H+	our pipeline without P-SBI	0.969 ± 0.010	0.657 ± 0.022	$\mathbf{0.964 \pm 0.004}$
DINOv3-H+	**PARTAWARE**	$\mathbf{0.980 \pm 0.001}$	$\mathbf{0.847 \pm 0.045}$	0.962 ± 0.002

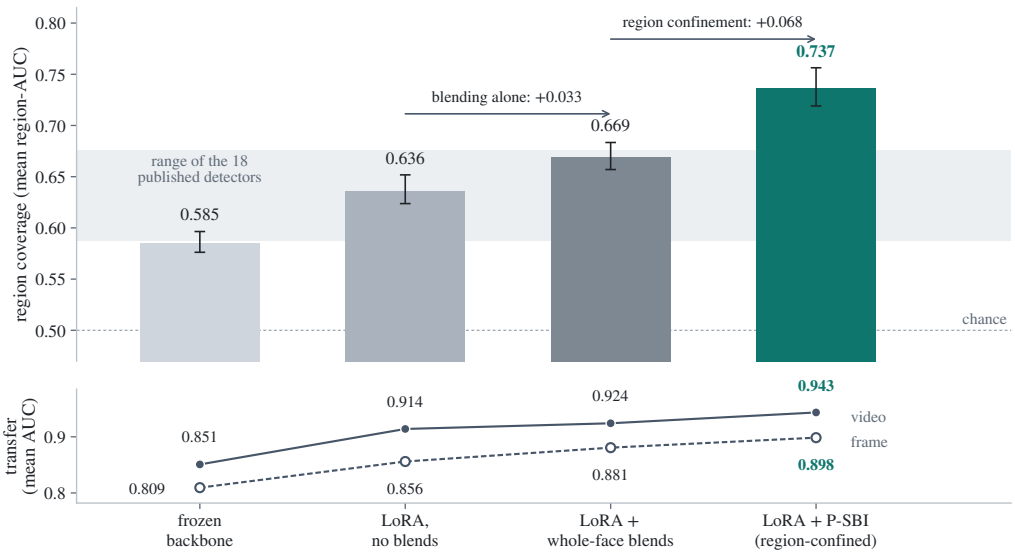

Figure F.2: **Training signal at fixed capacity.** Four DINOv3-B detectors with the same data, optimizer and seed, differing only in training signal, namely frozen backbone, LoRA on frame labels, LoRA with whole-face self-blends and LoRA with P-SBI. The top panel gives coverage with frame-level bootstrap 95% intervals, the band giving the range of the eighteen published detectors. The bottom panel gives cross-dataset transfer at video (solid) and frame level (dashed).

FaceForensics++ face holds no detail beyond the 256 px crop, so upsampling adds interpolation, not evidence.

BACKBONE FAMILY IS A FIRST-ORDER FACTOR. At matched L scale and rank, DINOv3-L leads both axes (0.970 transfer), ahead of the Perception Encoder (0.953) and CLIP-L. Scale within a weaker family does not close the gap, as a 1.89 B-parameter Perception Encoder, 2.2× our DINOv3-H+ model, falls 5.6 transfer points *below* it and below the 86 M base configuration. P-SBI still lifts coverage above the published band on three of four families, two of them unselected (PE-G 0.819, essentially DINOv3-L's 0.821, and CLIP-L 0.762), while SigLIP2 is the exception at 0.627, inside the band. The mechanism is thus mainly a property of the training signal, not of particular pretrained weights, yet not backbone-independent, since one family does not support it.

VALIDATION AUC SELECTS THE WRONG MODEL. The in-distribution/transfer inversion seen in the literature (Appendix L.1) recurs among our configurations, which fix data, training configuration, protocol and seed and vary only the backbone, so it is not an artifact of differing research groups. The PE-G configuration has our highest validation AUC (0.9905, above PARTAWARE's 0.9893) yet is the weakest large backbone on transfer, below the 86 M base configuration, so validation-AUC selection would have preferred it to a model transferring 5.6 points better. We select on transfer and use validation AUC only as a convergence check, making the four transfer benchmarks development benchmarks, and Appendix J reports a benchmark no decision used.

F.4 SEED VARIABILITY AND LORA RANK

Coverage is our noisiest axis, so single-seed differences need error bars. At rank 16, DINOv3-L to DINOv3-H+ lowers single-seed coverage while transfer rises, but three seeds of that configuration span 0.784–0.834 (mean 0.815 ± 0.028) with transfer varying only ± 0.002, so the drop is seed noise, not a capacity/coverage trade. Rank 32 at DINOv3-H+ improves both axes over rank 16, as at L scale. PARTAWARE's two repeats reach 0.877 and 0.868 coverage against the reported seed's 0.795 (Appendix Table F.4), so the reported model is the recipe's low coverage draw, while transfer and robustness barely move. The same three seeds trained without P-SBI (second block of Appendix Table F.4) separate the axes directly. Coverage falls to 0.657 ± 0.022, inside the published band, with the worst region a single eye at chance on every seed. Transfer falls by 0.011, about one control

standard deviation, in-distribution corrupted AUC is unchanged and corrupted Celeb-DF-v2 loses 0.023. Rank 32 is arbitrary within noise, not tuned, since the two ranks differ by 0.0011 in transfer, below the seed spread (0.980 ± 0.001 over seeds 123/456/789), and are level on corrupted AUC (seed s.d. 0.002). Rank 32 and seed 123 were fixed before any repeat and are reported unchanged. On the real region-confined forgeries of Appendix K, rank 16 leads.

Table F.4: Seeds of the PARTAWARE configuration (DINOv3-H+, LoRA rank 32) with and without P-SBI. Corrupted AUC is mask-free over the 21 conditions of Section 3.4. The paper reports seed 123, and $\pm$ is the sample standard deviation over three seeds.

Seed	Transfer	Coverage	Worst region	FF++ corr.	Celeb-DF-v2 corr.
123 (reported)	0.9797	0.795	0.684 (brows)	0.9604	0.8991
456	0.9816	0.877	0.810 (mouth)	0.9646	0.9118
789	0.9800	0.868	0.787 (mouth)	0.9621	0.8942
mean $\pm$ s.d.	0.980 ± 0.001	0.847 ± 0.045	–	0.962 ± 0.002	0.902 ± 0.009
Same configuration without P-SBI					
123	0.9800	0.682	0.540 (right eye)	0.9599	0.9032
456	0.9661	0.641	0.496 (right eye)	0.9672	0.8768
789	0.9613	0.649	0.508 (right eye)	0.9648	0.8570
mean $\pm$ s.d.	0.969 ± 0.010	0.657 ± 0.022	–	0.964 ± 0.004	0.879 ± 0.023

F.5 COMPUTE COST AGAINST CAPABILITY

Appendix Table F.5 reports parameters, trainable parameters and throughput of the full forward pass, including our residual computation, on one RTX A5000 at batch 16, 256 px, fp32. The 89 M DINOv3-B model trained with P-SBI already exceeds the coverage of every published detector at their throughput, the 303 M GenD DINOv3-L and CLIP-L Effort included. DINOv3-L has the best single-seed coverage of the three scales, while DINOv3-H+ leads transfer and robustness at a third of the B-scale throughput. P-SBI supervision, not backbone size, separates region-sensitive from region-insensitive detectors.

