CROSS-TASK GAINS WITHOUT AUXILIARY LABELS: ARCHITECTURE-PRESERVING ABLATIONS UNDER DISJOINT SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

When two tasks are labelled on different datasets and no example carries both, coupling them through a shared module often improves the target task, a gain credited to the auxiliary labels. But the module also changes the architecture available to the target task, so comparing coupled with separate heads cannot separate supervision transfer from architectural benefit, and none of the fourteen multi-objective methods and systems we audit reports a control that can. We resolve this attribution problem with architecture-preserving label ablations, which need no new model or data: the coupling module and its readout are kept while the auxiliary labels are removed, the module is frozen at random initialization, or the auxiliary task is dropped with the readout kept. A gain that survives without the labels is architectural by construction, and the observed gain splits exactly into architectural and label parts; an exact single-task tie certifies the split. Our testbed is vision-only deepfake localization, where spatial masks and temporal boundaries come from separate corpora, with frozen features and cross-task attention. The control overturns the default reading. Attention improves official AP@0.5 at every seed, by $+0.025$ at ten-fold budget, yet 86% of the pilot gain survives with patch access but no masks (and most of it on a second encoder and dataset pair). At ten-fold, attention without masks matches or exceeds full attention at every seed on the main encoder, and even with the spatial task dropped and the readout kept, most of the gain remains at both budgets (both claims post hoc). A parameter-matched module with masks but no patch access reaches half of the pilot gain; the tested output-space consistency term, a loss through which mask labels do reach the temporal head, gives no consistent advantage. The gain shrinks with budget, is not reliably retained under fine-tuning, and does not predict better localization of unseen manipulation types. A cross-task gain under disjoint supervision should pass this control before it is credited to labels; we release it with scoring code.

1 INTRODUCTION

Multi-task learning from disjoint supervision is how heterogeneous annotation is combined in practice: one dataset supervises one task, another a second, and no example carries both labels. Methods that couple predictions, align features, or distil task experts under such partial supervision (Li et al., 2022; Spinola et al., 2023; Lin et al., 2026; Sun et al., 2026) improve the target task, and the gain is read as supervision transfer from the auxiliary labels. That reading is hard to test: the auxiliary labels reach the target task through the coupling module, and the module also changes the architecture available to it, so coupled against separate heads changes two things at once, and none of the fourteen multi-objective methods and systems we audit reports that comparison for a second label source (Section 2). The answer decides which annotations to pay for.

We investigate this distinction in vision-only deepfake localization. Both label types are then read from one visual feature space, so the coupling module is the only path between the tasks; audio would give the temporal task a private input that no label ablation removes. A forensic system should localize both the manipulated region and its temporal extent, but masks and segment annotations are expensive and usually live in separate corpora. We combine FaceForensics++ masks (Rössler et al., 2019) with AV-Deepfake1M temporal boundaries (Cai et al., 2024). Our frozen-backbone testbed

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.

compares summed losses, output-space consistency, and feature-space cross-task attention in the OmniFD lineage (Liu et al., 2025a).

We study the frozen regime deliberately: frozen foundation features are the common way to add a task cheaply, and it is where such claims are cheapest to make and hardest to audit, since only heads and couplers train. It also makes the control exact: our architecture-preserving label ablations keep the coupling module and its readout while removing the auxiliary labels, freeze the module at its random initialization, or drop the auxiliary task with the readout kept, and because a frozen backbone gives the auxiliary loss no other path to the target head, the observed gain splits exactly into an architectural and a label part (Section 3.2), for the cost of one retrained arm.

The control overturns the default reading. Attention improves official temporal AP at every seed, yet each step that cuts the labels' route but not patch access leaves most of the gain. Attention without mask labels (the label-free arm) keeps most of it at the pilot budget, and at ten-fold (post hoc), is at or above full attention at every seed on the main encoder. Without the spatial task, a same-readout model (post hoc) keeps most of it at both budgets. Even frozen at random initialization, the module keeps much of it. The label part changes sign across pilot seeds. Without patch access, a parameter-matched module with masks reaches half. The distinction matters beyond attribution: the same in-domain improvement does not predict better localization of unseen manipulation types.

Our contributions are:

1. A control that resolves the attribution problem: architecture-preserving label ablations (labels removed, which none of the fourteen audited methods and systems reports; module frozen at random initialization; or a same-readout single-task model); a gain that survives without the labels is architectural by construction, and an exact tie with the single-task floor certifies the split. Under frozen features, cross-task attention keeps most of its gain with patch access but no mask supervision and, at both budgets, without the spatial task; a parameter-matched module with masks but no patch access reaches half of it (Table 2).
2. A regime map: the label-free gain holds at both budgets, on a second frozen encoder and a second temporal dataset, but is not reliably retained under backbone fine-tuning; a generalization audit (near-floor zero-shot transfer does not preserve the in-domain ranking; a teacher probe finds a source-associated response in the output-space coupling channel): a regime-dependent result, not a general failure of supervision transfer.

2 Related work

2.1 Multi-task learning from disjoint supervision

Cross-task consistency began as an output-space constraint (Zamir et al., 2020), extended to partial annotations by MTPSL (Li et al., 2022); later work pseudo-labels missing annotations via task tokens (Zhang et al., 2025) or across disjoint datasets (Spinola et al., 2023), aligns task features across domain partitions (Lin et al., 2026), or distils disjoint-dataset per-task experts into one backbone (Sun et al., 2026). Well-tuned summed-loss baselines (Kurin et al., 2022; Xin et al., 2022) motivate careful controls. Closest in structure to our label-free finding, Xu et al. (2026) show that a trainable target-side encoder over frozen source features suffices to prevent negative transfer; we ask the converse of a positive gain, whether it needs the auxiliary labels at all. Task-grouping studies measure whether one task's supervision helps another under shared parameters, by training with and without it (Standley et al., 2020) or from gradient affinity (Fifty et al., 2021); our label ablation holds the coupling architecture fixed: the loss goes, the data and the module it feeds stay. None of the fourteen multi-objective methods and systems we audit reports that comparison for a second label source; the nearest cases ablate an auxiliary loss over the same labels or none, or keep frozen random constraint networks in a loss-side coupling (Zamir et al., 2020) (Table 47).

2.2 Benchmark validity and rank transfer

Patel et al. (2026) argue that aggregate-score rankings of LLM-agent benchmarks do not transfer out of distribution and propose reporting predictive validity. We add a vision-forensics instance: validation rankings anti-correlate with joint where-and-when rankings and invert under zero-shot transfer, so we report per-readout rank-transfer statistics, not one leaderboard.

2.3 JOINT AND UNIFIED FORGERY MODELS

OmniFD (Liu et al., 2025a), closest to our feature-coupling arm, trains vision-only spatial-mask and temporal-segment heads jointly on ForgeryNet's disjoint image and video subsets: feature-space coupling within one dataset family (its LAV-DF row is a single-task instance trained on LAV-DF). Its released model zero-shot and its recipe retrained on our pair both appear as context rows in Table 1. Venus-DeFakerOne (GuangJian Team, 2026) unifies detection and localization at image level; DDL (Miao et al., 2025) and FakeParts (Liu et al., 2025b) ship disjoint-subset or subtype-level spatial and temporal labels; audio-visual systems (Anshul et al., 2025; Zheng et al., 2025; Koutlis & Papadopoulos, 2026; Li et al., 2026) pair detection with temporal localization; GEM-TFL (Zhu et al., 2026b) bridges label granularity within one dataset. Our distinction from OmniFD is the label-removal and regime analysis, not the existence of disjoint spatial and temporal supervision; our results concern a controlled frozen-backbone implementation and do not refute its end-to-end results. UVIF (Liu et al., 2026) improves partially forged video detection with annotated images via pseudo-labeling and feature alignment, studying detection rather than our joint localization readout.

2.4 TEMPORAL LOCALIZATION ON AV-DEEPFAKE1M

Audio-visual localizers lead AV-Deepfake1M at 92–97 AP@0.5 (Zhu et al., 2026a; Koutlis & Papadopoulos, 2026; Wang et al., 2026); published vision-only rows (Table 1) range from ActionFormer on VideoMAEv2 (Wang et al., 2023; Cai et al., 2024) to TE-TAD and TadTR as tabulated by DeformTrace (Kim et al., 2024; Liu et al., 2022; Zhu et al., 2026a) (provenance and backbones unstated), and GEM-TFL's retrained visual-only ActionFormer reaches 67.0 on ResNet-50 features. Two systems tabulate no vision-only row above VideoMAEv2's (Ni et al., 2026; Yu et al., 2026), one inferring a fundamental visual-only capability gap; our anchor, the same ActionFormer head on InternVideo2-1B features, shows stronger visual features substantially narrow it (Appendix A.7).

3 EXPERIMENTAL SETUP

3.1 DATA AND MODEL

No training clip carries both label types. The spatial task is supervised by **FaceForensics++** (Rössler et al., 2019) at c23 (per-frame masks whose geometry differs by family): 8 frames of each of the 2,880 manipulated training videos and 4 of its original (360 identity pairs in both orders, four methods; one third real; about 34,560 images at 224×224), official identity splits. The temporal task is supervised by the **visual stream** of **AV-Deepfake1M** (Cai et al., 2024): TalkLip lip-sync manipulations with ground-truth segments, audio discarded (threat model: Appendix A.4), clips labelled by their visual edits (`real`, `visual_modified`, `both_modified`; audio-only edits excluded), 32 frames at 224×224. Unless stated otherwise, runs use a pilot budget of 1,000 mixed batches $\times$ 10 epochs drawn from the 559,607-clip temporal pool, reshuffled each epoch; $3\times$ and $10\times$ multiply that budget, and because the spatial loader is exhausted first, the added batches are mostly temporal.

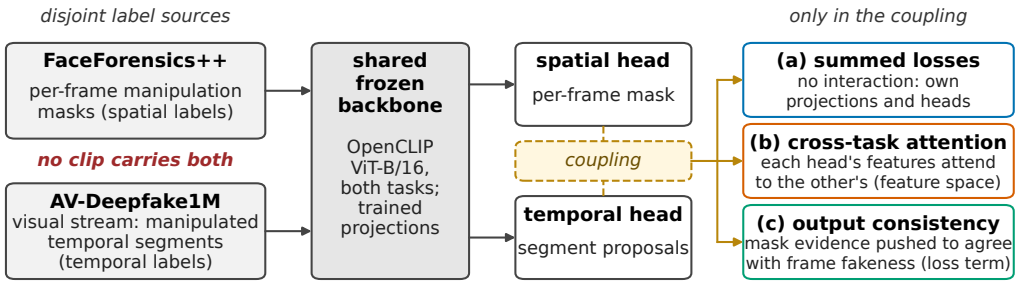


Figure 1: Study design (details in Section 3). The arms differ only in the coupling: (a) summed losses, (b) feature-space cross-task attention, (c) an output-space consistency term; backbone, heads, sampler and budget are shared, and no training clip carries both label types.

A **frozen OpenCLIP ViT-B/16** (Figure 1) image encoder, pre-trained on LAION-2B (Cherti et al., 2023; Schuhmann et al., 2022) and fine-tuned on ImageNet-12k/1k (timm `vit_base_patch16_clip_224`, weights `laion2b_ft_in12k_in1k`), embeds every frame; only two 256-d projections, the two heads and any coupling module train. The **spatial head** decodes per-frame masks from projected patch tokens; the **temporal head**, a BMN-style boundary-matching localizer (Lin et al., 2019), reads one pooled vector per frame (patch-token mean, max and standard deviation, with the CLS token) and emits up to 16 scored proposals per clip. FaceForensics++ images enter it as static 8-frame clips, so both heads see every batch. The sources interleave 50/50 (32 images or 8 clips per batch) with matched seeds (identical data order within one architecture); segments use the fused start · end · boundary-map decode of Lin et al. (2019), fixed across arms and readouts (exploratory seed-42 runs excepted; Appendix A.1).

3.2 Coupling mechanisms and controls

The arms change the feature coupling or add an output-space loss; because feature coupling changes parameterization and feature access, the controls below are essential for attribution:

- **Naive sum (control):** no shared trainable parameter (frozen backbone, own projections).
- **Cross-task attention:** a 4-head module lets each frame's pooled vector attend over that frame's projected patch tokens ($s{\to}t$, *patch access*) and broadcasts temporal features back to the spatial branch ($t{\to}s$) before decoding (the OmniFD family; Liu et al., 2025a).
- **Output-space consistency ($w = 0.1$; Zamir et al., 2020):** peak mask evidence (top-2% pixel mean, threshold $\tau = 0.3$) teaches the temporal head's frame fakeness on FaceForensics++ images; on AV-Deepfake1M clips, anchored on ground-truth frame labels, mask mass is penalized on real frames, its per-clip mean on fake frames must exceed that on real frames by a 0.2 margin (removing the empty-mask optimum), and adjacent real frames are smoothed.
- **Controls:** parameter-matched self-attention within each task (no patch access); a temporal-only single-task floor (naive, spatial loss weights zero) that any multi-task claim must beat (Wu et al., 2020); attention without mask labels (spatial loss zeroed); consistency without its real-mass anchoring term; and, post hoc, $s{\to}t$ attention only (patch access kept), with mask labels and without them; the label-free one has no trainable path from the spatial task, so it is a single-task model with patch access (the *single-task attention readout*).

Two further arms probe the space between: a cross-dataset pseudo-label arm (FaceForensics++ images take the *ground-truth* mask's verdict as their temporal target; an EMA teacher supervises masks on the temporal clips) and a soft reformulation of the consistency term. All arms share one loader, configuration schema and code path.

Attribution by ablation. Let $A(M, L)$ be the target-task score with the coupling module absent ($M{=}N$), trained ($M{=}C$) or frozen at random initialization ($M{=}C_0$), and the auxiliary loss on ($L{=}1$) or off ($L{=}0$). With a frozen backbone and no shared trainable parameter, that loss cannot reach the target head when $M{=}N$, so $A(N, 1) = A(N, 0)$ is a testable identity (Section 5.2 verifies it to the last digit) and the gain decomposes exactly,

$$A(C,1) - A(N,1) = \underbrace{A(C,0) - A(N,0)}_{\Delta_{\text{arch}}} + \underbrace{A(C,1) - A(C,0)}_{\Delta_{\text{labels}}}, \tag{1}$$

so the label-free arm identifies Δ_{arch} and the *retained fraction* $\Delta_{\text{arch}}/(A(C,1) - A(N,1))$ is the statistic a label claim must survive; $A(C_0, 1)$ bounds what needs a trained module, the single-task attention readout what needs the auxiliary task at all. The decomposition is operational (per matched seed, seed spread as uncertainty; Section 4): it measures what the auxiliary loss changes here, not information transfer.

3.3 Same-clip reference and context systems

FF++-stitched clips concatenate real and fake segments of one identity, so masks and segments share a clip (the temporal stitching of Saha et al., 2023, extended with masks); they serve as a supervision dose (0%, 25%, 100% of training batches) and a same-clip reference (Appendix A.5);

controls: stitched-only training, real-real stitch negatives, boundary-masked scoring. For the temporal task's range, Table 1 adds ActionFormer (Zhang et al., 2022) on frozen InternVideo2-1B features (Wang et al., 2024), DiGIT (Kim et al., 2025) and published rows.

4 EVALUATION PROTOCOL

Pre-registration. Reading rules for metrics, checkpoints, comparisons, thresholds and controls were recorded in dated notes before the corresponding results; Appendix A.8 gives the dates and lists the dated amendments and exceptions.

Metrics and scopes. In-domain temporal localization is segment AP at tIoU 0.5 (AP@0.5) on held-out AV-Deepfake1M validation under one fixed implementation and decode rule (Section 3.1), each budget tabulated in a separate block; that split shares its speakers and source recordings with training (Appendix A.7). One rule governs every table: rows are compared only within a block that shares training budget, backbone, metric family and protocol; blocks are printed side by side for scale and never ranked against each other, and rows for other systems are reported under their own protocol, never ranked against ours. Checkpoints are selected on that same validation window (capped at 8,000 clips at the pilot budget, the full 43,054 at ten-fold; Appendix A.1). Table 1 scores exported proposals with the official evaluation code (a recorded round-trip check reproduced the direct decode exactly from the exported proposals); internal numbers never share a column with published ones. The joint where-and-when readout scores 286 manipulated and 100 real validation clips against landmark-derived mouth-region pseudo-masks (tube-mAP@0.2); these weak references may bias both absolute scores and rankings. Transfer is strictly zero-shot (no target examples, tuning, calibration or threshold search): *(i) TVIL* (Zhang et al., 2023), AP@0.5 with AR@10 as the designated secondary; *(ii) DF40* (Yan et al., 2024), 39 of its 40 generators in four paradigms (37 readable), video-level detection AUC under a score mapping frozen on FaceForensics++ validation before any DF40 video was scored, reals being FaceForensics++ test originals (an asymmetry stated wherever DF40 numbers appear), pooled over the three *face-local* paradigms (swap, reenactment, editing; whole-image synthesis has no local region), and read as descriptive patterns, not transfer claims, wherever that mapping is near chance on FaceForensics++ validation (*evidence* AUC ≤ 0.55); *(iii) OpenForensics* (Le et al., 2021), whole-image forged-face union-mask IoU. Definitions and thresholds, including the in-domain mask metrics: Appendix A.10.

Seeds and uncertainty. Three matched seeds with identical data order within an architecture make comparisons paired: we report mean $\pm$ sample sd, paired same-seed deltas with their sign pattern, and 95% percentile bootstrap intervals ($B = 1000$; Appendix A.10), which condition on the trained weights and do not bound run-to-run spread (Section 6.1). Each transfer readout carries an **inversion statistic**: the probability, over resamples and seeds, that the official-harness in-domain leader also leads under the readout (exact ties, none occurring, counted as retained), computed per block with per-seed fractions; low values are inversions, and it is descriptive, not a test.

5 RESULTS

Budgets (pilot, $10\times$): Section 3.1. Temporal and label-control results use official AP@0.5; internal sampled-grid AP: Appendix A.1.

5.1 IN-DOMAIN GAINS AND THEIR DEPENDENCE ON TRAINING REGIME

Budgets and checkpoint selection. Attention leads naive at every matched seed and tested budget. The largest budget is unconverged: two attention runs and one naive run were still improving at their final epoch. The decreasing gap is consistent with an optimization or sample-efficiency advantage but does not establish its asymptotic value. Last-epoch selection (post hoc) preserves the attention-minus-naive signs across budgets and encoders; the second dataset pair is less robust (Appendix A.1).

Official harness. Re-scored with the official evaluation code, attention's paired gain over naive shrinks monotonically with budget — $+0.070$ AP@0.5 at pilot, $+0.053$ at three-fold, $+0.025$ at

Table 1: Temporal localization AP@0.5 on AV-Deepfake1M, official evaluation code; every row we score uses the same 57,340-record split (upstream validation), second AP column its 43,054-record visual subset (records without an audio-only edit: 28,819 visually manipulated and 14,235 real). *Top:* our joint arms, ten-fold budget, frozen OpenCLIP ViT-B/16 (Section 3.1), mean ± sd over three seeds and paired same-seed deltas; the no-mask-label row is post hoc; sign counts are positive seeds on the full split (visual subset in brackets where different). *Bottom:* context rows under different protocols (Section 4); the InternVideo2-1B anchor's backbone gap is expected (Appendix A.7); "vis. GT": visually-edited segments only, "all GT": every annotated segment; the two OmniFD rows differ only in training (Appendix A.7).

Joint arm (10×, frozen OpenCLIP ViT-B/16)	AP@0.5 full (0–1)	AP@0.5 vis. subset (0–1)	n / sign
naive (summed per-task losses)	0.146 ± 0.008	0.219 ± 0.011	3
cross-task attention	0.171 ± 0.001	0.256 ± 0.002	3
attention, no mask labels	0.181 ± 0.010	0.270 ± 0.015	3
consistency regularizer (w=0.1)	0.145 ± 0.007	0.218 ± 0.010	3
paired: attention − naive	$+0.025 \pm 0.008$	$+0.037 \pm 0.012$	3/3
paired: attention − no mask labels	-0.010 ± 0.009	-0.014 ± 0.014	0/3 (1/3)
paired: regularizer − naive	-0.001 ± 0.004	-0.001 ± 0.006	2/3

Context row (own protocol)	Backbone / protocol	AP@0.5 (%)	vis. (%)
ActionFormer anchor, vis. GT	InternVideo2-1B, temporal-only, 10% train split	64.6%	95.5%
ActionFormer anchor, all GT	as above, paper-parity GT	75.9%	94.0%
DiGIT, vis. GT	InternVideo2-1B; 5 epochs (undertrained)	59.8%	88.7%
DiGIT, all GT	as above	61.7%	87.3%
ActionFormer + InternVideo	published (dataset paper)	36.08%	—
ActionFormer + VideoMAEv2	published (dataset paper)	20.24%	—
TE-TAD	DeformTrace table; source, backbone unstated	71.4%	—
TadTR	DeformTrace table; source, backbone unstated	60.5%	—
BA-TFD audio-visual	published AV reference (dataset paper)	37.37%	—
Released systems scored here — 0–1 as the harness reports them; both scored on the first 256 frames			
OmniFD, zero-shot	end-to-end, ForgeryNet-trained	0.228	0.343
OmniFD, retrained on our pair	released recipe, public init; ≈2× our pilot clips, ≈8× our pilot images	0.523	0.767

Table 2: Pilot and ten-fold controls on the official AV-Deepfake1M full-split AP@0.5 (0–1), frozen OpenCLIP, seeds 1–3 (4–5: Appendix A.2). Post-hoc scoring; the $s \to t$, single-task-readout, shuffled-mask and ten-fold control arms (Sections 3.2 and 5.2) were also trained post hoc; the frozen-attention arm keeps its patch projection trainable. The two budgets are separate blocks, never compared by level; —: not run at that budget. Differences use unrounded AP; pilot intervals: Tables 20 and 21; other controls: Appendix A.2.

	pilot budget			ten-fold budget		
Arm	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
naive / temporal-only	0.081	0.081	0.083	0.138	0.152	0.149
cross-task attention	0.146	0.148	0.160	0.171	0.170	0.172
attention, no mask labels	0.128	0.156	0.141	0.171	0.180	0.191
attention, shuffled mask labels (post hoc)	0.139	0.158	0.127	—	—	—
parameter-matched self-attention	0.107	0.120	0.124	—	—	—
attention, frozen module	0.117	0.131	0.136	0.165	0.192	0.176
$s \to t$ attention only	0.134	0.148	0.133	—	—	—
single-task attention readout	0.139	0.131	0.133	0.169	0.166	0.166
full attention − no mask labels	+0.017	−0.008	+0.019	−0.000	−0.010	−0.019
full attention − single-task readout	+0.006	+0.017	+0.027	+0.002	+0.004	+0.006

ten-fold — positive at every seed of every budget, and at ten-fold on AR@10 alike, while the regularizer has small, mixed-sign differences from naive at the largest budget (Table 1). Absolute values are far below the internal metric's, a level gap a fixed-prediction audit (post hoc) attributes mostly to sampled-grid annotation and coordinate changes, not AP integration (Appendix A.1; Table 3). Every attention-minus-naive difference keeps its sign in 95% of 5% subsamples (Table 56); evaluation resolution does not bound run-to-run spread (Appendix A.2). For context, OmniFD's recipe retrained on our pair reaches 0.523 on the same harness against our 0.171, a level gap described,

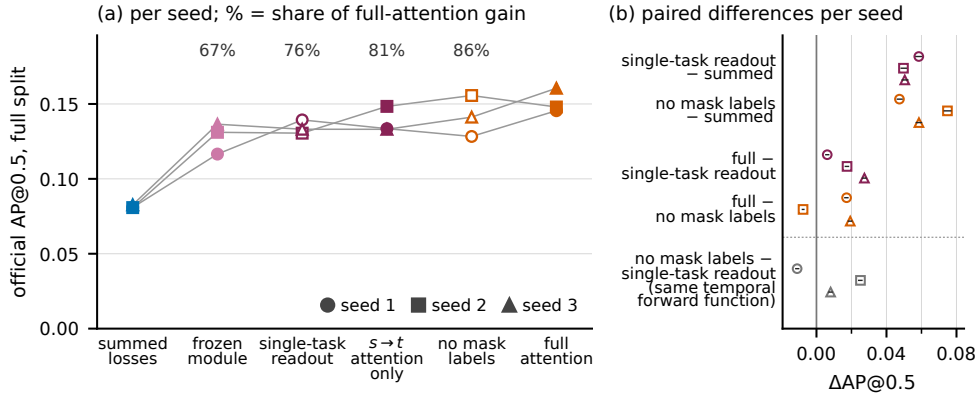


Figure 2: Label and direction controls on the official metric (pilot budget, full split, seeds 1–3; post hoc). (a) AP@0.5 per arm and seed (hollow: no mask labels; arms as in Table 2; temporal-only equals summed losses to all digits); percentages: ratios of mean paired gains (Section 3.2). (b) Paired differences per seed; each drawn full-split 95% video-bootstrap interval is at most 0.0046 wide (all intervals: Tables 20 and 21); the "same temporal forward function" row pairs arms differing in data order and optimizer trajectory, which such intervals do not bound.

not attributed (Appendix A.7). Removing that recipe's mask loss (post hoc) changes its AP@0.5 by -0.0071 ± 0.0077 over three seeds, lower at every seed (a sign pattern, no interval); the contrast is the recipe's whole effect of mask supervision, not the architecture-preserving split of Section 5.2 (Appendix A.7).

Robustness. The ordering holds at every seed on a second frozen backbone (DINOv2, Oquab et al., 2024), where attention without mask labels retains 98% of the pilot official gain and 73% at ten-fold, above naive at every seed and below full attention at ten-fold (post hoc; Appendix A.3). On a second dataset pair (FaceForensics++ masks, TVIL segments) attention leads naive at every seed, attention without mask labels retains 89% of that gain (post hoc, internal metric, positive at every seed; Appendix A.3), and the regularizer is unresolved (beyond ±0.02 at every seed, both signs); orderings only. Main pair, official (post hoc): attention exceeds naive at 3/3 frozen seeds, 2/3 when fine-tuned at one hundredth or one tenth of the heads' rate (internal AP: 2/3 and 1/3). The one-tenth count includes resumed seed 1; its uninterrupted rerun gives 1/3 (Appendix A.3). Neither recipe consistently retains the advantage; the regularizer is unresolved. Full minus label-free attention stays within 0.01 at every seed under both recipes; contribution 1 is scoped to frozen backbones.

5.2 Controls: What the Gain Is Made Of

Single-task tie. The temporal-only control reproduces naive's validation AP to all 16 digits at every seed: naive is the single-task floor.

How much requires mask labels? On the official evaluation code at pilot budget (post hoc), attention without mask supervision retains 86% of the mean attention-minus-naive gain (91% over seeds 1–5, post hoc; Appendix A.2), and 138% at ten-fold budget (post hoc), where full attention does not exceed it at any seed on the full split (the margin is 0.00006 at seed 1; Table 13). Freezing the attention module at initialization while training its patch projection retains 67% (Figure 2; Table 2). Full attention minus the label-free arm, $+0.010 \pm 0.015$ at pilot budget, changes sign across seeds (two of three positive; intervals: Table 20). Permuting the mask targets within each batch (spatial gradient kept, image–mask correspondence destroyed) leaves the label-free level within the seed spread (shuffled minus label-free $+0.011$ / $+0.002$ / -0.014; 86% retained over naive, positive at every seed; post hoc, Table 2), so the residual effect does not track label content. Removing temporal-to-spatial attention retains 81%; the single-task attention readout, which also removes mask supervision and so has no trainable path from the spatial task, retains 76%, positive at every seed, with every interval of that gain over naive excluding zero (Table 21). Full attention exceeds it at every seed;

that gap is not attributed (Section 6.1). At ten-fold (post hoc; Table 2) the single-task readout retains 85% and the frozen module 126%, each above naive at every seed: the budgets order the two arms oppositely. Mask dose moves the gain little: with a tenth or a hundredth of the mask pool re-iterated to match exposure, attention keeps $+0.244$ / $+0.250$ of the $+0.271$ internal-metric gain over naive (full minus reduced positive at 3/3 and 2/3 seeds; post hoc); unmatched, the reduced arms gain more at every seed (full minus reduced -0.0476 to -0.0843), an exposure effect of $+0.06$ to $+0.10$ (Appendix A.3). Other controls: Appendix A.2.

5.3 Generalization audit: joint readout and zero-shot transfer

The in-domain gain does not carry to the other readouts. On the joint where-and-when set (Section 4) the in-domain ranking does not carry over (Table 31, Appendix A.4): the regularizer tops tube-mAP@0.2 at two of three pilot seeds, the $10\times$ ordering flips between seeds, and in-domain AP selects different checkpoints from this weak-reference criterion (inversion statistic, Section 4: 0.294 at pilot and 0.634 at $10\times$, each mixing near-0 and near-1 seeds). *TVIL (video inpainting).* At pilot scale attention trails naive at every seed on AP@0.5 and on the designated secondary metric AR@10 (mean -0.0232 AP@0.5; inversion statistic 0.029; Table 32, Appendix A.4); at ten-fold (mean -0.0153; statistic 0.260) it leads only at seed 1, with a paired interval including zero. Absolute AP is coarse (short inpainted segments, sparse temporal sampling); ranks and paired deltas are the readout. The deficit needs no mask labels: the label-free attention arm trails the temporal-only control at every seed, and full attention is within 0.010 of the label-free arm (Table 22; Appendix A.10).