Table F.5: Compute against capability within frozen foundation-model adaptation. Trainable parameters are all weights updated in training, for PARTAWARE 7.93 M of LoRA (0.93% of the backbone) plus 2.91 M in stem, fusion and heads (1.27% of the whole model), for Effort its LoRA, and for GenD its LayerNorm parameters, unmarked in its release. Best in bold, second best underlined.

Detector	Params (M) ↓	Trainable (M) ↓	Img/s ↑	Coverage ↑	Transfer ↑
GenD DINOv3-L (Yermakov et al., 2026b)	303	–	**61.4**	0.676	0.9567
Effort (CLIP-L) (Yan et al., 2025)	303	0.20	<u>56.7</u>	0.646	0.9570
DINOv3-B configuration	89	3.23	56.5	0.737	0.9434
DINOv3-L configuration	308	5.00	35.0	**0.821**	<u>0.9700</u>
PARTAWARE (DINOv3-H+)	851	10.84	18.3	<u>0.795</u>	**0.9797**

G CORRUPTION ROBUSTNESS BY DETECTOR FAMILY

For the robustness half of RQ2, one corrupted-AUC number cannot separate an architectural choice from a detector's era. The breakdown by corruption (Appendix Table G.1) shows robustness tracking era, far more so zero-shot. In distribution the older families look competitive, their content familiar and only the corruption new, whereas zero-shot they separate. Over the seven corruptions, spatial CNNs fall from 0.890 to 0.771 and frequency CNNs from 0.896 to 0.780, while frozen foundation-model adaptation loses only $0.917 \rightarrow 0.903$ and PARTAWARE $0.960 \rightarrow 0.931$. Under noise, CNN-era detectors average ≈ 0.73 in distribution and ≈ 0.67 zero-shot, CLIP-scale detectors ≈ 0.88 and ≈ 0.87, ours 0.957 and 0.940. Whole-face blending alone *improves* zero-shot, as LAA-Net and FakeFormer score poorly on FaceForensics++ itself (0.867 and 0.789 clean) and well on Celeb-DF-v2. PARTAWARE is the best group on all seven corruptions in distribution and six of seven zero-shot. Noise falls in the band our frequency stem reads, the obvious place for it to be fragile, but our detectors

are the most noise-robust tested, so the self-supervised backbone and P-SBI training outweigh any fragility it adds. Pixelation hurts every model, ours included (0.68–0.88 across the board), and there frozen foundation-model adaptation edges PARTAWARE zero-shot. Appendix Figure G.1 resolves each family by severity.

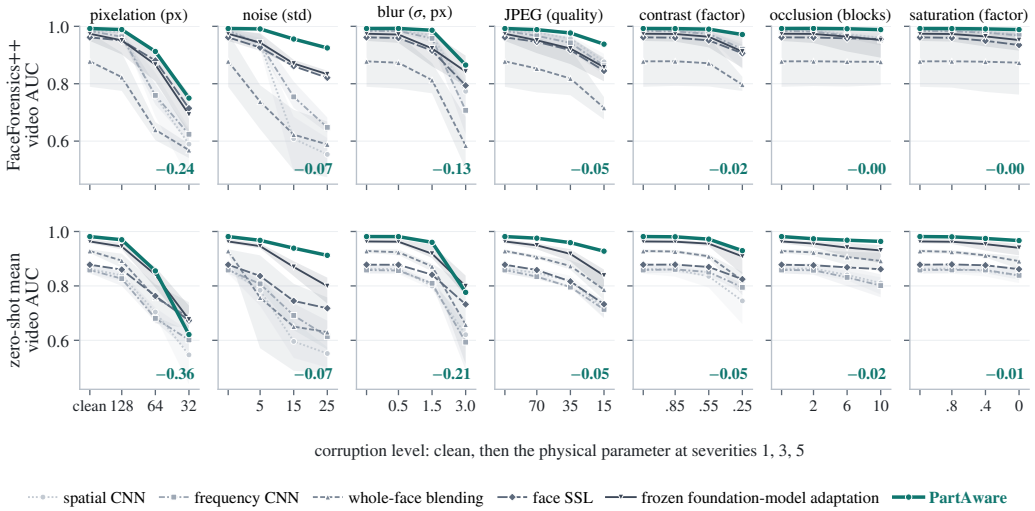

Figure G.1: Video AUC against corruption level, in distribution (FaceForensics++, top) and zero-shot (mean of the four unseen benchmarks, bottom). Each family is drawn as its mean with a min–max band, and PARTAWARE is the teal line, with the teal number the AUC it loses from clean to the highest severity. Panels are ordered by average damage, most damaging first, and ticks give the physical parameter of each level, as in Appendix Table G.1. Pixelation is most damaging (-0.24 in distribution, -0.36 zero-shot for PARTAWARE), whereas under noise PARTAWARE loses 0.07 while the spatial CNNs fall to 0.55.

Table G.1: Corrupted video AUC per corruption (mean of severities 1, 3, 5) and family, on FaceForensics++ and averaged over the four unseen benchmarks. *Level* gives the parameter per severity, namely downsample size from a 256 px crop (pixelation), noise standard deviation on a 0–255 scale, blur σ in pixels, JPEG quality, factor with 0 fully flat or gray (contrast, saturation), and number of 24 px black blocks (occlusion). Entries average each family's detectors (CNN 5, Freq 3, Blend 3, SSL 1, Adapt 6), best group per block in bold and second best underlined.