DF40 (detection AUC). The regularizer family's pooled face-local AUC is near 0.001 at ten-fold budget and inverted at every pilot seed, against naive and attention near or below chance (Appendix A.4), consistent with a source-associated response: these checkpoints score real and fake FaceForensics++ input near 1 (median 0.92–1.00 on both classes, evidence AUC 0.50–0.53) while the DF40 reals are FaceForensics++ originals (Section 4), so source and authenticity are confounded in the pooled AUC (Section 6.2; Figure 5). The inversion statistic is 0.001 at pilot and 0.307 at ten-fold, where the naive-versus-attention ordering is seed-unstable; no ranking conclusion is drawn. *OpenForensics (in-the-wild swaps, spatial).* The baseline keeps a usable mask while the regularizer's collapses: a paired IoU@0.5 deficit of -0.120 ± 0.058 on Test-Dev, same sign at every pilot seed, reproduced on Test-Challenge, intervals excluding zero at every seed on both splits. This amplifies the in-domain deficit rather than inverting it (inversion statistic 0.561): the regularizer retains 4–8% of its in-domain mask quality, the baseline 23–44% (Appendix A.4). Attention shows no consistent direction against the baseline.

6 Analysis

6.1 What the feature-space controls establish

With a frozen backbone, auxiliary labels can change the target head only through a parameter both losses update; the naive arm has none, the exact tie of Section 5.2 is that condition's signature (Equation 1), and a coupling module is the only way to create such a parameter. Adding to naive one normalization both losses update (post hoc; Appendix A.2) breaks the exact tie at every seed, so labels reach the head through the shared path; removing them moves official AP by -0.00048 / -0.00093 / $+0.00001$ (positive at one seed of three). Cross-task attention adds a residual spatial-to-temporal attention readout over each frame's projected patch tokens to the pooled frame vector the temporal head reads; its temporal-to-spatial direction feeds only the mask decoder and, with mask labels, carries mask-loss gradients into that readout. The label-free control keeps both and so isolates Δ_{labels} of Equation 1 per seed, with A the official temporal AP.

The residual label effect changes sign across pilot seeds, per-seed intervals excluding zero (Table 20). Self-attention matches parameter count but lacks patch access; frozen attention leaves the patch projection trainable. The single-task attention readout (post hoc) retains 76% at pilot and 85% at ten-fold with no trainable path from the spatial task. It shares one temporal forward function with the label-free attention arm (whose temporal-to-spatial attention never feeds the temporal head), yet their AP differs by up to 0.025 on the full split, every interval excluding zero (Appendix A.2): data order and optimizer trajectory alone move AP that much, so seed spread, not a per-seed interval,

is what arm comparisons must clear. Neither that spread nor the $+0.006$ to $+0.027$ gap from full attention is attributed. One- and two-head readouts retain 56% and 88% (post hoc, above naive at every seed; Appendix A.2). Linear CKA of the features the temporal head reads (Appendix A.2, post hoc): same-seed attention against its label-free twin is 0.822 / 0.968 / 0.787 and against the single-task readout 0.806 / 0.832 / 0.789, within or above the cross-seed range 0.753–0.788 at every seed, so neither mask supervision nor the spatial task moves them beyond seed spread.

6.2 What the output-space probes establish

The tested regularizer has a small, inconsistent temporal effect (Section 5.1); its spatial signature is value agreement rather than mask extent: at seed 1 the fraction of pixels predicted within 0.01 of ground truth is 0.92 versus 0.15 for naive, while mask IoU@0.5 falls at every seed (Appendix A.5).

A post-hoc probe of checkpoints finds temporal scores near 1 on real and fake FaceForensics++ images, whereas naive and attention score 0.001–0.003: a source-associated response bearing on the source-confounded DF40 result (Section 5.3), not its training-time cause: teacher targets can disagree with the student, and no mask-side ablation cuts the teaching path (Appendix A.6).

A ground-truth-mask pseudo-label arm differs little from naive ($+0.0035$ internal AP, 3 seeds, post hoc) but also changes loss form and weight, and stitched supervision (Appendix A.5) changes the input distribution; neither being a clean teacher-quality intervention, these results constrain the tested recipes without showing teacher quality irrelevant or output-space transfer impossible.

6.3 What a joint test measures, and what predicts out-of-domain rank

Within the evaluated checkpoint collection, validation AP anti-correlates with the joint metric ($\tau = -0.29$) and inpainting transfer ($\tau = -0.50$; Figure 10(c), Appendix A.11): descriptive associations among related checkpoints, the joint one resting on weak mouth masks. No tested diagnostic gives a reliable selection rule: its top checkpoint is never the target's in the 28 diagnostic–target cells and the strongest associations do not persist on the held-out partition (Appendix A.11). Manipulated lip motion is over-smoothed, and smoother fakes tend to be localized better (Appendix A.5).

7 Limitations

The main attention result uses frozen CLIP and DINOv2 ViT-Bs, one head design and FaceForensics++ masks in both dataset pairs. Post hoc, the label-free arm is above naive at every DINOv2 seed and two of three pilot-budget fine-tuned seeds per recipe; the frozen-module and $s{\to}t$-only controls are single-encoder, the latter post hoc without last-epoch or zero-shot readouts. Three seeds per untuned recipe (resumed-run caveat, Section 5) establish neither convergence nor loss of advantage; small differences do not show zero label benefit. Adaptive task weighting (Kendall et al., 2018; Chen et al., 2018; Liu et al., 2019) was not tried (balance probed by weight sweep, soft coupling and term ablations). No budget above ten-fold was run; seeds 4–5 are post hoc label ablations (Appendix A.2); AV-Deepfake1M++ validation has one diffusion generator, 199 Diff2Lip clips from an unseen source: zero-shot AP is below 0.022 there, as on its unseen-source TalkLip slice (Appendix A.7).

DF40 scores cannot separate generator generalization from source (Section 5.3); TVIL scenes are largely non-facial (Appendix A.3), so its deficit is as much domain as manipulation-type shift; output-space results cover the tested losses, weights and teachers, not learned cross-task mappings.

8 Conclusion

Under frozen features, most of the attention gain survives with patch access but no masks or spatial task at both budgets (a parameter-matched module with masks but no patch access reaches half): the readout carries it, not primarily the masks. Budget and fine-tuning dependence and weak zero-shot inpainting transfer limit its reach. Crediting a gain to labels requires an architecture-preserving label ablation, a same-readout single-task arm, seed-spread uncertainty and headline-budget attribution.

REPRODUCIBILITY STATEMENT

We release, as supplementary material, an anonymized code snapshot containing: the frozen joint where-and-when evaluation-set definition and its tube-mAP scorer; the cross-generator detection probe with its frozen generator list and frozen score mapping; the zero-shot temporal-transfer recipe; the converters that place model proposals on the dataset's official scoring harness; every training configuration used in the paper, together with the guard tests that pin each ablation arm to its parent configuration; and the pre-registration notes whose dates the appendix timeline table reports. The main joint-model experiments train over frozen backbone features (OpenCLIP ViT-B/16 (Radford et al., 2021; Cherti et al., 2023), LAION-2B pre-trained and ImageNet-12k/1k fine-tuned, or DINOv2 ViT-B/14) with the seeds stated in the tables. Explicitly marked fine-tuning controls and context systems use their own recipes; temporal proposals use one fixed decode (product of start, end, and boundary-map scores; top 16 of all start–end pairs, with no non-maximum suppression) throughout, except the exploratory seed-42 runs of Table 5 (Appendix A.1), which use the boundary-map score alone. The datasets (FaceForensics++, AV-Deepfake1M, AV-Deepfake1M++, DF40, TVIL, OpenForensics) are used for non-commercial academic research only, under their published licences and end-user agreements. Neither dataset media, dataset annotations nor clip identifiers are redistributed, apart from the FaceForensics++ maintainers' public split lists and seeded subsets of them: the snapshot ships code, configurations and the frozen evaluation-set definition, the last of which carries only SHA-256 digests of the identifiers of the 300 manipulated and 100 real AV-Deepfake1M validation clips it selects (286 manipulated clips remain after the landmark-failure exclusion applied when the pseudo-masks are generated; those are the clips scored), and the seed and rule that drew them. Selection is deterministic, so the complete set — its ground-truth segments included — is regenerated by the released script from a licensed copy of the dataset and checked against the hash the snapshot records. FaceForensics++, AV-Deepfake1M and AV-Deepfake1M++ are access-gated: each requires a signed end-user licence agreement and must be obtained from its own maintainers, who alone distribute it. The AV-Deepfake1M and AV-Deepfake1M++ agreement further requires that publications using the data cite Cai et al. (2024), Cai et al. (2023) and Cai et al. (2022), and that the terms of VoxCeleb2 (Chung et al., 2018), from which those videos are derived, are observed as well. Each dataset is cited at the point of use so that the terms attaching to it can be followed.

ETHICS STATEMENT

This work builds forensic tooling for detecting and localizing manipulated video, a dual-use area: a model that localizes manipulations also maps what a generator must improve. We collected no new personal data and identified no individuals; all face imagery comes from existing public research corpora used under their licences and end-user agreements. We report the negative and inversion results in full — including where a detector's scores separate the source a model was trained on from unseen sources rather than manipulation from authenticity — precisely so that such systems are not over-trusted in deployment.

AI USE STATEMENT

An autonomous research assistant built on Claude (Anthropic), operated by the authors, was used throughout this project.

In this work, we used generative AI tools for the following tasks with required disclosure: refining hypotheses and consulting supplementary hypotheses and candidate research directions (which the authors then accepted, narrowed or rejected at each pre-registered gate); helping develop the conceptual framework through which this paper separates feature-space from output-space coupling; helping develop experiments and giving feedback on them, including drafting the pre-registration documents; implementing the methods, the experiment code and its tests; cleaning and reformatting datasets, including dataset-reader validation; supporting qualitative analysis of experiment logs; and interpreting results, including drafting result interpretations and sections of this manuscript.

We have not used generative AI tools for the following task with required disclosure: generating synthetic data sets. No reported experimental result is model-generated; every number comes from a training or scoring run on the public corpora named in the reproducibility statement. AI-assisted

revisions included recommendations on the title, keywords and claim scope; the authors retain responsibility for final submission decisions. The remaining required-disclosure tasks are not applicable to this work: it states and proves no mathematical claims, so formulating mathematical claims, providing critical ingredients for proving them, and assisting in the writing of proofs do not arise; and the manuscript was written in English throughout, so no translation was involved.

Additionally, we used generative AI tools for the following tasks with recommended disclosure: identifying and searching for relevant literature and summarizing and analysing it; brainstorming; suggesting experimental parameters and probe designs; creating and editing software code; creating artifacts released with this paper; creating and modifying the scientific figures in this paper from their frozen specifications; drafting parts of this paper and suggesting its structure; formatting references; and editing the manuscript for readability. Two recommended-disclosure tasks do not arise here: this work runs no surveys or interviews and transcribes no recorded research material.

The authors are responsible for reviewing and verifying all AI-assisted text, interpretations and code before submission. Generated code is checked with the released tests, including configuration guards for new controls. The source ledger links main-text numerical results to experiment artifacts or dated notes. New post-hoc controls enter as evidence only after their outputs have been scored and identity-checked. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. Next-frame feature prediction for multimodal deepfake detection and temporal localization. *arXiv preprint arXiv:2511.10212*, 2025.

Peijun Bao, Anwei Luo, Gang Pan, Alex C. Kot, and Xudong Jiang. ActivityForensics: A comprehensive benchmark for localizing manipulated activity in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2604.03819.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? Content Driven Audio-Visual deepfake dataset and multimodal method for temporal forgery localization. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–10, Sydney, Australia, 2022. doi: 10.1109/DICTA56598.2022.10034605.

Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the Matrix: A large scale benchmark for Content Driven Audio-Visual forgery detection and localization. *Computer Vision and Image Understanding*, 236:103818, 2023. doi: 10.1016/j.cviu.2023.103818.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7414–7423, 2024. doi: 10.1145/3664647.3680795.

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. AV-Deepfake1M++: A large-scale audio-visual deepfake benchmark with real-world perturbations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13686–13691, 2025. doi: 10.1145/3746027.3761979. arXiv:2507.20579.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. arXiv:2010.06595.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. arXiv:1711.02257.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for

contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, pp. 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929.

Soumyya Kanti Datta, Shan Jia, and Siwei Lyu. LoCC: Detection and localization of lip-syncing deepfakes via counterfactual frame consistency. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2026. arXiv:2606.22772.

Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. arXiv:2109.04617.

GuangJian Team. Venus-DeFakerOne: Unified fake image detection & localization. *arXiv preprint arXiv:2605.14091*, 2026.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. arXiv:1705.07115.

Ho-Joong Kim, Jung-Ho Hong, Heejo Kong, and Seong-Whan Lee. TE-TAD: Towards full end-to-end temporal action detection via time-aligned coordinate expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18837–18846, 2024.

Ho-Joong Kim, Yearang Lee, Jung-Ho Hong, and Seong-Whan Lee. DiGIT: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Nicholas Klein, Hemlata Tak, James Fullwood, Krishna Regmi, Leonidas Spinoulas, Ganesh Sivaraman, Tianxiang Chen, and Elie Khoury. Pindrop it! Audio and Visual deepfake countermeasures for robust detection and fine-grained localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13700–13706, 2025. doi: 10.1145/3746027.3761981. arXiv:2508.08141.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. arXiv:1905.00414.

Christos Koutlis and Symeon Papadopoulos. AuViRe: Audio-visual speech representation reconstruction for deepfake temporal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. arXiv:2511.18993.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12169–12183, 2022.

Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10097–10107, 2021.

Qingcao Li, Miao He, Liang Yi, Qing Wen, Yitao Zhang, Hongshuo Jin, Peng Cheng, Zhongjie Ba, et al. Divide and conquer: Multimodal video deepfake detection via cross-modal fusion and localization. *arXiv preprint arXiv:2602.00209*, 2026.

Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Fangzhou Lin, Yuping Wang, Yuliang Guo, Zixun Huang, Xinyu Huang, Haichong Zhang, Kazunori Yamada, Zhengzhong Tu, Liu Ren, and Ziming Zhang. NexusFlow: Unifying disparate tasks under partial supervision via invertible flow networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2512.06251.

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. arXiv:1907.09702.

Haotian Liu, Haoyu Chen, Chenhui Pan, You Hu, Guoying Zhao, and Xiaobai Li. OmniFD: A unified model for versatile face forgery detection. *arXiv preprint arXiv:2512.01128*, 2025a.

Haotian Liu, Yang Liu, Guoying Zhao, and Xiaobai Li. Learning unified video and image representation for video face forgery detection. *arXiv preprint arXiv:2608.13064*, 2026.

Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. arXiv:1803.10704.

Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441, 2022. doi: 10.1109/TIP.2022.3195321. arXiv:2106.10271.

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. *arXiv preprint arXiv:2508.21052*, 2025b.

Changtao Miao, Yi Zhang, Weize Gao, Zhiya Tan, Weiwei Feng, Man Luo, Jianshu Li, Ajian Liu, et al. DDL: A large-scale datasets for deepfake detection and localization in diversified real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025.

Jingchen Ni, Cangjin Yu, Dan Jiang, Quan Zhang, Keyu Lv, Shannan Yan, Linyue Pan, Ke Zhang, and Chun Yuan. MG-RWKV: Multi-grained context-aware RWKV for temporal forgery localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Youngmin Oh, Hyung-Il Kim, and Jung Uk Kim. Task prototype-based knowledge retrieval for multi-task learning from partially annotated data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2601.07474.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. arXiv:2304.07193.

Dhaval C. Patel, Kaoutar El Maghraoui, Shuxin Lin, Yusheng Li, Tianjun Feng, Chun-Yi Tsai, Yihan Sun, et al. Beyond static leaderboards: Predictive validity for the evaluation of LLM agents. *arXiv preprint arXiv:2606.19704*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00020.

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.

Sanjay Saha, Rashindrie Perera, Sachith Seneviratne, Tamasha Malepathirana, Sanka Rasnayaka, Deshani Geethika, Terence Sim, and Saman Halgamuge. Undercover deepfakes: Detecting fake segments in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023. arXiv:2305.06564.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 25278–25294, 2022.

Federica Spinola, Philipp Benz, Minhyeong Yu, and Tae-hoon Kim. Knowledge Assembly: Semi-supervised multi-task learning from multiple datasets with disjoint labels. *arXiv preprint arXiv:2306.08839*, 2023.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. arXiv:1905.07553.

Yihong Sun, Seoung Wug Oh, Jiahui Huang, Bharath Hariharan, and Joon-Young Lee. Unified video dense prediction from disjoint data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. arXiv:2607.21592.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2303.16727.

Tianyi Wang, Xi Shao, Harry Cheng, Yinglong Wang, and Mohan Kankanhalli. Word-anchored temporal forgery localization. *arXiv preprint arXiv:2603.06220*, 2026.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, et al. InternVideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. arXiv:2403.15377.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:2005.00944.

Derrick Xin, Behrooz Ghorbani, Ankush Garg, Orhan Firat, and Justin Gilmer. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems*, volume 35, pp. 13597–13609, 2022.

Yichen Xu, Ryumei Nakada, Linjun Zhang, and Lexin Li. Residual feature integration is sufficient to prevent negative transfer. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.11771.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, et al. DF40: Toward next-generation deepfake detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 29387–29434, 2024.

Qilin Yin, Wei Lu, Xiangyang Luo, Rui Yang, and Xiaochun Cao. Context-aware TFL: A universal context-aware contrastive learning framework for temporal forgery localization. *IEEE Transactions on Dependable and Secure Computing*, 23:8146–8157, 2026. doi: 10.1109/TDSC.2026.3680621. arXiv:2506.08493.

Cangjin Yu, Quan Zhang, Dan Jiang, and Ke Zhang. UniSkip-Mamba: A frequency-aware state space model for audio-visual temporal forgery localization. *arXiv preprint arXiv:2607.04498*, 2026.

Amir Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2006.04096.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Jingdong Zhang, Hanrong Ye, Xin Li, Wenping Wang, and Dan Xu. Multi-task label discovery via hierarchical task tokens for partially annotated dense predictions. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 719–728, 2025. doi: 10.1145/3746027.3755727. arXiv:2411.18823.

Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. UMMAFormer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8749–8759, 2023.

Chende Zheng, Ruiqi Suo, Zhoulin Ji, Jingyi Deng, Fangbin Yi, Chenhao Lin, and Chao Shen. Multi-modal deepfake detection and localization with FPN-transformer. *arXiv preprint arXiv:2511.08031*, 2025.

Xiaodong Zhu, Suting Wang, Yuanming Zheng, Junqi Yang, Yangxu Liao, Yuhong Yang, Weiping Tu, and Zhongyuan Wang. DeformTrace: A deformable state space model with relay tokens for temporal forgery localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a. arXiv:2603.04882.

Xiaodong Zhu, Yuanming Zheng, Suting Wang, Junqi Yang, Yuhong Yang, Weiping Tu, and Zhongyuan Wang. GEM-TFL: Bridging weak and full supervision for forgery localization through EM-guided decomposition and temporal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. arXiv:2603.05095.

The appendix expands the evidence for claims stated in the main text; no claim appears here for the first time. A few readouts appear only here and support no main-text claim: the spatial-only twin of the temporal-only control (Appendix A.2), a bound reported for completeness; the speaker-exposure quartiles and the proposal-capped OmniFD scores (Appendix A.7); and the per-method TVIL slice (Appendix A.4). Each block names the main-text section it supports, and each condensed main-text table names its full version here. Figure 1 (Section 3.1) summarizes the study design.

A.1 Full per-seed result tables

Supports Sections 5.1 and 5.3. Tables 4–9 give internal AP@0.5 at the best epoch, per seed, beside the same epoch's FaceForensics++ mask metrics (IoU@0.5, value agreement, L1) for the three arms at each budget and for the exploratory single-seed variants; Tables 10–15 give the official evaluation code's AP@0.5 and AR@10 per seed (AP@0.5 only for the DINOv2 table). Blocks are separated by training budget ($1\times$, $3\times$, $10\times$), backbone and metric family (our internal all-point AP implementation vs the official evaluation code); numbers from different blocks are reported side by side but never ranked against each other, because both the evaluated population and the AP implementation differ across blocks. The internal validation window is capped by the batch budget (8,000 clips at the pilot budget, all 43,054 at $10\times$), and the best epoch is selected on that same window. Post hoc (not pre-registered), we re-read every in-domain internal-AP comparison as it stood on 2026-09-12 under last-epoch selection instead (rows added afterwards — among them the $s\to t$-only, shuffled-mask and seed 4–5 rows, the later label-free rows at ten-fold, on DINOv2, unfrozen and on the second pair, and the spatial-only twin — were not re-read, nor were the official-code and zero-shot readouts): no attention-minus-naive sign changes at any budget or backbone, and no control verdict changes (the exact single-task tie; attention minus the label-free arm inside ± 0.02, a description of spread; the frozen module above $+0.05$; the 1% dose above $+0.15$). What does change: the regularizer's in-domain deltas change sign at some seeds, each such delta inside the seed-noise band at its best epoch, and the second dataset pair's attention-minus-naive gain, outside that band at every seed at its best epoch, changes sign at one seed, which under that pair's pre-registered rule would read as not replicating.

Why the internal and official levels differ (post hoc). A fixed-prediction audit of the ten-fold attention checkpoint uses the same visual-subset videos throughout (Table 3). The original-time official score is 0.255909. Changing only the coordinate and annotation representation to the reader's sampled grid, while retaining official matching and integration, raises it to 0.831724. Internal matching then raises it to 0.882748; switching integration gives 0.882781, reproducing the training readout. Integration alone leaves the original-time score unchanged. The annotation transform drops

7,661 of 38,161 segments and merges 100 more, leaving 30,400. Coordinate quantization, dropping and merging are changed together here, not independently isolated. Thus the large level gap mainly reflects a different localization target in this audit, not AP integration. Prediction round-trip parity does not establish annotation equivalence. Input hashes, aggregate counts and the CPU driver are released; benchmark annotations remain unmodified.

Table 3: Post-hoc evaluation reconciliation, fixed ten-fold attention predictions on 43,054 visual-subset videos. Original: original-time segments; grid: rounded, sampled and merged reader targets with remapped predictions. Official matching is label-major with strict IoU; internal matching is proposal-major with inclusive IoU. These diagnostic transforms do not replace official benchmark scoring.

Coordinates and GT	Matching	Official integration	Envelope integration
Original	Official	0.255909	0.255909
Original	Internal	0.265599	0.265599
Sampled grid	Official	0.831724	0.831757
Sampled grid	Internal	0.882748	0.882781

Behind the pilot budget of Section 5.1 and the two further arms of Section 3.2: the single-seed pilot weight sweep of the output-space regularizer (seed 42, scored with the boundary-map score alone; Table 5) is monotone past its selected weight — against the naive baseline, $w = 0.5$ gives $+0.0016$ and $w = 2.0$ gives -0.0451, the latter strictly below the baseline on both heads; in the same single-seed block the cross-dataset pseudo-labeling rival is -0.006 against the baseline; the soft-coupling variant of the regularizer, run once as pre-registered, reaches 0.4668 against the attention bar of 0.6954 (Table 6). On the official harness at $10\times$ the regularizer is paired against the baseline at -0.0008 (full split) / -0.0010 (visual-manipulation subset), a mean a fifth the size of its own seed spread, with the same per-seed sign pattern as under the internal metric (Table 12). Mask IoU within temporal true positives at 0.5, the second readout of the joint evaluation set behind Table 31, per seed: pilot budget naive 0.1683 / 0.1506 / 0.1698, attention 0.1458 / 0.1381 / 0.1748, regularizer 0.1946 / 0.1545 / 0.1063; $10\times$ naive 0.1181 / 0.1460 / 0.1232, attention 0.0745 / 0.1661 / 0.1583, regularizer 0.1230 / 0.1194 / 0.1075.

Table 4: Internal AP@0.5 at the best epoch, pilot budget (1,000 batches $\times$ 10 epochs), frozen OpenCLIP ViT-B/16, fused start $\cdot$ end $\cdot$ boundary-map decode, 8,000-clip validation window, seeds 1–3. Spatial columns are FaceForensics++ mask metrics at the same epoch; value agreement is the fraction of pixels whose soft value lies within 0.01 of the ground truth. Differences are paired by seed. Mean rows: level mean $\pm$ sample sd and mean paired difference. †Best epoch is the last.

Seed	Best epoch	AP@0.5	Δ vs naive	IoU@0.5	Value agreement	L1
Naive sum						
1	9†	0.4403	—	0.4072	0.1530	0.0306
2	9†	0.4049	—	0.4403	0.0068	0.0423
3	9†	0.4270	—	0.3976	0.1718	0.0307
Mean		0.4241 $\pm$ 0.0179	—			
Cross-task attention						
1	9†	0.6919	+0.2516	0.4587	0.1115	0.0322
2	8	0.6984	+0.2935	0.4475	0.0054	0.0504
3	9†	0.6959	+0.2689	0.4026	0.0779	0.0315
Mean		0.6954 $\pm$ 0.0033	+0.2713			
Output-space consistency (w=0.1)						
1	9†	0.4674	+0.0271	0.2586	0.9239	0.0186
2	7	0.4072	+0.0023	0.2116	0.2906	0.0276
3	7	0.4348	+0.0078	0.2213	0.5721	0.0253
Mean		0.4365 $\pm$ 0.0301	+0.0124			

Table 5: Exploratory single-seed pilot runs (seed 42), internal AP@0.5 at the best epoch, pilot budget. These runs use a boundary-map-only decode, unlike every other internal table, so they are compared only with each other. Every run peaks at its last epoch (9). Differences are against the seed-42 naive run; attention plus consistency lies -0.0207 below attention alone. Spatial columns as in Table 4; —: not recorded.

Run (seed 42)	Best epoch	AP@0.5	Δ vs naive	IoU@0.5	Value agreement	L1
naive sum	9	0.3803	—	0.446	0.016	0.038
cross-dataset pseudo-labeling	9	0.3746	-0.006	0.424	—	0.019
cross-task attention	9	0.7026	$+0.3223$	0.416	0.0197	0.037
Output-space consistency, weight sweep						
w=0.1	9	0.3983	$+0.0180$	0.388	0.916	0.021
w=0.5	9	0.3819	$+0.0016$	0.293	0.925	0.019
w=2.0	9	0.3352	-0.0451	0.132	0.932	0.019
attention + consistency (w=0.1)	9	0.6819	$+0.3016$	0.057	0.896	0.022

Table 6: Exploratory single-seed pilot runs (seed 1), internal AP@0.5 at the best epoch, fused decode. Differences are against the naive run trained on the same data mix and seed; the soft-coupled variant lies -0.0006 below hard-coupled consistency at seed 1 (Table 4). The stitched mix changes the training data, so its rows are not ranked against the standard mix. Spatial columns as in Table 4. †Best epoch is the last.

Run (seed 1)	Best epoch	AP@0.5	Δ vs naive	IoU@0.5	Value agreement	L1
Standard two-source mix, 10 epochs						
soft-coupled consistency (w=0.1)	9†	0.4668	$+0.0265$	0.2315	0.9252	0.0188
Mix with 25% stitched same-clip FaceForensics++ data, 10 epochs						
naive sum	9†	0.2938	—	0.3121	0.9064	0.0202
output-space consistency (w=0.1)	9†	0.3260	$+0.0321$	0.1404	0.9265	0.0184
cross-task attention	8	0.5736	$+0.2797$	0.3650	0.5621	0.0265
Same stitched mix, continued to 20 epochs (fresh optimizer at epoch 10)						
naive sum	19†	0.4380	—	0.3417	0.9362	0.0158
output-space consistency (w=0.1)	19†	0.4583	$+0.0203$	0.1772	0.9407	0.0160

Table 7: Cross-task attention plus output-space consistency (w=0.05), internal AP@0.5 at the best epoch, pilot budget, fused decode. The seed-42 run that prompted the confirmation is shown apart, not averaged in. Against same-seed attention alone the differences are mixed in sign and each within ± 0.02, so no temporal direction is read; the joint-readout tube-mAP@0.2 beats consistency alone at every seed, by strongly seed-dependent margins, so the combination is reported as seed-sensitive.

Seed	Best epoch	AP@0.5	Δ vs attention	IoU@0.5	Value agreement	L1
Exploratory origin run, single seed						
42	9†	0.7135	—	0.1405	0.9118	0.0216
Confirmation, seeds 1–3 (attention alone: Table 4)						
1	8	0.6816	-0.0103	0.0248	0.8361	0.0247
2	8	0.6816	-0.0168	0.2953	0.7636	0.0251
3	8	0.6974	$+0.0015$	0.1358	0.8454	0.0234
Mean ± sample sd		0.6869 ± 0.0091	-0.0085 ± 0.0093			

†Best epoch is the last epoch.

Table 8: Internal AP@0.5 at the best epoch, three-fold budget (3,000 batches × 10 epochs), frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3; only naive sum and cross-task attention were trained at this budget. The internal validation window is capped by the same batch count, so levels are not comparable across budgets; only the paired difference is. Columns as in Table 4; mean ± sample sd. †Best epoch is the last.

Seed	Best epoch	AP@0.5	Δ vs naive	IoU@0.5	Value agreement	L1
Naive sum						
1	9†	0.5847	—	0.2942	0.9394	0.0154
2	8	0.5779	—	0.3658	0.9308	0.0173
3	5	0.5635	—	0.3470	0.9122	0.0200
Mean		0.5754 ± 0.0108	—			
Cross-task attention						
1	8	0.8164	+0.2318	0.2663	0.9397	0.0157
2	9†	0.8082	+0.2303	0.3400	0.9383	0.0157
3	7	0.7946	+0.2311	0.3443	0.9361	0.0166
Mean		0.8064 ± 0.0110	+0.2311 ± 0.0008			

Table 9: Internal AP@0.5 at the best epoch, ten-fold budget (10,000 batches × 10 epochs), frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3. At this budget the validation cap no longer binds, so AP is computed on the full 43,054-clip filtered validation set rather than the pilot's 8,000-clip window; levels are not comparable across budgets. Columns as in Table 4; mean ± sample sd. †Best epoch is the last.