		FaceForensics++ (in-distribution)						**Four unseen benchmarks** (zero-shot)					
Corruption	**Level** sev. 1/3/5	CNN	Freq	Blend	SSL	Adapt	**Ours**	CNN	Freq	Blend	SSL	Adapt	**Ours**
pixelation	128/64/32	0.775	0.782	0.677	0.849	0.837	**0.884**	0.696	0.703	0.778	0.765	**0.821**	0.815
noise	5/15/25	0.701	0.781	0.649	0.869	0.882	**0.957**	0.644	0.704	0.679	0.766	0.872	**0.940**
blur	0.5/1.5/3.0	0.910	0.883	0.757	0.889	0.912	**0.948**	0.761	0.753	0.818	0.817	0.893	**0.906**
JPEG	70/35/15	0.937	0.924	0.796	0.903	0.911	**0.968**	0.795	0.781	0.855	0.803	0.902	**0.954**
contrast	.85/.55/.25	0.956	0.959	0.849	0.939	0.950	**0.985**	0.813	0.835	0.885	0.858	0.942	**0.961**
occlusion	2/6/10	0.973	0.967	0.877	0.958	0.963	**0.991**	0.838	0.830	0.907	0.869	0.942	**0.968**
saturation	.8/.4/.0	0.980	0.976	0.876	0.948	0.963	**0.991**	0.852	0.852	0.909	0.872	0.952	**0.974**

H PER-REGION COVERAGE, ATTRIBUTION MAPS AND OPERATING POINTS

H.1 COVERAGE PER REGION AND DETECTOR

Averaged over the eighteen published detectors, region-AUC rises with the area of the edited region, from 0.524 on either single eye to 0.807 on the full face, and 17 of the 18 stay below 0.56 on both eyes (Appendix Table H.1). The rise is not purely geometric, since the forehead, at 10% of the crop, scores 0.566, below the mouth (0.602) and the nose (0.619), which are smaller. PARTAWARE gains most where the field is weakest, 0.80 against 0.52 on a single eye, and least where it is already strong,

as on the full-face recipe the best published detector reaches 0.911 against our 0.903. What separates the methods is region sensitivity, not blend detection in general.

Table H.1: Region-AUC (frame level) per recipe for all nineteen detectors, columns roughly smallest region first. Chance is ≈ 0.50, best in bold and second best underlined. ‡ marks training beyond FaceForensics++ c23.

Detector	**Fam.**	**Mean**$^{\uparrow}$	L-eye	R-eye	Eyes	Brows	Nose	Fore.	L-cheek	R-cheek	Mouth	Chin	Inner	Full
Xception	CNN	0.602	0.507	0.491	0.523	0.493	0.623	0.472	0.692	0.656	0.618	0.655	0.718	0.770
EfficientNet-B4	CNN	0.595	0.508	0.530	0.534	0.552	0.591	0.507	0.646	0.614	0.592	0.631	0.686	0.743
CORE	CNN	0.606	0.525	0.510	0.549	0.528	0.636	0.494	0.656	0.647	0.587	0.625	0.727	0.782
RECCE	CNN	0.608	0.511	0.507	0.532	0.511	0.584	0.491	0.689	0.680	0.610	0.670	0.705	0.810
UCF	CNN	0.614	0.523	0.506	0.535	0.567	0.645	0.518	0.684	0.662	0.588	0.643	0.725	0.767
F3Net	Freq	0.599	0.526	0.505	0.532	0.550	0.623	0.518	0.641	0.661	0.579	0.617	0.692	0.739
SPSL	Freq	0.593	0.527	0.508	0.548	0.530	0.607	0.467	0.633	0.666	0.590	0.610	0.720	0.709
SRM	Freq	0.592	0.498	0.484	0.496	0.479	0.620	0.450	0.674	0.680	0.609	0.630	0.707	0.773
LAA-Net‡	Blend	0.672	0.542	0.545	<u>0.596</u>	<u>0.613</u>	0.637	<u>0.684</u>	0.764	<u>0.758</u>	0.593	0.689	0.777	0.867
FakeFormer‡	Blend	0.623	0.545	0.541	0.579	0.571	0.590	0.664	0.597	0.622	0.558	0.607	0.782	0.819
ProDet‡	Blend	0.644	0.508	0.530	0.562	0.570	0.652	0.634	0.649	0.682	0.615	0.626	0.788	**0.911**
FSFM	SSL	0.587	0.524	0.511	0.543	0.549	0.552	0.568	0.588	0.586	0.590	0.606	0.686	0.747
Effort	Adapt	0.646	<u>0.552</u>	<u>0.567</u>	0.581	0.595	0.603	0.649	0.669	0.668	0.587	0.669	0.749	0.861
ForensicsAdapter	Adapt	0.637	0.529	0.554	0.572	0.550	0.640	0.557	0.689	0.685	<u>0.639</u>	0.629	0.787	0.809
Unlocking-CLIP	Adapt	0.655	0.528	0.545	0.560	0.595	0.621	0.623	0.719	0.723	0.631	0.711	0.768	0.832
GenD CLIP-L‡	Adapt	0.654	0.532	0.553	0.562	0.601	0.607	0.626	0.724	0.721	0.622	0.710	0.757	0.838
GenD DINOv3-L‡	Adapt	<u>0.676</u>	0.531	0.541	0.558	0.606	<u>0.671</u>	0.616	<u>0.766</u>	0.739	0.625	<u>0.784</u>	<u>0.804</u>	0.870
GenD PE-L‡	Adapt	0.660	0.512	0.512	0.527	0.576	0.632	0.642	0.749	0.749	0.611	0.777	0.762	0.876
PARTAWARE	**Adapt**	**0.795**	**0.801**	**0.797**	**0.712**	**0.684**	**0.780**	**0.724**	**0.881**	**0.870**	**0.711**	**0.787**	**0.886**	<u>0.903</u>

H.2 COVERAGE AND ATTRIBUTION MAPS

Appendix Figure H.1 places those values on a face, one tile per detector ordered by coverage. The pattern holds across five families and a decade of architectures, as the small central and lateral regions sit at or below chance while the cheeks and the full face carry whatever sensitivity a detector has.

Appendix Figure H.2 repeats the layout for the post-hoc Shapley readout, whose chance level is $1/12$. Detectors at chance on coverage still name the edited region well above chance there, which is the *oracle assistance* analyzed in Appendix L.2, where the pristine pixels restored by the intervention, not the detector, reveal the edited region. Restoring pristine pixels leaves every unedited region with zero marginal contribution, so the intervention locates the edit even where the detector cannot. The two map sets are read together, never the attribution alone.