Seed	Best epoch	AP@0.5	Δ vs naive	IoU@0.5	Value agreement	L1
Naive sum						
1	6	0.6761	—	0.2833	0.9313	0.0170
2	8	0.6762	—	0.3805	0.9305	0.0169
3	9†	0.7055	—	0.3625	0.9353	0.0160
Mean		0.6859 ± 0.0169	—			
Cross-task attention						
1	9†	0.8828	+0.2067	0.2319	0.9400	0.0159
2	8	0.8729	+0.1967	0.2933	0.9386	0.0159
3	9†	0.8776	+0.1721	0.3424	0.9431	0.0157
Mean		0.8778 ± 0.0049	+0.1918 ± 0.0178			
Output-space consistency (w=0.1)						
1	7	0.6957	+0.0196	0.0032	0.9696	0.0195
2	8	0.6895	+0.0133	0.0064	0.9696	0.0192
3	5	0.6882	−0.0172	0.0182	0.9657	0.0190
Mean		0.6911 ± 0.0040	+0.0052 ± 0.0197			

Table 10: Official AV-Deepfake1M evaluation code, pilot budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3: AP@0.5 and AR@10 on the 57,340-record split of Table 1 and its 43,054-record visual-manipulation subset; consistency weight w=0.1. Differences are paired by seed; mean and sample sd over seeds; —: not aggregated. Never ranked against other budgets, the DINOv2 block, or internal AP.

Seed	Naive sum	Cross-task attention	Output-space consistency	Attention − naive	Consistency − naive
AP@0.5, full split					
1	0.080899	0.145505	0.088406	+0.064606	+0.007507
2	0.080805	0.147976	0.085255	+0.067171	+0.004450
3	0.082663	0.160491	0.082507	+0.077828	−0.000156
Mean	0.08146	0.15132	0.08539	+0.06987	+0.00393
Sample sd	0.00105	0.00803	0.00295	0.00701	0.00386
AP@0.5, visual-manipulation subset					
1	0.121979	0.217344	0.133314	+0.095365	+0.011335
2	0.121551	0.221117	0.128636	+0.099566	+0.007085
3	0.124516	0.239611	0.124440	+0.115095	−0.000076
Mean	0.12268	0.22602	0.12880	+0.10334	+0.00611
Sample sd	0.00160	0.01192	0.00444	0.01039	0.00577
AR@10, full split					
1	0.182374	0.201351	0.183098	+0.018977	+0.000724
2	0.183686	0.200298	0.183506	+0.016612	−0.000180
3	0.184659	0.201531	0.178459	+0.016872	−0.006200
Mean	—	—	—	+0.01749	—
Sample sd	—	—	—	0.00130	—
AR@10, visual-manipulation subset					
1	0.259545	0.289468	0.260355	+0.029923	+0.000810
2	0.261898	0.287728	0.261919	+0.025830	+0.000021
3	0.263499	0.289531	0.254896	+0.026032	−0.008603
Mean	—	—	—	+0.02726	—
Sample sd	—	—	—	0.00231	—

Table 11: Official AV-Deepfake1M evaluation code, three-fold budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3; only naive sum and cross-task attention were trained at this budget. Metrics, split and subset as in Table 10; differences paired by seed; —: not aggregated. Never ranked against other budgets or internal AP.

Seed	Naive sum	Cross-task attention	Attention − naive
AP@0.5, full split			
1	0.124355	0.171739	+0.047384
2	0.111067	0.160038	+0.048971
3	0.102324	0.163574	+0.061250
Mean	0.11258	0.16512	+0.05254
Sample sd	0.01109	0.00600	0.00759
AP@0.5, visual-manipulation subset			
1	0.186643	0.256364	+0.069721
2	0.166986	0.238896	+0.071910
3	0.154104	0.244175	+0.090071
Mean	0.16924	0.24648	+0.07723
Sample sd	0.01639	0.00896	0.01117
AR@10, full split			
1	0.193943	0.206118	+0.012175
2	0.192394	0.206531	+0.014137
3	0.190688	0.205184	+0.014496
Mean	—	—	+0.01360
Sample sd	—	—	0.00125
AR@10, visual-manipulation subset			
1	0.276408	0.296981	+0.020573
2	0.274414	0.296502	+0.022088
3	0.271806	0.296009	+0.024203
Mean	—	—	+0.02229
Sample sd	—	—	0.00182

Table 12: Official AV-Deepfake1M evaluation code, ten-fold budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3: the per-seed rows behind the top block of Table 1 (its label-free arm: Table 13). Metrics, split and subset as in Table 10; consistency weight w=0.1; differences paired by seed; —: not aggregated. Never ranked against other budgets, the context rows of Table 1, or internal AP.

Seed	Naive sum	Cross-task attention	Output-space consistency	Attention − naive	Consistency − naive
AP@0.5, full split					
1	0.137526	0.171020	0.139309	+0.033494	+0.001783
2	0.151900	0.170083	0.152982	+0.018183	+0.001082
3	0.149005	0.172449	0.143772	+0.023444	−0.005233
Mean	0.14614	0.17118	0.14535	+0.02504	−0.00079
Sample sd	0.00760	0.00119	0.00697	0.00778	0.00386
AP@0.5, visual-manipulation subset					
1	0.205926	0.255909	0.208727	+0.049983	+0.002801
2	0.226981	0.254150	0.228793	+0.027169	+0.001812
3	0.222877	0.257875	0.215343	+0.034998	−0.007534
Mean	0.21859	0.25598	0.21762	+0.03738	−0.00097
Sample sd	0.01116	0.00186	0.01023	0.01159	0.00570
AR@10, full split					
1	0.197870	0.209504	0.199386	+0.011634	+0.001516
2	0.200054	0.208934	0.200593	+0.008880	+0.000539
3	0.199890	0.209142	0.198987	+0.009252	−0.000903
Mean	—	—	—	+0.00992	—
Sample sd	—	—	—	0.00149	—
AR@10, visual-manipulation subset					
1	0.284086	0.302830	0.285598	+0.018744	+0.001512
2	0.286342	0.302804	0.286712	+0.016462	+0.000370
3	0.286169	0.302717	0.284282	+0.016548	−0.001887
Mean	—	—	—	+0.01725	—
Sample sd	—	—	—	0.00129	—

Table 13: Official AV-Deepfake1M evaluation code, ten-fold budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3: the per-seed rows behind the *attention, no mask labels* row of Table 1 (post hoc; Appendix A.8, item (viii)). Split and subset as in Table 10; the naive and attention references are the rows of Table 12; differences paired by seed; mean and sample sd over seeds; no interval (points only). Never ranked against other budgets, the context rows of Table 1, or internal AP.

Seed	Attention, no mask labels	No mask labels − naive	Attention − no mask labels
AP@0.5, full split			
1	0.171079	+0.033553	−0.000059
2	0.179768	+0.027868	−0.009685
3	0.191418	+0.042413	−0.018969
Mean	0.18075	+0.03461	−0.00957
Sample sd	0.01021	0.00733	0.00946
AP@0.5, visual-manipulation subset			
1	0.255758	+0.049832	+0.000151
2	0.268208	+0.041227	−0.014058
3	0.285657	+0.062780	−0.027782
Mean	0.26987	+0.05128	−0.01390
Sample sd	0.01502	0.01085	0.01397
AR@10, full split			
1	0.208932	+0.011062	+0.000572
2	0.208960	+0.008906	−0.000026
3	0.208101	+0.008211	+0.001041
Mean	0.20866	+0.00939	+0.00053
Sample sd	0.00049	0.00149	0.00053
AR@10, visual-manipulation subset			
1	0.302631	+0.018545	+0.000199
2	0.303352	+0.017010	−0.000548
3	0.301027	+0.014858	+0.001690
Mean	0.30234	+0.01680	+0.00045
Sample sd	0.00119	0.00185	0.00114

Table 14: Official AV-Deepfake1M evaluation code, ten-fold budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3: the per-seed rows behind the *attention, frozen module* and *single-task attention readout* cells of the ten-fold block of Table 2 (post hoc; Appendix A.8, item (xiii)). Split and subset as in Table 10; the naive and attention references are the rows of Table 12; differences paired by seed; mean and sample sd over seeds; no interval (points only). Never ranked against other budgets, the context rows of Table 1, or internal AP.

Seed	Attention, frozen module	Single-task readout	Frozen − naive	Single-task − naive	Attention − frozen	Attention − single-task
AP@0.5, full split						
1	0.164680	0.169375	+0.027154	+0.031849	+0.006340	+0.001645
2	0.191947	0.166448	+0.040047	+0.014548	−0.021864	+0.003635
3	0.176250	0.166265	+0.027245	+0.017260	−0.003801	+0.006184
Mean	0.17763	0.16736	+0.03148	+0.02122	−0.00644	+0.00382
Sample sd	0.01369	0.00175	0.00742	0.00931	0.01429	0.00228
AP@0.5, visual-manipulation subset						
1	0.246083	0.253432	+0.040157	+0.047506	+0.009826	+0.002477
2	0.286157	0.248690	+0.059176	+0.021709	−0.032007	+0.005460
3	0.262987	0.248761	+0.040110	+0.025884	−0.005112	+0.009114
Mean	0.26508	0.25029	+0.04648	+0.03170	−0.00910	+0.00568
Sample sd	0.02012	0.00272	0.01099	0.01385	0.02120	0.00332
AR@10, full split						
1	0.208059	0.207931	+0.010189	+0.010061	+0.001445	+0.001573
2	0.208461	0.208159	+0.008407	+0.008105	+0.000473	+0.000775
3	0.206836	0.207828	+0.006946	+0.007938	+0.002306	+0.001314
Mean	—	—	+0.00851	+0.00870	+0.00141	+0.00122
Sample sd	—	—	0.00162	0.00118	0.00092	0.00041
AR@10, visual-manipulation subset						
1	0.300904	0.303158	+0.016818	+0.019072	+0.001926	−0.000328
2	0.300891	0.303367	+0.014549	+0.017025	+0.001913	−0.000563
3	0.298928	0.299282	+0.012759	+0.013113	+0.003789	+0.003435
Mean	—	—	+0.01471	+0.01640	+0.00254	+0.00085
Sample sd	—	—	0.00203	0.00303	0.00108	0.00224

Table 15: Official AV-Deepfake1M evaluation code, frozen DINOv2 backbone, pilot budget (upper blocks) and, post hoc (Appendix A.8, item (xxi)), ten-fold budget for the naive, attention and label-free attention arms (lower blocks; the consistency arm was not run there), fused decode, seeds 1–3; AP@0.5 only (AR@10 not tabulated for this backbone). Split and subset as in Table 10; consistency weight w=0.1; differences paired by seed; mean and sample sd over seeds. The label-free attention column and its two paired rows are post hoc (Appendix A.8, item (x)), without intervals. A separate backbone block: only its ordering is compared with the OpenCLIP blocks, never its levels.

Seed	Naive sum	Cross-task attention	Output-space consistency	Attention − naive	Consistency − naive	Attention, no mask labels
AP@0.5, full split						
1	0.084962	0.152640	0.080307	+0.067678	−0.004655	0.155908
2	0.079858	0.148761	0.076974	+0.068903	−0.002884	0.148755
3	0.092807	0.155891	0.088243	+0.063084	−0.004564	0.149425
Mean	0.08588	0.15243	0.08184	+0.06656	−0.00403	0.15136
Sample sd	0.00652	0.00357	0.00579	0.00307	0.00100	0.00395
No mask labels − naive	+0.070946 / +0.068897 / +0.056618			+0.06549 ± 0.00775 (3/3)		
Attention − no mask labels	−0.003268 / +0.000006 / +0.006466			+0.00107 ± 0.00495 (2/3)		
AP@0.5, visual-manipulation subset						
1	0.127437	0.227889	0.120489	+0.100452	−0.006948	0.233034
2	0.119752	0.222195	0.115398	+0.102443	−0.004354	0.221902
3	0.139496	0.232504	0.132631	+0.093008	−0.006865	0.222668
Mean	0.12890	0.22753	0.12284	+0.09863	−0.00606	0.22587
Sample sd	0.00995	0.00516	0.00885	0.00497	0.00147	0.00622
No mask labels − naive	+0.105597 / +0.102150 / +0.083172			+0.09697 ± 0.01208 (3/3)		
Attention − no mask labels	−0.005145 / +0.000293 / +0.009836			+0.00166 ± 0.00758 (2/3)		
Ten-fold budget (post hoc, item (xxi)): AP@0.5, full split						
1	0.155454	0.217769	—	+0.062315	—	0.211217
2	0.148347	0.191500	—	+0.043153	—	0.172383
3	0.158388	0.199863	—	+0.041475	—	0.186116
Mean	0.15406	0.20304	—	+0.04898	—	0.18991
Sample sd	0.00516	0.01342	—	0.01158	—	0.01969
No mask labels − naive	+0.055763 / +0.024036 / +0.027728			+0.03584 ± 0.01735 (3/3)		
Attention − no mask labels	+0.006552 / +0.019117 / +0.013747			+0.01314 ± 0.00630 (3/3)		
Ten-fold budget (post hoc, item (xxi)): AP@0.5, visual-manipulation subset						
1	0.232195	0.324998	—	+0.092803	—	0.315141
2	0.221703	0.285664	—	+0.063961	—	0.257332
3	0.236249	0.298289	—	+0.062040	—	0.277603
Mean	0.23005	0.30298	—	+0.07293	—	0.28336
Sample sd	0.00751	0.02008	—	0.01723	—	0.02933
No mask labels − naive	+0.082946 / +0.035629 / +0.041354			+0.05331 ± 0.02582 (3/3)		
Attention − no mask labels	+0.009857 / +0.028332 / +0.020686			+0.01963 ± 0.00928 (3/3)		

A.2 CONTROL ROWS

Supports Sections 5.2 and 6.1. Three controls probe what the joint arms add. The temporal-only control (no spatial task) fixes the single-task baseline the joint arms must beat for the co-supervision claim. The attention-without-mask-labels control separates the architectural contribution of the cross-task module from its use of mask supervision; post-hoc $s{\to}t$-only controls also remove its temporal-to-spatial direction, with and without mask supervision (Table 21). The real-mass ablation removes one consistency term at a time to test whether a single term drives the transfer behaviour (it does not; Appendix A.6). Table 19 gives every control per seed in domain and Table 22 its zero-shot transfer companions. The temporal-only control reproduces the naive arm's validation AP to sixteen digits at every seed, consistent with the absence of a trainable path between tasks in the naive implementation. Its spatial mirror image, naive with the three temporal loss weights set to zero (post hoc; Appendix A.8, item (xii)), ends training with FaceForensics++ mask metrics that differ from naive's by $+0.0048$ / -0.0005 / $+0.0038$ in IoU@0.5, -0.0450 / $+0.0005$ / -0.0044 in value agreement and at most $+0.0002$ in L1 at seeds 1–3, tracking naive at every epoch in IoU and L1; this is a bound, not the exact tie of the temporal path, because the mask decoder's bilinear upsampling has a nondeterministic GPU backward and no deterministic-algorithm flag is set, so two runs of the naive configuration itself need not reproduce each other's mask metrics; the differences are not attributed.

With the exact single-task tie, the attention gain over naive decomposes per seed as pathway + labels: $+0.2516 = (+0.2567) + (-0.0052)$, $+0.2935 = (+0.3085) + (-0.0150)$, $+0.2689 = (+0.2599) + (+0.0090)$ (each term rounded separately from unrounded AP). The frozen-module control (attention module at its random initialization, never trained; its patch projection and the pooled-feature projection the temporal head reads stay trainable and still receive mask gradients through the frozen module) has a paired gain over naive of $+0.163$ / $+0.170$ / $+0.187$ with paired-bootstrap 95% intervals ($B = 1000$, percentile, over 8,000 validation clips) of $[+0.1547, +0.1722]$ / $[+0.1617, +0.1788]$ / $[+0.1775, +0.1956]$ — entirely above $+0.15$ at every seed — while the full attention arm exceeds the frozen module by $+0.0887$ / $+0.1237$ / $+0.0821$. The per-seed decomposition of the parameter-matched capacity control (the same attention block attending only to own-task features, matching the attention arm's parameter count) retains ≈32–36% of the internal-metric gain and leaves a residual of ≈64–68%. This is descriptive: the control also removes patch-token access, so these fractions do not isolate capacity from cross-task interaction.

The controls on the official metric (post hoc, not pre-registered). Every control above is read on our internal validation metric. Scored with the official evaluation code at the pilot budget (Table 16; same driver, split and decode as Table 1), the same picture holds: the temporal-only control reproduces the naive arm to all six printed digits at every seed; attention minus the label-free arm is $+0.0096 \pm 0.0150$, inside the ± 0.02 band at every seed (a description of spread, not equivalence); the label-free arm carries $+0.0603 \pm 0.0138$ of attention's $+0.0699 \pm 0.0070$ gain over naive (86%, sign 3/3); the frozen random-initialized module carries $+0.0466 \pm 0.0096$ (67%, sign 3/3); and the post-hoc $s{\to}t$-only arms carry $+0.0569 \pm 0.0093$ with mask supervision (81%, sign 3/3) and $+0.0529 \pm 0.0049$ without (76%, sign 3/3; Table 21). The visual-manipulation subset gives the same sign patterns. The shuffled-mask control (post hoc, item (xviii)), which computes the mask loss against targets permuted across the images of each batch so that the spatial gradient is kept and the image–mask correspondence is destroyed, carries $+0.0600 \pm 0.0166$ (86%, sign 3/3); shuffled minus the label-free arm is $+0.011161$ / $+0.002453$ / -0.014306 (2/3) and attention minus shuffled $+0.006036$ / -0.010119 / $+0.033595$ (2/3); the visual subset gives the same sign pattern and fraction (86%). The parameter-matched self-attention control (scored post hoc, item (xxii); the attention module's parameter count attending only to own-task features, no patch-token access) carries $+0.0358 \pm 0.0083$ (51%, sign 3/3), and attention exceeds it at every seed ($+0.0340 \pm 0.0056$, sign 3/3): a parameter-matched module without patch-token access reaches half of the official gain, against a third of the internal-metric gain above and 67% for the random-initialized module that keeps patch access; because the control removes both the cross-task flow and patch-token access, the share does not isolate capacity from access. The visual subset gives the same signs and share (52%). The attention readout with one or two heads instead of four (post hoc, item (xvi); parameter count unchanged) carries $+0.0389 \pm 0.0098$ (56%, sign 3/3) and $+0.0616 \pm 0.0079$ (88%, sign 3/3); attention minus the one-head readout is $+0.034902$ / $+0.017939$ / $+0.040127$ (3/3) and minus the two-head readout $+0.008422$ / -0.003483 / $+0.019737$ (2/3); the visual subset gives the same signs and fractions (56%, 88%).

Table 16: Official AV-Deepfake1M evaluation code, pilot budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3: AP@0.5 on the full 57,340-record split for the control arms, and paired same-seed differences. Post hoc (not pre-registered; Appendix A.8). Temporal-only equals naive sum to all printed digits, as it does on the internal metric. $s \to t$-only arms, both populations and intervals: Table 21. Own metric family: never ranked beside internal-AP tables. Shuffled masks: mask targets permuted across the images of each batch before the mask loss (post hoc, item (xviii)). Param.-matched: the attention module's parameter count attending only to own-task features, no patch-token access (scored post hoc, item (xxii)). One and two heads: the attention module with one or two heads instead of four, parameter count unchanged (post hoc, item (xvi)).

Seed	Temporal-only (single task)	Naive sum	Attention	Attention, no mask labels	Attention, frozen module	Attention, shuffled masks	Self-attention, param.-matched
1	0.080899	0.080899	0.145505	0.128308	0.116540	0.139469	0.107251
2	0.080805	0.080805	0.147976	0.155642	0.131075	0.158095	0.120224
3	0.082663	0.082663	0.160491	0.141202	0.136483	0.126896	0.124356
Attention readout with one / two heads (post hoc, item (xvi)); seed 1 / 2 / 3)							
One head	0.110603 / 0.130037 / 0.120364						
Two heads	0.137083 / 0.151459 / 0.140754						
Paired same-seed differences (seed 1 / 2 / 3; mean ± sample sd)							
Attention − naive	$+0.064606$ / $+0.067171$ / $+0.077828$	$+0.0699 \pm 0.0070$ (3/3)					
No mask labels − naive	$+0.047409$ / $+0.074837$ / $+0.058539$	$+0.0603 \pm 0.0138$ (3/3)					
Frozen module − naive	$+0.035641$ / $+0.050270$ / $+0.053820$	$+0.0466 \pm 0.0096$ (3/3)					
Attention − no mask labels	$+0.017197$ / -0.007666 / $+0.019289$	$+0.0096 \pm 0.0150$ (2+/1−)					
Shuffled masks − naive	$+0.058570$ / $+0.077290$ / $+0.044233$	$+0.0600 \pm 0.0166$ (3/3)					
Shuffled masks − no mask labels	$+0.011161$ / $+0.002453$ / -0.014306	-0.0002 ± 0.0129 (2+/1−)					
Attention − shuffled masks	$+0.006036$ / -0.010119 / $+0.033595$	$+0.0098 \pm 0.0221$ (2+/1−)					
Param.-matched − naive	$+0.026352$ / $+0.039419$ / $+0.041693$	$+0.0358 \pm 0.0083$ (3/3)					
Attention − param.-matched	$+0.038254$ / $+0.027752$ / $+0.036135$	$+0.0340 \pm 0.0056$ (3/3)					
One head − naive	$+0.029704$ / $+0.049232$ / $+0.037701$	$+0.0389 \pm 0.0098$ (3/3)					
Two heads − naive	$+0.056184$ / $+0.070654$ / $+0.058091$	$+0.0616 \pm 0.0079$ (3/3)					
Attention − one head	$+0.034902$ / $+0.017939$ / $+0.040127$	$+0.0310 \pm 0.0116$ (3/3)					
Attention − two heads	$+0.008422$ / -0.003483 / $+0.019737$	$+0.0082 \pm 0.0116$ (2+/1−)					

Seeds 4–5 (post hoc). Two further matched seeds of the naive, attention and label-free attention arms were launched on 2026-09-13 and read under a reading rule dated 2026-09-15, written before any of their values was read (Appendix A.8, item (ix)); Table 17 lists every seed. Over seeds 1–5 on the full split, attention − naive is $+0.0703 \pm 0.0070$ (5/5), the label-free arm − naive $+0.0639 \pm 0.0137$ (5/5), and attention − label-free $+0.0064 \pm 0.0119$, positive at three of five seeds; the label-free arm carries 91% of the mean attention gain (86% over seeds 1–3). The visual-manipulation subset gives the same sign counts and 91%. No interval is computed for seeds 4–5; the three-seed rows above, Table 2 and Figure 2 are unchanged.

Shared-path condition test (post hoc). Section 6.1 states the condition under which the ablation is exact: auxiliary labels reach the target head only through a parameter both losses update, and the naive arm has none. Adding one affine LayerNorm to naive, applied to the pooled vector the temporal head reads and to every projected patch token the mask decoder reads, creates exactly such a parameter. Trained with and without mask labels at seeds 1–3 under a reading rule dated 2026-09-15 (item (xvii)), the internal-AP tie that is exact for naive breaks at every seed (Table 18); on the official full split the label effect through that path, no labels minus labels, is -0.000476 / -0.000928 / $+0.000011$ (positive at one seed of three; visual subset -0.000640 / -0.001564 / $+0.000011$, one of three), while the normalization itself lowers naive by -0.003259 / -0.001877 / -0.003707 (zero of three positive). No fraction is defined for these rows (the common fraction is taken against attention − naive), and no magnitude is claimed beyond the printed values.

Mask-label efficiency (Table 26; summarized in Section 5.2): the 10% dose is 36 video pairs and the 1% dose 4 pairs of the FF++ mask pool; the paired temporal gains over naive are $+0.3321 \pm 0.0288$ (10%) and $+0.3348 \pm 0.0227$ (1%), sign 3/3; the full-dose gain minus the reduced-dose gain lies between -0.0476 and -0.0843 across seeds and doses (both reduced doses above the full dose at every seed); the mask head degrades with dose — IoU@0.5 0.29–0.37 at the reduced doses vs 0.40–0.46 at full, L1 up to 0.39 at 1%. Exposure confounds the dose comparison: the smaller mask pools run out early in each epoch, after which every batch is temporal, so the reduced-dose arms train on more temporal batches than the full-dose arm. Re-iterating the reduced pool so that the spatial and temporal batch counts match the full-dose arm's (post hoc, item (xix)) gives 0.6805 / 0.6477 / 0.6758 at 10% and 0.7014 / 0.6687 / 0.6515 at 1%: full-dose minus matched-dose is $+0.0114$ /

Table 17: Official AV-Deepfake1M evaluation code, pilot budget, frozen OpenCLIP ViT-B/16, fused decode: the three label-ablation arms at seeds 1–5 (seeds 4–5 post hoc, no intervals; Appendix A.8), AP@0.5 on the full split and the visual-manipulation subset, paired same-seed differences, and means with sample sd over seeds 1–3 and 1–5.

Seed	Naive sum	Attention	Attention, no mask labels	Attention − naive	No labels − naive	Attention − no labels
AP@0.5, full split						
1	0.080899	0.145505	0.128308	+0.064606	+0.047409	+0.017197
2	0.080805	0.147976	0.155642	+0.067171	+0.074837	−0.007666
3	0.082663	0.160491	0.141202	+0.077828	+0.058539	+0.019289
4	0.075573	0.139521	0.133446	+0.063948	+0.057873	+0.006075
5	0.079423	0.157409	0.160447	+0.077986	+0.081024	−0.003038
Mean, seeds 1–3	0.08146	0.15132	0.14172	+0.06987	+0.06026	+0.00961
Sample sd, seeds 1–3	0.00105	0.00803	0.01367	0.00701	0.01379	0.01500
Mean, seeds 1–5	0.07987	0.15018	0.14381	+0.07031	+0.06394	+0.00637
Sample sd, seeds 1–5	0.00266	0.00864	0.01389	0.00704	0.01369	0.01193
Positive, seeds 1–5				5/5	5/5	3/5
AP@0.5, visual-manipulation subset						
1	0.121979	0.217344	0.191947	+0.095365	+0.069968	+0.025397
2	0.121551	0.221117	0.232666	+0.099566	+0.111115	−0.011549
3	0.124516	0.239611	0.211419	+0.115095	+0.086903	+0.028192
4	0.113655	0.208224	0.199146	+0.094569	+0.085491	+0.009078
5	0.119487	0.234917	0.239642	+0.115430	+0.120155	−0.004725
Mean, seeds 1–3	0.12268	0.22602	0.21201	+0.10334	+0.08933	+0.01401
Sample sd, seeds 1–3	0.00160	0.01192	0.02037	0.01039	0.02068	0.02218
Mean, seeds 1–5	0.12024	0.22424	0.21496	+0.10400	+0.09473	+0.00928
Sample sd, seeds 1–5	0.00409	0.01288	0.02071	0.01045	0.02046	0.01766
Positive, seeds 1–5				5/5	5/5	3/5

Table 18: Shared-path condition test, pilot budget, frozen OpenCLIP ViT-B/16, seeds 1–3 (post hoc; Appendix A.8, item (xvii)): the naive arm with one affine normalization shared by both heads, with mask labels (shared norm) and with both mask-loss weights zero (no labels). Internal validation AP@0.5 at the selected epoch (the tie check; naive and temporal-only are identical to all printed digits) and official AP@0.5 on the full 57,340-record split and the visual-manipulation subset; paired same-seed differences against the naive rows of Table 16; no interval.

	Internal AP@0.5		Official AP@0.5, full		Official AP@0.5, visual	
Seed	shared norm	no labels	shared norm	no labels	shared norm	no labels
1	0.410270	0.415524	0.077640	0.077164	0.116817	0.116177
2	0.415447	0.393516	0.078928	0.078000	0.118718	0.117154
3	0.414043	0.411927	0.078956	0.078967	0.118977	0.118988
Paired same-seed differences, official full split (seed 1 / 2 / 3; mean ± sample sd)						
Shared norm − naive	−0.003259 / −0.001877 / −0.003707			−0.00295 ± 0.00095 (0/3)		
No labels − shared norm	−0.000476 / −0.000928 / +0.000011			−0.00046 ± 0.00047 (1/3)		
No labels − naive	−0.003735 / −0.002805 / −0.003696			−0.00341 ± 0.00053 (0/3)		
Visual-manipulation subset						
No labels − shared norm	−0.000640 / −0.001564 / +0.000011			−0.00073 ± 0.00079 (1/3)		

$+0.0507$ / $+0.0201$ (3/3) and -0.0095 / $+0.0297$ / $+0.0444$ (2/3), and unmatched minus matched, the exposure effect at the same labels, is $+0.0591$ / $+0.1009$ / $+0.1044$ and $+0.0586$ / $+0.0971$ / $+0.0994$ (3/3 each); the mask head still degrades with dose (IoU@0.5 0.28–0.35 and 0.18–0.22).