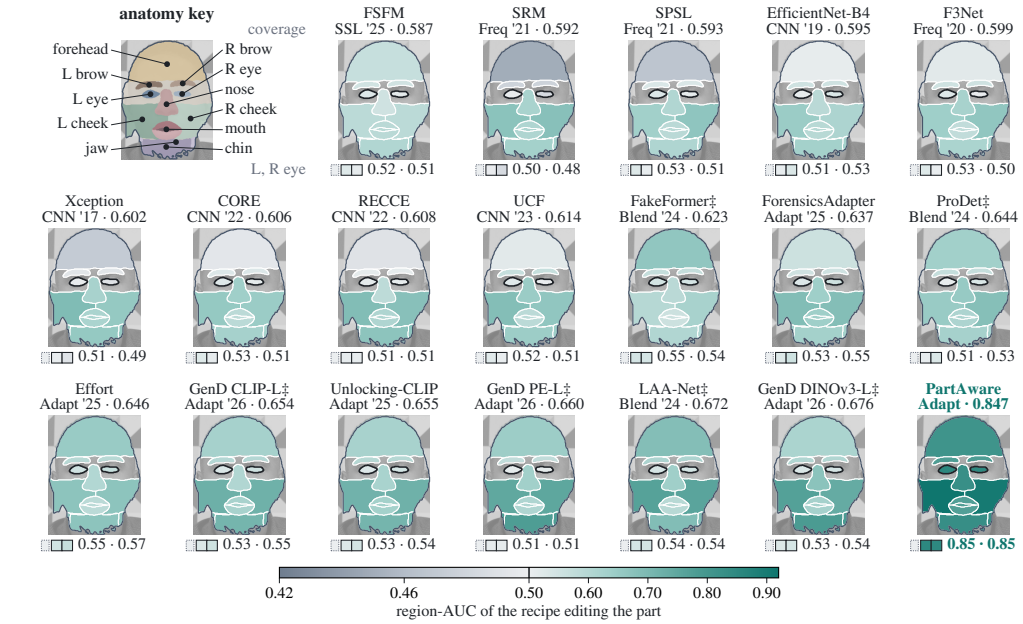

Figure H.1: **Where each detector can see a manipulation.** The first tile is the anatomy key, coloring all fourteen parts and labeling the eleven that fit (left and right refer to the image), and every other tile colors each part by that detector's region-AUC for the recipe editing it, grouped recipes coloring all their parts, with family and venue year in the title. The single eyes (0.4% of the crop each) are too small to read as color, so they are ringed and repeated as swatches beside a dashed cell filled at chance. Panels show the eighteen published detectors by coverage, then PARTAWARE (mean of three seeds). The scale is centered on chance and slate marks region-AUC below it. A double dagger marks a released checkpoint not trained on FaceForensics++ c23 alone (Appendix B).

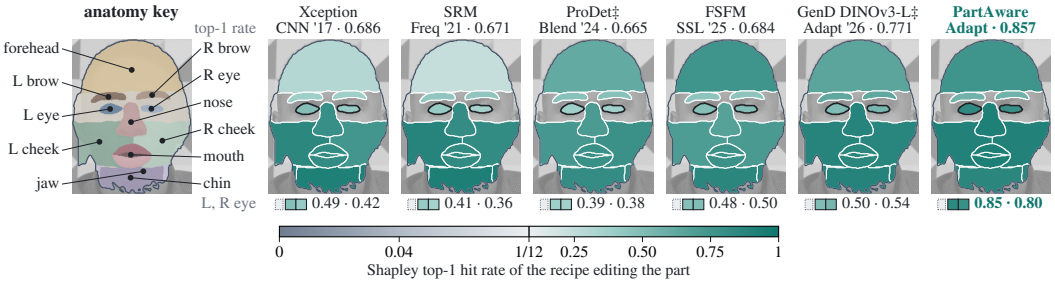

Figure H.2: **Does post-hoc attribution name the edited part?** Laid out as Figure 3, but each part shows the Shapley top-1 hit rate of the recipe editing it, on a 0 to 1 scale whose chance level is $1/12$, and the dashed cell beside each eye pair is filled at that level. Every detector is far above chance on every part, including the eyes it cannot detect at all, which is why these rates are a consistency check on the intervention rather than evidence of localization (Appendix L.2). Appendix Table A.2 gives per-detector rates for the full field. A double dagger marks a released checkpoint not trained on FaceForensics++ c23 alone (Appendix B).

H.3 ROC CURVES AND OPERATING POINTS

Coverage is threshold-free, so Appendix Figure H.3 gives the curves behind it and Appendix Table H.2 two operating points on them. At a 10% false-positive rate on clean frames PARTAWARE detects 0.58 and 0.57 of left- and right-eye blends against 0.12 to 0.18 for the two strongest published detectors, and the ordering is unchanged at 5%.

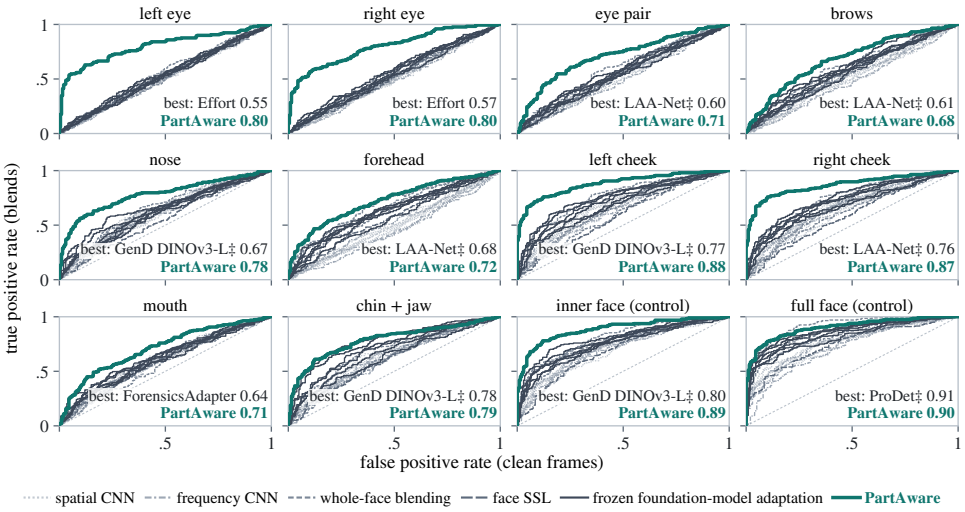

Figure H.3: Per-region ROC curves from the coverage runs (positives are that recipe's blended frames and negatives the pooled clean frames). Published detectors are drawn by family (gray level and line style), PARTAWARE in teal, and each panel names the second-best detector, the strongest published one except on the full face, where ProDet leads and PARTAWARE is second, and prints both AUCs. At a 10% false-positive rate no published detector exceeds 0.19 true-positive rate on either single eye, against 0.58 and 0.57 for PARTAWARE. A double dagger marks a released checkpoint not trained on FaceForensics++ c23 alone (Appendix B).

Table H.2: True-positive rate on blends at 5% / 10% false-positive rate on clean frames, per region.

Region	PARTAWARE	GenD DINOv3-L	ForensicsAdapter
left eye	0.54 / 0.58	0.08 / 0.12	0.09 / 0.13
right eye	0.49 / 0.57	0.09 / 0.13	0.12 / 0.18
eye pair	0.25 / 0.37	0.09 / 0.16	0.11 / 0.16
brows	0.18 / 0.28	0.08 / 0.18	0.11 / 0.15
nose	0.43 / 0.56	0.21 / 0.32	0.26 / 0.32
forehead	0.32 / 0.42	0.08 / 0.18	0.11 / 0.15
left cheek	0.67 / 0.73	0.34 / 0.45	0.30 / 0.36
right cheek	0.68 / 0.76	0.28 / 0.39	0.35 / 0.38
mouth	0.26 / 0.37	0.16 / 0.25	0.25 / 0.30
chin and jaw	0.44 / 0.53	0.37 / 0.47	0.19 / 0.27
inner face	0.64 / 0.74	0.42 / 0.57	0.49 / 0.54
full face	0.69 / 0.78	0.61 / 0.70	0.55 / 0.57

I INDEPENDENT BENCHMARK OF DIFFUSION-INPAINTED REGION EDITS

The coverage protocol shares its construction with P-SBI training. A second benchmark therefore varies manipulation family, image source and region truth together, since every edit is a diffusion inpainting.

CONSTRUCTION. The source is the last 300 CelebAMask-HQ images (ids $29{,}000$ and above), absent from every training and evaluation set here. Seven recipes (`left_eye`, `right_eye`, `eye_pair`, `brows`, `nose`, `mouth`, `inner_face`) take each region from the dataset's *own* human-annotated part mask, dilated by 6 px at 512×512, and the five without an annotated counterpart (both cheeks, forehead, chin and jaw, full face) are omitted. Kandinsky 2.2 decoder inpainting (30 steps, guidance 7.5, neutral prompt) re-synthesizes the region, and SDXL inpainting with the same settings is a second generator on a further 150 images. Pixels outside the dilated region are restored from the pristine image, confining the edit, so region truth is independent of our parser and our blend and the edit carries generator texture rather than a blend seam.

SCORING. As in the main protocol, each face is detected once on the pristine image and the same landmarks align every edit, so crops are pixel-registered, and 268 of 300 Kandinsky and 132 of 150 SDXL images pass the frontend. Region-AUC uses the pooled estimator with frame-level bootstrap intervals over images (half-width at most 0.035 for Kandinsky, 0.045 for SDXL). Coverage averages the six local recipes (inner-face control shown, not averaged), so only its ordering, not its value, is comparable with Table 2. Since our parser was trained on CelebAMask-HQ labels, PARTAWARE is also scored mask-free.

RESULT. In Appendix Table I.1 every published detector separates the inner-face edit almost perfectly (0.97–1.00, ProDet and FSFM excepted), handles the nose well, and falls toward chance on the single eyes (0.50–0.57), the size-graded failure seen on blends (Appendix Table H.1). PARTAWARE reaches 0.66 and 0.70 on the eyes and 0.762 coverage against 0.694 for GenD PE-L, the best published detector, and mask-free scoring gives 0.754, so the advantage is not the mask. The margin is smaller than on blends (0.068 against 0.119), the expected cost of an unseen manipulation family, but its sign and its concentration on small lateral regions carry over, whereas the published ordering does not, with GenD PE-L leading rather than GenD DINOv3-L. SDXL reproduces this, with PARTAWARE at 0.746 (0.743 mask-free) against 0.677 for GenD PE-L, the same leader, 0.70 on each eye against at most 0.57. ProDet is at or near chance on every recipe, FSFM on all but nose (0.703) and inner face (0.842), because their pristine scores span nearly the unit interval, so a small edit rarely shifts a score beyond the between-image spread.

NAMING AND LOCALIZING AN UNSEEN MANIPULATION. Neither head was trained on a diffusion edit. On the Kandinsky edits the part head ranks the edited region first in 0.639 of 1876 cases against a chance rate of 0.184 (SDXL 0.647 of 924, chance 0.183). Nose and inner-face edits are named almost always (0.996, 1.000) and single eyes least often (0.37 and 0.42, chance 0.07), with side accuracy 0.62 (0.73 on SDXL). The localization head, which reads no mask, reaches pixel AUC 0.937 (0.931) against the changed pixels, and its peak falls inside an edit covering 1.9% of the crop in 0.39 of cases (0.32), about twenty times the area rate. Both heads carry over to an unseen manipulation family, the part head below its accuracy on blends (0.693).

Table I.1: Region-AUC on diffusion-inpainted CelebAMask-HQ faces, region truth from the dataset's annotations, per recipe for Kandinsky 2.2 (n=268 images) and as coverage for SDXL (n=132, last column). Coverage averages the six local recipes (Inner shown, not averaged), best in bold and second best underlined.

Detector	Fam.	L-eye	R-eye	Eyes	Brows	Nose	Mouth	Inner	Cov.$^{\uparrow}$	Cov. SDXL$^{\uparrow}$
ProDet‡	Blend	0.506	0.501	0.511	0.503	0.476	0.529	0.500	0.504	0.507
FSFM	SSL	0.503	0.504	0.513	0.494	0.703	0.575	0.842	0.549	0.523
Effort	Adapt	0.551	0.558	<u>0.700</u>	0.647	0.830	0.728	0.992	0.669	0.659
ForensicsAdapter	Adapt	0.538	0.541	0.632	0.599	0.820	**0.760**	0.967	0.648	0.631
Unlocking-CLIP	Adapt	0.546	<u>0.565</u>	0.670	0.591	0.895	**0.760**	0.989	0.671	0.642
GenD CLIP-L‡	Adapt	<u>0.553</u>	0.560	0.653	0.616	0.884	0.714	0.982	0.663	0.639
GenD DINOv3-L‡	Adapt	0.512	0.514	0.546	0.586	0.951	0.680	0.989	0.632	0.629
GenD PE-L‡	Adapt	0.543	0.551	0.652	<u>0.709</u>	<u>0.964</u>	0.743	**0.999**	<u>0.694</u>	<u>0.677</u>
PARTAWARE	**Adapt**	**0.657**	**0.704**	**0.719**	**0.758**	**0.994**	0.741	**0.999**	**0.762**	**0.746**
PARTAWARE (mask-free)	**Adapt**	0.651	0.693	0.710	0.758	0.992	0.722	0.999	0.754	0.743

J HELD-OUT IN-THE-WILD EVALUATION ON DEEPFAKE-EVAL-2024

Backbone family and scale were chosen on the four transfer benchmarks of Appendix Table A.1 (Appendix F.3), so a lead measured there carries that optimism. Deepfake-Eval-2024 (Chandra et al., 2026) informed no decision and was scored once, after every configuration was fixed. It is 2024 social-media footage with unknown modern generators, upload compression, varied framing and multiple people. It is out of distribution on generator, capture and compression at once, and its own paper reports state-of-the-art detectors near chance on it. All 2036 labeled videos are scored under the protocol of every other benchmark (the dataset's terms permit evaluation only), with its train/test designation carried through and reported separately.