Table 19: Pilot-budget controls: internal AP@0.5 at the best epoch (AV-Deepfake1M validation, frozen OpenCLIP ViT-B/16) per seed, and the paired same-seed difference from naive sum. Temporal-only equals naive sum to sixteen digits. Parameter-matched self-attention attends only to own-task features. Mask-label fractions are 36 and 4 of the 360 FaceForensics++ training pairs. $s{\to}t$-only rows are post hoc (Table 21). The historical ± 0.02 band describes seed spread, not equivalence; runs sharing a temporal forward function differ by up to 0.0589 here (attention without mask labels minus $s{\to}t$ only, no labels (the single-task attention readout of Section 5.2); contrast (3) of Table 21).

	AP@0.5			Δ vs naive sum		
Arm	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
naive sum	0.4403	0.4049	0.4270	—	—	—
temporal-only (no spatial task)	0.4403	0.4049	0.4270	0 (exact tie at every seed)		
Attention variants						
cross-task attention	0.6919	0.6984	0.6959	$+0.2516$	$+0.2935$	$+0.2689$
attention without mask labels	0.6971	0.7134	0.6869	$+0.2567$	$+0.3085$	$+0.2599$
attention module frozen at random init.	0.6031	0.5746	0.6138	$+0.1628$	$+0.1697$	$+0.1867$
$s{\to}t$ attention only (post hoc)	0.6861	0.6989	0.6770	$+0.2458$	$+0.2940$	$+0.2500$
$s{\to}t$ only, no labels (post hoc)	0.7004	0.6546	0.6919	$+0.2601$	$+0.2496$	$+0.2649$
attention, 10% of mask labels	0.7395	0.7487	0.7802	$+0.2992$	$+0.3438$	$+0.3532$
attention, 1% of mask labels	0.7600	0.7658	0.7508	$+0.3197$	$+0.3609$	$+0.3238$
parameter-matched self-attention	0.5212	0.5100	0.5198	$+0.0809$	$+0.1051$	$+0.0928$
Output-space consistency						
output-space consistency (w=0.1)	0.4674	0.4072	0.4348	$+0.0271$	$+0.0023$	$+0.0078$
without real-mass term	0.4612	0.4224	0.4064	$+0.0209$	$+0.0174$	-0.0207
without rank term	0.4687	0.4145	0.3834	$+0.0284$	$+0.0095$	-0.0436
without smoothness term	0.4670	0.4156	0.4197	$+0.0267$	$+0.0106$	-0.0074
without rank and smoothness	0.4122	0.4152	0.3936	-0.0281	$+0.0102$	-0.0334

Official label-control uncertainty (post hoc). Paired video bootstraps for every pilot seed and both official populations resolve full attention minus label-free attention in opposite directions (Table 20). Thus a difference within the historical internal noise band is not evidence of equivalence: the fitted model with mask supervision scores higher at some seeds and lower at another. The label-free gain over naive is positive throughout. These intervals condition on the trained models; they do not establish an expected effect over new training seeds, nor bound differences between runs that differ in data order and optimizer trajectory (Table 21, contrast (3)).

Table 20: Post-hoc official AP@0.5 differences and paired video-bootstrap 95% intervals, $B = 1000$, pilot budget. No-label denotes the full attention architecture with spatial loss zeroed. Every contrast on this closed list is shown; the $s{\to}t$-control list is in Table 21. Full and visual populations are resampled separately. Stable original-video ordering and duplicate-video equivalence are checked; every unresampled AP matches the independent official scorer within 10^{-6}.

Population	Seed	Full attention − no-label	No-label − naive
full	1	$+0.017197$ $[+0.015818, +0.018375]$	$+0.047409$ $[+0.045806, +0.049301]$
full	2	-0.007666 $[-0.008447, -0.006793]$	$+0.074836$ $[+0.072592, +0.077167]$
full	3	$+0.019289$ $[+0.017822, +0.020578]$	$+0.058539$ $[+0.056640, +0.060426]$
visual	1	$+0.025397$ $[+0.023340, +0.027078]$	$+0.069968$ $[+0.067454, +0.072772]$
visual	2	-0.011549 $[-0.012772, -0.010215]$	$+0.111115$ $[+0.107887, +0.114616]$
visual	3	$+0.028192$ $[+0.026196, +0.030150]$	$+0.086903$ $[+0.083831, +0.089792]$

Spatial-to-temporal attention only (post hoc). In every cross-task attention arm the temporal head reads a pooled frame vector, as in the baseline, plus a residual spatial-to-temporal ($s{\to}t$) attention readout over that frame's projected patch tokens; the temporal-to-spatial direction feeds only the mask decoder. Two post-hoc arms remove that direction, one with mask supervision, whose loss still reaches the shared trainable patch projection, and one without (Table 21; protocol and dates

in Appendix A.8). On the official full split, $s{\to}t$ attention only retains 81% of the mean attention gain over naive and $s{\to}t$ only without mask labels 76% (visual subset: 81% and 76%), positive at every seed, with every interval of the latter's gain over naive (contrast (0)) above zero on both populations. Full attention exceeds the no-label $s{\to}t$ arm at every seed with intervals excluding zero (contrast (1)); that gap is not attributed. Within the $s{\to}t$ architecture, whose two label settings share data order (Section 4), the label effect (contrast (2)) changes sign and its seed-3 intervals include zero. Contrast (4) is not a clean direction contrast: with mask labels, mask-loss gradients reach the attention readout, its two normalizations and the pooled-feature projection through the temporal-to-spatial direction, which exists only in the full arm, so (4) measures temporal-to-spatial attention together with mask-loss gradients into the attention readout; it is $+0.0120$ / -0.0004 / $+0.0274$, with seed-2 intervals including zero.

Contrast (3) compares two arms with the same temporal forward function. Without mask loss the temporal-to-spatial weights receive no nonzero gradient, stay bit-identical to their initialization at every seed, and never feed the temporal head. The two arms differ in data order, which is matched only within an architecture, and in optimizer trajectory: on mask batches, zero-valued gradients keep the attention readout, its normalizations and the pooled-feature projection in the optimizer's steps in the full architecture only. Data order is shared only within an architecture because each loader's per-epoch shuffle seed is drawn from the global random generator after the model is constructed, and construction consumes an architecture-dependent number of draws; an explicit per-loader generator would remove that difference. Contrast (3) nonetheless reaches 0.025 on the full split and 0.038 on the visual subset, with every interval excluding zero. Per-seed video-bootstrap intervals therefore do not bound differences between runs that differ in data order and optimizer trajectory, so contrasts (1) and (4) are not attributed. On internal AP (Table 19), full attention minus the no-label $s{\to}t$ arm is -0.0085 / $+0.0438$ / $+0.0040$, so the two metric families disagree in sign at seed 1. Internal contrast (3) is -0.0034 / $+0.0589$ / -0.0050, so the historical ± 0.02 band does not bound such differences either. Neither $s{\to}t$ arm was re-read at the last epoch or scored zero-shot.

Parameter-matched control, zero-shot (post hoc). The parameter-matched self-attention control, scored under the same frozen zero-shot protocols for the rank-association analysis, transfers best of all arms on TVIL (0.0680 / 0.0788 / 0.0787), and its DF40 rows carry the near-chance evidence flag (Table 22). It lacks patch-token access; that the arms with patch access transfer worse is consistent with that access, rather than added capacity, carrying the transfer cost.

Feature similarity of the pilot arms (post hoc). Linear CKA (Kornblith et al., 2019) between the per-frame features the temporal head reads, captured after any spatial-to-temporal attention from each arm's best checkpoint on the first 512 validation clips in file order (32 frames each: 16,384 rows of 256 features, columns centred; Appendix A.8, item (xv)). Identity checks: $\mathrm{CKA}(x, x) = 1.000$ for every arm, and CKA(naive, temporal-only)$= 1.000$ at every seed, the exact tie of Section 5.2, so the three temporal-only rows duplicate the naive rows and are omitted from Table 23. Same-seed attention against its label-free twin is 0.822 / 0.968 / 0.787 and against the single-task readout 0.806 / 0.832 / 0.789; the cross-seed attention-to-attention cells are 0.784 / 0.788 / 0.753 (label-free arm: 0.786 / 0.801 / 0.832). Both contrasts are at or above the attention range's minimum at every seed (Section 6.1). Without attribution: attention against naive is 0.543 / 0.451 / 0.445 and against the frozen module 0.569 / 0.523 / 0.537, whose features sit nearer naive's (0.854 / 0.828 / 0.815); the two $s{\to}t$-only label settings, which share data order, are among the most similar cross-arm pairs (0.955 / 0.937 / 0.943). Descriptive only; no interval; one similarity index, never a selection criterion.

Table 21: Spatial-to-temporal ($s \to t$) attention-only controls, post hoc (not pre-registered; Appendix A.8): official AV-Deepfake1M AP@0.5, pilot budget, frozen OpenCLIP ViT-B/16, fused decode, seeds 1–3, on the full 57,340-record split and its 43,054-record visual subset. $s \to t$ arms: temporal-to-spatial attention removed; with mask supervision on, the mask loss still reaches the shared trainable patch projection. No mask labels: mask-loss weights zero. Differences use unrounded AP. Paired video-bootstrap 95% intervals, $B = 1000$, percentile, RNG root 20260913, full and visual populations resampled separately; every cell of the closed five-contrast list is shown, and every unresampled AP matches the official scorer within 10^{-6}. Intervals condition on each trained seed and do not bound differences between runs that differ in data order and optimizer trajectory (contrast (3)).

	Full split			Visual subset		
Arm	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
Naive sum	0.080899	0.080805	0.082663	0.121979	0.121551	0.124516
Full attention	0.145505	0.147976	0.160491	0.217344	0.221117	0.239611
Attention, no mask labels	0.128308	0.155642	0.141202	0.191947	0.232666	0.211419
Attention, frozen module	0.116540	0.131075	0.136483	0.174175	0.195769	0.203868
$s \to t$ attention only	0.133519	0.148385	0.133084	0.199520	0.221731	0.198946
$s \to t$ only, no labels	0.139361	0.130505	0.133111	0.208402	0.194759	0.199099

Seed	Full split: Δ [95% interval]	Visual subset: Δ [95% interval]
(0) $s \to t$ only, no labels − naive		
1	$+0.058461$ $[+0.056398, +0.060429]$	$+0.086422$ $[+0.083491, +0.089383]$
2	$+0.049700$ $[+0.048009, +0.051689]$	$+0.073208$ $[+0.070655, +0.076012]$
3	$+0.050447$ $[+0.048603, +0.052597]$	$+0.074584$ $[+0.071920, +0.077731]$
(1) Full attention − $s \to t$ only, no labels		
1	$+0.006144$ $[+0.004938, +0.007373]$	$+0.008942$ $[+0.007182, +0.010698]$
2	$+0.017471$ $[+0.016102, +0.018702]$	$+0.026358$ $[+0.024310, +0.028475]$
3	$+0.027381$ $[+0.025937, +0.028750]$	$+0.040511$ $[+0.038166, +0.042380]$
(2) $s \to t$ attention only − $s \to t$ only, no labels (label effect; shared data order)		
1	-0.005841 $[-0.006673, -0.004869]$	-0.008881 $[-0.010128, -0.007427]$
2	$+0.017880$ $[+0.016759, +0.018875]$	$+0.026972$ $[+0.025358, +0.028328]$
3	-0.000027 $[-0.000985, +0.000715]$	-0.000154 $[-0.001523, +0.000972]$
(3) Attention, no mask labels − $s \to t$ only, no labels (same temporal forward function; different data order and optimizer trajectory)		
1	-0.011053 $[-0.012207, -0.009766]$	-0.016455 $[-0.017967, -0.014466]$
2	$+0.025136$ $[+0.023733, +0.026381]$	$+0.037907$ $[+0.035668, +0.039866]$
3	$+0.008092$ $[+0.006726, +0.009303]$	$+0.012319$ $[+0.010244, +0.014110]$
(4) Full attention − $s \to t$ attention only (temporal-to-spatial attention together with mask-loss gradients into the attention readout)		
1	$+0.011985$ $[+0.010575, +0.013216]$	$+0.017823$ $[+0.015792, +0.019658]$
2	-0.000409 $[-0.001623, +0.000790]$	-0.000613 $[-0.002410, +0.001357]$
3	$+0.027407$ $[+0.026068, +0.028780]$	$+0.040665$ $[+0.038605, +0.042645]$

Table 22: Zero-shot transfer of the pilot-budget controls (no tuning on either target): TVIL internal AP@0.5 and DF40 pooled face-local AUC (140 FaceForensics++ test originals as reals). d: the frozen score mapping barely separates real from fake in distribution, so the row is reported but not ranked. "= naive": identical to naive sum at every digit. The frozen-module and $s \to t$-only controls were scored in domain only.

	TVIL AP@0.5			DF40 pooled face-local AUC		
Arm	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
naive sum	0.0490	0.0487	0.0513	0.547^d	0.529^d	0.503^d
temporal-only (no spatial task)	= naive	= naive	= naive	= naive	= naive	= naive
Attention variants						
cross-task attention	0.0337	0.0281	0.0177	0.481	0.482	0.404^d
attention without mask labels	0.0345	0.0299	0.0276	0.549^d	0.403^d	0.394^d
attention, 10% of mask labels	0.0370	0.0415	0.0330	0.473^d	0.455^d	0.409^d
attention, 1% of mask labels	0.0335	0.0379	0.0291	0.647^d	0.465^d	0.365^d
parameter-matched self-attention	0.0680	0.0788	0.0787	0.1470^d	0.2602^d	0.3241^d
Output-space consistency						
output-space consistency (w=0.1)	0.0296	0.0323	0.0412	0.004^d	0.008^d	0.003^d
without real-mass term	0.0228	0.0336	0.0378	0.0070^d	0.0023^d	0.0027^d
without rank term	0.0245	0.0261	0.0480	0.0163^d	0.0031^d	0.0053^d
without smoothness term	0.0245	0.0347	0.0480	0.0024^d	0.0169^d	0.0012^d
without rank and smoothness	0.0388	0.0503	0.0406	0.0146^d	0.0052^d	0.0025^d

Table 23: Linear CKA between the features the temporal head reads, pilot arms, seeds 1–3, 512 validation clips (post hoc; Appendix A.8, item (xv)). N: naive sum; A: cross-task attention; L: attention without mask labels; F: attention with the module frozen at initialization; S: $s \to t$ attention only; U: single-task attention readout ($s \to t$ only, no mask labels); digit: seed. Temporal-only equals N at every seed and is omitted. Symmetric; descriptive only.