ACCEPTANCE. The shared frontend finds a usable face in 809 of 2036 videos (40%), far below the 81% of the transfer benchmarks, and acceptance is class-asymmetric, at 33% of real against 47% of fake videos. All detectors are scored on the same 809 (the eighteen published ones reuse PARTAWARE's crop cache), so the comparison is identical-set but restricted to videos where a face was found, while the excluded 60%, largely text, crowds or non-face content, is where a frame-level face detector cannot help.

RESULT. Appendix Table J.1 gives video AUC on all 809 videos and on the 325 test-designated ones. Every detector is near chance, as the dataset's authors report for the field, and PARTAWARE ranks first on both (0.624, 0.645), ahead of Effort (0.610) on the full set and Xception (0.624) on the test split, by margins inside their paired bootstrap intervals, +0.013 (95% CI −0.018 to +0.046, 2000 resamples over videos) and +0.020 (−0.040 to +0.077). The ordering of Table 1 therefore holds on data no decision touched and the lead's sign is not a selection artifact, though no significant margin is established or claimed. LAA-Net and FakeFormer rank last and the order within the adaptation family is not preserved (Effort and Unlocking-CLIP lead it), consistent with those detectors sitting within noise of each other.

Table J.1: Deepfake-Eval-2024 video AUC on the 809 videos in which the shared frontend finds a face and on the 325 of them in the test designation, an identical set for all nineteen detectors, sorted by the first column. Best in bold, second best underlined. ‡ marks training beyond FaceForensics++ c23.

Detector	Family	Ref.	AUC, all (n=809) ↑	AUC, test (n=325) ↑
PARTAWARE (ours)	**Adapt**	–	**0.624**	**0.645**
Effort	Adapt	(Yan et al., 2025)	<u>0.610</u>	0.607
Unlocking-CLIP	Adapt	(Yermakov et al., 2025)	0.606	0.596
GenD PE-L‡	Adapt	(Yermakov et al., 2026b)	0.595	0.621
GenD CLIP-L‡	Adapt	(Yermakov et al., 2026b)	0.586	0.593
GenD DINOv3-L‡	Adapt	(Yermakov et al., 2026b)	0.577	0.604
Xception	CNN	(Chollet, 2017)	0.572	<u>0.624</u>
ForensicsAdapter	Adapt	(Cui et al., 2025)	0.554	0.564
EfficientNet-B4	CNN	(Tan & Le, 2019)	0.551	0.582
FSFM	SSL	(Wang et al., 2025)	0.547	0.616
SPSL	Freq	(Liu et al., 2021)	0.535	0.576
RECCE	CNN	(Cao et al., 2022)	0.528	0.576
CORE	CNN	(Ni et al., 2022)	0.527	0.584
UCF	CNN	(Yan et al., 2023a)	0.525	0.584
F3Net	Freq	(Qian et al., 2020)	0.520	0.574
ProDet‡	Blend	(Cheng et al., 2024)	0.519	0.532
SRM	Freq	(Luo et al., 2021)	0.515	0.587
LAA-Net‡	Blend	(Nguyen et al., 2024b)	0.491	0.506
FakeFormer‡	Blend	(Nguyen et al., 2024a)	0.467	0.466

K ZERO-SHOT TRANSFER TO REAL LIPSYNC FORGERIES

Coverage is measured on our own construction, so the strongest test of the training signal is a manipulation we did not construct that, like P-SBI, confines each edit to one region. Lip-synchronization generators are the only such family with public benchmarks, since they edit only the mouth, are absent from FaceForensics++ training and are distributed as evaluation sets, whereas face swaps and reenactment alter the whole face. We evaluate on the FakeAVCeleb wav2lip subset (500+500) (Khalid et al., 2021), AVLips (400+400) (Liu et al., 2024), and the identity-disjoint test splits of DeepSpeak v1 (811+811, wav2lip and video-retalking) and v2 (469+469, Diff2Lip and LatentSync) (Barrington et al., 2026), all as distributed. One variant targets this family with a **cross-frame** recipe, in which the `mouth` donor comes from another frame of the same video without warping, so the injected artifact is purely a mismatch in time (the "cross-frame mouth donor" row of Appendix Table F.2). Appendix Table K.1 scores all nineteen detectors on identical videos (acceptance in its caption). On FakeAVCeleb's wav2lip forgeries PARTAWARE (0.9673) ties GenD DINOv3-L (0.9677), while the

CLIP-L adaptations fall to 0.79–0.88 there yet reach 0.93–0.98 on DeepSpeak's own wav2lip fakes, so that gap is attributable to FakeAVCeleb's footage, not the generator. On AVLips the detectors span 0.60 to 0.79, led by GenD PE-L (0.7891), with PARTAWARE at 0.7616. On DeepSpeak v2, built with modern generators, PARTAWARE (0.929) and GenD DINOv3-L (0.925) tie at the top, every CLIP-L adaptation is at or below 0.83, and LatentSync is the harder generator for eighteen of nineteen detectors.

Table K.1: Zero-shot lipsync detection (video AUC). FakeAVCeleb-wav2lip and AVLips contribute the 610 and 511 videos with a detected face (61% and 64% of those distributed, 303/307 and 259/252 fake/real), and DeepSpeak (Barrington et al., 2026) the 2409 of 2560 distributed v1 and v2 identity-disjoint test videos (94%), by generator. Every detector scores the same videos, best in bold and second best underlined. ‡ marks training beyond FaceForensics++ c23.

				DeepSpeak v1			DeepSpeak v2		
Detector	**Fam.**	**FakeAVCeleb**↑	**AVLips**↑	All	Retalking	Wav2Lip	All	Diff2Lip	LatentSync
Xception	CNN	0.8029	0.7085	0.8988	0.8649	0.9370	0.8793	0.9094	0.8582
EfficientNet-B4	CNN	0.7834	0.6291	0.8041	0.7214	0.8978	0.6960	0.7195	0.6794
CORE	CNN	0.8539	0.7388	0.8623	0.8034	0.9290	0.8205	0.8860	0.7744
RECCE	CNN	0.7643	0.7322	0.9424	0.9139	0.9746	0.8404	0.8678	0.8212
UCF	CNN	0.7841	0.6705	0.9032	0.8685	0.9425	0.8791	0.9030	0.8622
F3Net	Freq	0.8113	0.6555	0.8572	0.7999	0.9220	0.7580	0.8155	0.7175
SPSL	Freq	0.7279	0.5970	0.7680	0.6948	0.8508	0.7897	0.8343	0.7582
SRM	Freq	0.8439	0.6773	0.8773	0.8249	0.9365	0.7922	0.8265	0.7681
LAA-Net‡	Blend	0.9098	0.6635	0.6707	0.4920	0.8730	0.7003	0.8164	0.6186
FakeFormer‡	Blend	0.8651	0.6534	0.7342	0.5609	0.9304	0.7049	0.7387	0.6811
ProDet‡	Blend	0.8470	0.6716	0.8454	0.7696	0.9312	0.7933	0.8150	0.7780
FSFM	SSL	0.7687	0.7544	0.8677	0.8015	0.9427	0.7237	0.7109	0.7327
Effort	Adapt	0.8193	0.6719	0.8182	0.7224	0.9266	0.6621	0.8096	0.5583
ForensicsAdapter	Adapt	0.7890	0.7077	0.8643	0.7995	0.9377	0.7922	0.8687	0.7384
Unlocking-CLIP	Adapt	0.8779	0.7561	<u>0.9501</u>	<u>0.9242</u>	0.9794	0.8249	0.9112	0.7643
GenD CLIP-L‡	Adapt	0.8835	0.7105	**0.9505**	**0.9243**	<u>0.9802</u>	0.8289	0.9032	0.7766
GenD DINOv3-L‡	Adapt	**0.9677**	0.7580	0.9204	0.8769	0.9695	<u>0.9251</u>	<u>0.9668</u>	<u>0.8957</u>
GenD PE-L‡	Adapt	0.9186	**0.7891**	0.9041	0.8605	0.9534	0.8726	0.9194	0.8396
PARTAWARE	**Adapt**	<u>0.9667</u>	<u>0.7615</u>	0.9379	0.8943	**0.9873**	**0.9331**	**0.9798**	**0.9002**

Covering all eighteen published detectors adds two findings. The whole-face blending detectors are the weakest of the nineteen on DeepSpeak v1 retalking, with LAA-Net at 0.492, below chance, and FakeFormer 0.561, against 0.69–0.92 for the rest. Both are nonetheless competitive on FakeAVCeleb (0.910, 0.865), so whole-face blend supervision does not transfer to a mouth-confined edit, the failure P-SBI is designed to avoid. Nor does recency order this benchmark, as Xception, from 2017, reaches 0.879 on DeepSpeak v2, third of nineteen and ahead of every CLIP-L adaptation, while Effort reaches 0.662. Lipsync transfer tracks neither backbone scale nor recency, and is reported apart from the corruption sweep.

ALTERNATIVE CONFIGURATIONS TRADE AVLIPS AGAINST DEEPSPEAK V2. The three configurations most likely to beat PARTAWARE on a region-confined generator were also scored. They reach 0.7865 (rank-32 DINOv3-L), 0.7829 (cross-frame mouth donor) and 0.7812 (rank-16 DINOv3-H+) on AVLips, against 0.7616 for PARTAWARE, and 0.9049, 0.9087 and 0.9571 on DeepSpeak v2, against 0.9290. Only the rank-16 configuration leads on both (Appendix F.4).

K.1 LIMITS OF THE CROSS-FRAME RECIPE ON REAL LIPSYNC FORGERIES

The cross-frame result (Appendix K) is informative as a partial success. The cross-frame mouth donor teaches the intended local mechanism, reaching the best mouth-region coverage measured here (0.831), yet does not close the gap on AVLips (0.7829), built from a real, modern lipsync generator. The recipe injects *temporal content mismatch*, namely a mouth from another moment composited without warp, a real artifact family but not the one a lipsync generator leaves. A generator conditioned on audio leaves artifacts of generation itself, such as spectral inconsistencies in the synthesized region, jitter correlated with phoneme timing and blending specific to its compositing step. The result separates an axis a synthetic signal can move (mouth-region *localization*) from one it does not (detecting a *particular deployed generator*), and leaves the concrete open problem of training a signal that transfers to AVLips-style forgeries.

L In-Distribution AUC, Attribution Faithfulness and the Localization Head

Scale alone does not reach the observed coverage. Coverage is *not* independent of capacity, since across the eighteen published detectors it correlates strongly with cross-dataset transfer ($r = +0.86$, $R^2 = 0.75$, with coverage $= 0.143 + 0.532\times$ transfer) and with zero-shot corrupted AUC ($r = +0.76$, $R^2 = 0.58$), as any fit to Appendix Tables A.1 and A.2 shows. The slope is too shallow to matter, since at our transfer of 0.980 the fit predicts coverage 0.664 against a measured 0.795, a residual of $+0.130$ or 8.4 residual standard deviations, and $+0.130$ (6.5 s.d.) on the robustness fit. Reaching 0.795 on either trend requires a transfer AUC of 1.22 or a corrupted AUC of 1.27, both unattainable (Appendix Figure L.1). Scale raises coverage but, on this trend, does not reach the observed value, although no controlled scaling study was run.

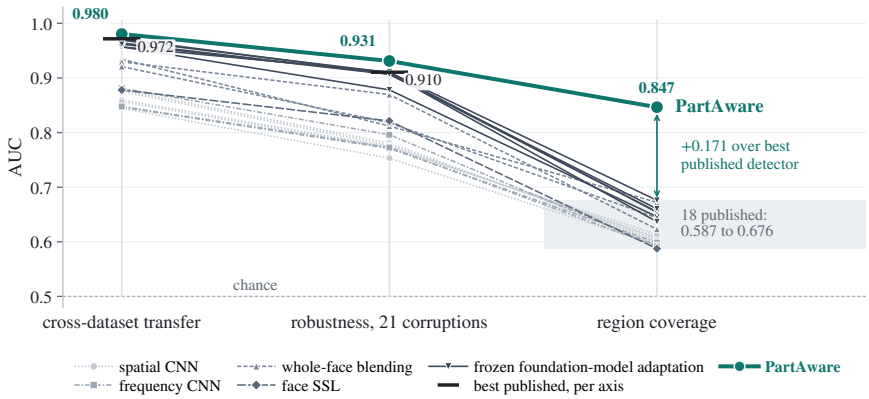

Figure L.1: **Detection is saturated but localization is not.** One line per detector across three capabilities on one AUC axis (transfer and robustness at video level, coverage at frame level, with 0.50 chance on each). The eighteen published detectors spread on transfer (0.844–0.972) and zero-shot robustness (0.753–0.910), then converge into a band 0.09 wide on coverage (0.587–0.676). PartAware (0.980, 0.931, 0.847, where transfer and coverage are means of three seeds and robustness is seed 123) leads on all three, by 0.008, 0.021 and 0.171 over the second-best detector named under each axis, so the first two show that the region signal costs nothing and the third is what it adds.

Robustness and coverage come from different sources. GenD DINOv3-L shares the backbone of our DINOv3-L configuration but trains without P-SBI. It nearly matches that configuration on corruption robustness (0.9469 against 0.9505) yet falls well short on coverage (0.676 against 0.821, Appendix Table F.3), and the frozen and no-P-SBI controls of Appendix Table F.2 show the reverse. With respect to RQ3, robustness derives from the pretrained representation and coverage from the training signal, although the two models also differ in tuning, so this bounds the effect rather than isolating it.

Boundary supervision does not supply region sensitivity. ForensicsAdapter, the strongest published transfer baseline and trained with pixel-level boundary supervision, reaches only 0.637 coverage. A boundary marks *that* a seam exists without naming its region, which would explain the gap, although we have not tested this directly.

L.1 In-Distribution AUC as an Anti-Signal for Deployment

Effort, with the lowest in-distribution score of its family and competitive transfer, instantiates a general pattern. Across the full comparison (Appendix A), 13 of 19 detectors fall within a 1.7-point in-distribution band while spreading over thirteen points of transfer (Spearman $\rho = -0.49$ over the eighteen published detectors, Section 5.1), so ranking by in-distribution score is close to random with respect to deployment. This reflects any benchmark optimized until only its own artifacts remain to fit, not a limit on generalization, and GenD's finding that diverse training data yields benchmark-general transfer (Yermakov et al., 2026b) is complementary. In-distribution AUC belongs in a sanity check,

not a ranking criterion, and headline comparisons should use the slower-saturating axes, namely cross-dataset transfer and, on the evidence of Sections 5.2 and 5.3, region coverage and corruption robustness.

L.2 Faithfulness and Oracle Assistance of the Shapley Attribution

Appendix Table L.1 separates the three ways this paper asks which region was edited, namely two post-hoc interventions, occlusion (the score drop when one part alone is restored) and Shapley attribution, which restore pixels from the pristine frame and cannot run without it, and the model's own part head, which reads the edited image alone. The floor rows show how much of a post-hoc score this protocol concedes to a statistic carrying no anatomy at all.

Table L.1: Three ways of asking which region was edited, on the 200×12 blends of Table 2. The floor is the parameter-free statistic of Table 2, carrying no anatomy, and the whole-face control's part head never saw a confined region. Top-1 chance is $1/12$ and side chance is 0.5. Best in bold, second best underlined.

Readout	**Pristine frame**	**Top-1**$^{\uparrow}$	**Side acc.**$^{\uparrow}$
Floor, a parameter-free statistic without anatomy			
occlusion	required	0.732	0.854
Shapley	required	0.706	0.828
PartAware			
occlusion	required	<u>0.812</u>	0.902
Shapley	required	**0.857**	**0.925**
part head, one forward pass	not required	0.693	<u>0.920</u>
Whole-face control			
part head, one forward pass	not required	0.318	0.596

Table L.2: The dense localization head against the exact blend mask on the 200×12 blends of Table 2, with IoU and Dice at threshold 0.5, pointing as the map maximum falling inside the edited region, and mass as the map fraction inside it. The bottom block gives PartAware per recipe, with region area as a fraction of the crop.

Model	**Pixel AUC** ↑	**IoU@0.5** ↑	**Dice** ↑	**Pointing** ↑	**Mass in region** ↑
whole-face control (DINOv3-B)	0.879	0.143	0.206	0.220	0.152
base configuration (DINOv3-B)	0.916	0.152	0.208	0.451	0.225
PartAware	0.927	0.192	0.266	0.474	0.197
Per recipe PartAware					
left eye (0.4%)	0.970	0.122	0.177	0.459	0.032
right eye (0.4%)	0.969	0.132	0.188	0.480	0.034
eye pair (0.7%)	0.949	0.030	0.048	0.251	0.026
brows (1.2%)	0.909	0.034	0.059	0.158	0.036
nose (1.8%)	0.918	0.141	0.216	0.390	0.074
mouth (1.2%)	0.973	0.100	0.151	0.319	0.047
forehead (11.4%)	0.815	0.188	0.247	0.452	0.264
left cheek (6.1%)	0.935	0.420	0.531	0.720	0.308
right cheek (6.4%)	0.923	0.396	0.503	0.651	0.300
chin and jaw (3.7%)	0.864	0.117	0.192	0.200	0.101
inner face (3.9%)	0.939	0.188	0.294	0.534	0.152
full face (44.1%)	0.962	0.376	0.504	0.980	0.890

Two properties support reading the result as a capability rather than an artifact of the construction. Efficiency, that per-region attributions sum exactly to $f(\text{full}) - f(\text{all regions ablated})$, is an identity satisfied numerically here, which rules out accounting errors in the sums though not errors in labels or region indexing. The held-out-operator control (Section 6) shows operator-specific signal to matter only at small capacity, which also answers the critique that self-blending-trained detectors reduce to finding the signature of alpha compositing (Yermakov et al., 2026a), since a detector holding only that signature would fail it. Side accuracy cannot support the argument, for the reason below.

ATTRIBUTION UNDER PRISTINE RESTORATION IS ORACLE-ASSISTED. Restoring an unedited region changes nothing, so every unedited region has zero marginal contribution and the edited one carries the whole score change. In the spirit of saliency sanity checks (Adebayo et al., 2018), two scorers with no learned parameters and no anatomy test the consequence, namely mean face-masked high-frequency energy, and one minus that energy. The first falls on a feathered edit (coverage 0.472) and names the region 0.339 of the time at side accuracy 0.172. The second is barely a detector (coverage 0.528, worst region 0.493) yet attributes at top-1 0.706, top-3 0.861 and side accuracy 0.828 (occlusion top-1 0.732), within the published range (0.64–0.81 and 0.71–0.91) and close to PARTAWARE's 0.857 and 0.925. Top-1 and side accuracy therefore measure whether a score responds with the right sign, not whether a detector can locate the edit, and for a k-part recipe have chance level $k/14$. The direct tests are the part head (Table 2) and the localization head (Appendix L.3).

L.3 DIRECT EVALUATION OF THE DENSE LOCALIZATION HEAD

Attribution needs a pristine counterpart, whereas the localization head does not. Appendix Table L.2 scores the localization map against the exact blend mask on the 200×12 blends of Table 2, from one image. It ranks pixels inside the edited region above those outside at AUC 0.927 (0.97 on single eyes, 0.973 on the mouth) and its maximum falls inside the region 0.47 of the time, but it is a diffuse map rather than a segmentation, since IoU at 0.5 is 0.19 against regions covering 0.4–44% of the crop. Region naming therefore rests on the part head (0.693 top-1, and 0.808 and 0.816 for the other two seeds, Appendix Table L.1), the map being supporting evidence. The whole-face control points inside the region 0.22 of the time against 0.45 for the base configuration, so confinement teaches the map where to concentrate, and every number is unchanged under mask-free scoring.