	N1	N2	N3	A1	A2	A3	L1	L2	L3	F1	F2	F3	S1	S2	S3	U1	U2	U3
N1	1.000	0.909	0.915	0.543	0.513	0.463	0.486	0.505	0.560	0.854	0.832	0.829	0.497	0.490	0.496	0.479	0.491	0.480
N2	0.909	1.000	0.897	0.530	0.451	0.456	0.455	0.447	0.525	0.841	0.828	0.840	0.484	0.453	0.461	0.464	0.465	0.443
N3	0.915	0.897	1.000	0.538	0.482	0.445	0.453	0.480	0.532	0.829	0.799	0.815	0.469	0.465	0.451	0.450	0.485	0.448
A1	0.543	0.530	0.538	1.000	0.784	0.788	0.822	0.804	0.831	0.569	0.565	0.574	0.804	0.786	0.761	0.806	0.822	0.792
A2	0.513	0.451	0.482	0.784	1.000	0.753	0.779	0.968	0.830	0.478	0.523	0.482	0.781	0.845	0.813	0.756	0.832	0.809
A3	0.463	0.456	0.445	0.788	0.753	1.000	0.806	0.771	0.787	0.488	0.495	0.537	0.808	0.799	0.792	0.809	0.813	0.789
L1	0.486	0.455	0.453	0.822	0.779	0.806	1.000	0.786	0.801	0.523	0.495	0.490	0.820	0.765	0.816	0.818	0.749	0.816
L2	0.505	0.447	0.480	0.804	0.968	0.771	0.786	1.000	0.832	0.475	0.522	0.482	0.781	0.850	0.815	0.760	0.853	0.817
L3	0.560	0.525	0.532	0.831	0.830	0.787	0.801	0.832	1.000	0.543	0.554	0.554	0.818	0.817	0.812	0.813	0.812	0.819
F1	0.854	0.841	0.829	0.569	0.478	0.488	0.523	0.475	0.543	1.000	0.922	0.923	0.505	0.474	0.495	0.494	0.483	0.485
F2	0.832	0.828	0.799	0.565	0.523	0.495	0.495	0.522	0.554	0.922	1.000	0.921	0.518	0.485	0.509	0.506	0.501	0.498
F3	0.829	0.840	0.815	0.574	0.482	0.537	0.490	0.482	0.554	0.923	0.921	1.000	0.510	0.498	0.486	0.498	0.520	0.476
S1	0.497	0.484	0.469	0.804	0.781	0.808	0.820	0.781	0.818	0.505	0.518	0.510	1.000	0.792	0.817	0.955	0.776	0.816
S2	0.490	0.453	0.465	0.786	0.845	0.799	0.765	0.850	0.817	0.474	0.485	0.498	0.792	1.000	0.805	0.761	0.937	0.805
S3	0.496	0.461	0.451	0.761	0.813	0.792	0.816	0.815	0.812	0.495	0.509	0.486	0.817	0.805	1.000	0.810	0.783	0.943
U1	0.479	0.464	0.450	0.806	0.756	0.809	0.818	0.760	0.813	0.494	0.506	0.498	0.955	0.761	0.810	1.000	0.752	0.816
U2	0.491	0.465	0.485	0.822	0.832	0.813	0.749	0.853	0.812	0.483	0.501	0.520	0.776	0.937	0.783	0.752	1.000	0.801
U3	0.480	0.443	0.448	0.792	0.809	0.789	0.816	0.817	0.819	0.485	0.498	0.476	0.816	0.805	0.943	0.816	0.801	1.000

A.3 BACKBONE REPLICATION, DATA SCALING, LABEL EFFICIENCY

Supports Sections 5.1 and 5.2. DINOv2 (Oquab et al., 2024), three arms and three seeds on a frozen backbone, reproduces the main paired orderings (Tables 24 and 15). The label ablation repeated on this backbone (post hoc, seeds 1–3, no intervals; Table 15) keeps 98% of the mean official attention gain without mask labels, the label-free arm above naive at every seed and attention − label-free -0.003268 / $+0.000006$ / $+0.006466$ on the full split (two of three positive). At the ten-fold budget (post hoc, item (xxi); lower blocks of Table 15) the label-free arm keeps 73% of the mean official attention gain: label-free − naive $+0.055763$ / $+0.024036$ / $+0.027728$ and attention − naive $+0.062315$ / $+0.043153$ / $+0.041475$, each positive at every seed. The two budgets order attention against its label-free twin differently: attention − label-free is $+0.006552$ / $+0.019117$ / $+0.013747$ at ten-fold, positive at every seed, against two of three at pilot; on the OpenCLIP backbone at ten-fold the label-free arm is at or above full attention at every seed (Table 1), so that ordering differs between the two backbones run. Orderings are compared across backbones, levels never are, since absolute AP is not comparable across feature spaces. The training-budget rows ($1\times \rightarrow 3\times \rightarrow 10\times$) support the single cross-budget object in the paper — the paired attention-minus-naive delta as a function of budget (Table 25): $+0.2713 \rightarrow +0.2311 \pm 0.0008 \rightarrow +0.1918 \pm 0.0178$ internally, a ≈29% relative reduction across a tenfold increase in training budget, positive at all nine seeds; two of the three $10\times$ attention runs, and one naive run, were still improving at their final epoch, so the $10\times$ point may not be converged. The mask-label efficiency rows (100% $\rightarrow$ 10% $\rightarrow$ 1% of spatial labels; Table 26) support the label-budget observation. On the second backbone the paired attention-minus-naive gain is $+0.274$ / $+0.320$ / $+0.310$ (mean $+0.301$) and the regularizer-minus-naive difference -0.002 / -0.005 / $+0.014$ (mean $+0.002$); its cross-generator detection probe orders baseline above attention far above the regularizer, and its inversion statistics, computed against its own leaderboard, are 0.004 (joint readout), 0.044 (generator shift) and 0.192 (inpainting transfer) (Figure 3). The second backbone reproduces the out-of-domain picture as well: the regularizer costs cross-paradigm transfer at all three seeds (-0.052 ± 0.002 paired against the baseline on the inpainting benchmark, the tightest negative in the study); the joint where-and-when readout puts it first at all three seeds while the in-domain leaderboard puts it level with the baseline (last at two of three seeds); and its cross-generator sign inversion — near-perfect separation with the score flipped — appears on all three second-backbone models, so it is observed on every one of the nine frozen-backbone models that carry it, across two backbones and two training budgets.

Two further replications support the scope statements of Section 5.1, each in its own block. **Second dataset pair** (Table 27). The three arms were retrained with FaceForensics++ masks paired with the video-inpainting benchmark's training split (3,471 clips of object removal, a face present in roughly one sampled clip in twenty) and read on its validation split (474 clips), pilot budget, three seeds. Paired differences from naive: attention $+0.0800 \pm 0.0230$, positive at every seed; regularizer $+0.0144 \pm 0.0340$, outside the ± 0.02 band at every seed, on both sides. Post hoc (Appendix A.8, item (xi)), attention without mask labels reaches $+0.0277$ / $+0.0918$ / $+0.0938$ over naive on the same readout, positive at every seed and 89% of the mean attention gain; full attention minus the label-free arm is $+0.0262$ / $+0.0059$ / -0.0053 (two of three positive); no interval; on the mirror readout below it sits between naive and full attention at every seed. The same nine models scored zero-shot on the lip-sync validation pool (43,054 clips) they never saw put the naive baseline first at every seed, the attention advantage reversed, all at near-floor levels that are read for ordering only. These models train on the inpainting benchmark and are therefore absent from Table 32's zero-shot block; their values are never placed in a column with the first pair's. **Unfrozen backbone** (Table 28). The three arms with the OpenCLIP backbone trained end to end at one tenth of the heads' learning rate, pilot schedule, three seeds; two of the nine runs (attention and regularizer at seed 1) resumed from their own checkpoints after interruption, with the optimizer state rebuilt. On the internal metric the tested recipe yields similar scores (paired differences from naive within ± 0.012 at every seed), and the same-seed lift over each arm's frozen value — the one paired quantity that crosses the frozen/unfrozen boundary — is $+0.21$ to $+0.22$ for attention against $+0.45$ to $+0.51$ for the other two arms. Scored post hoc with the official evaluation code (full split; driver and decode of Table 10), the paired differences from naive are mixed in sign and reach 0.037 in magnitude: $+0.0065$ / -0.0160 / $+0.0371$ for attention, whose frozen-feature difference (mean $+0.0699$) is positive at every seed, and $+0.0196$ / -0.0215 / $+0.0213$ for the regularizer. With a seed standard deviation of 0.027 for attention, three seeds cannot separate retention from loss of the advantage. Post hoc (Appendix A.8, item (x)), attention trained without mask labels under

Table 24: Frozen DINOv2 ViT-B/14 replication, pilot budget: in-domain internal AP@0.5, zero-shot TVIL AP@0.5, DF40 pooled face-local AUC and joint tube-mAP@0.2 (weak spatial reference: rankings may also be biased). Orderings, never levels, are compared with the OpenCLIP rows. d: the frozen score mapping barely separates real from fake in distribution, so the row is reported but not ranked. Mean ± sample sd over seeds; —: not reported.

Arm	Seed	Val AP@0.5	TVIL AP@0.5	DF40 AUC	tube-mAP@0.2
naive sum	1	0.4424	0.1175	0.415^{d}	9.1×10^{-5}
	2	0.4200	0.0834	0.479^{d}	1.5×10^{-4}
	3	0.4713	0.0996	0.490^{d}	5.9×10^{-3}
	mean	0.4446 ± 0.0258	0.1002 ± 0.0171	—	0.0020 ± 0.0033
cross-task attention	1	0.7167	0.0753	0.393^{d}	2.1×10^{-4}
	2	0.7403	0.0848	0.325	1.1×10^{-5}
	3	0.7809	0.0758	0.327^{d}	1.0×10^{-4}
	mean	0.7460 ± 0.0325	0.0786 ± 0.0053	—	0.00011 ± 0.00010
output-space	1	0.4402	0.0657	0.00047^{d}	2.8×10^{-3}
consistency (w=0.1)	2	0.4149	0.0297	0.00116^{d}	0.3362
	3	0.4852	0.0490	0.00079^{d}	0.3072
	mean	0.4468 ± 0.0356	0.0482 ± 0.0180	—	0.2154 ± 0.1847
Paired same-seed difference					
attention − naive	1	$+0.2743$	-0.0423	-0.022	—
	2	$+0.3203$	$+0.0014$	-0.154	—
	3	$+0.3096$	-0.0239	-0.163	—
	mean	$+0.3014 \pm 0.0241$	-0.0216 ± 0.0219	—	—
consistency − naive	1	-0.0022	-0.0518	-0.414	—
	2	-0.0051	-0.0536	-0.477	—
	3	$+0.0138$	-0.0506	-0.489	—
	mean	$+0.0022 \pm 0.0102$	-0.0520 ± 0.0015	—	—

Table 25: Paired same-seed attention-minus-naive difference in temporal AP@0.5 at each training-data budget ($1\times$, $3\times$, $10\times$ the pilot's), the only quantity compared across budgets. The internal and official-code blocks use different AP definitions and are never compared numerically. Mean ± sample sd over three paired seeds; the difference is positive at every seed in every block.

Budget	n	seed 1	seed 2	seed 3	mean ± sd
Internal AP@0.5, AV-Deepfake1M validation					
$1\times$	3	$+0.2516$	$+0.2935$	$+0.2689$	$+0.2713 \pm 0.0211$
$3\times$	3	$+0.2318$	$+0.2303$	$+0.2311$	$+0.2311 \pm 0.0008$
$10\times$	3	$+0.2067$	$+0.1967$	$+0.1721$	$+0.1918 \pm 0.0178$
Official evaluation code, AP@0.5, full split (57,340 records)					
$1\times$	3	$+0.064606$	$+0.067171$	$+0.077828$	$+0.06987 \pm 0.00701$
$3\times$	3	$+0.047384$	$+0.048971$	$+0.061250$	$+0.05254 \pm 0.00759$
$10\times$	3	$+0.033494$	$+0.018183$	$+0.023444$	$+0.02504 \pm 0.00778$
Official evaluation code, AP@0.5, visual-manipulation subset (43,054 records)					
$1\times$	3	$+0.095365$	$+0.099566$	$+0.115095$	$+0.10334 \pm 0.01039$
$3\times$	3	$+0.069721$	$+0.071910$	$+0.090071$	$+0.07723 \pm 0.01117$
$10\times$	3	$+0.049983$	$+0.027169$	$+0.034998$	$+0.03738 \pm 0.01159$

Table 26: Mask-label efficiency, pilot budget, frozen OpenCLIP ViT-B/16: cross-task attention trained with all 360, 36 (10%) or 4 (1%) FaceForensics++ mask pairs. Temporal internal AP@0.5 with paired same-seed differences; mask IoU@0.5 and L1 on the full FaceForensics++ validation split at the selected epoch. Validation data are identical across doses; the smaller mask pools run out early in each epoch, after which every batch is temporal, so reduced-dose arms train on more temporal batches. The two exposure-matched blocks (post hoc, item (xix)) re-iterate the reduced pool so that the spatial and temporal batch counts match the full-dose arm's. Mean ± sample sd over seeds in the last block.

Arm (mask pairs)	Seed	AP@0.5	Δ vs naive	full − this	IoU@0.5	L1
naive sum (360)	1	0.4403	—	—	0.4072	0.0306
	2	0.4049	—	—	0.4403	0.0423
	3	0.4270	—	—	0.3976	0.0307
attention (360)	1	0.6919	+0.2516	—	0.4587	0.0322
	2	0.6984	+0.2935	—	0.4475	0.0504
	3	0.6959	+0.2689	—	0.4026	0.0315
attention (36, 10%)	1	0.7395	+0.2992	−0.0476	0.2948	0.1986
	2	0.7487	+0.3438	−0.0503	0.3652	0.2315
	3	0.7802	+0.3532	−0.0843	0.3559	0.1839
attention (4, 1%)	1	0.7600	+0.3197	−0.0681	0.3254	0.3427
	2	0.7658	+0.3609	−0.0674	0.3507	0.3938
	3	0.7508	+0.3238	−0.0549	0.3115	0.3163
attention (36, 10%), exposure-matched	1	0.6805	+0.2402	+0.0114	0.2859	0.0399
	2	0.6477	+0.2428	+0.0507	0.3482	0.0874
	3	0.6758	+0.2488	+0.0201	0.2792	0.0422
attention (4, 1%), exposure-matched	1	0.7014	+0.2611	−0.0095	0.2130	0.0525
	2	0.6687	+0.2638	+0.0297	0.2217	0.0782
	3	0.6515	+0.2245	+0.0444	0.1836	0.0712

Paired same-seed difference	mean ± sd	sign
attention (360) − naive	$+0.2713 \pm 0.0211$	3/3 positive
attention (36) − naive	$+0.3321 \pm 0.0288$	3/3 positive
attention (4) − naive	$+0.3348 \pm 0.0227$	3/3 positive
attention (360) − attention (36)	-0.0607 ± 0.0205	3/3 negative
attention (360) − attention (4)	-0.0635 ± 0.0074	3/3 negative
attention (36, matched) − naive	$+0.2439 \pm 0.0044$	3/3 positive
attention (4, matched) − naive	$+0.2498 \pm 0.0220$	3/3 positive
attention (360) − attention (36, matched)	$+0.0274 \pm 0.0206$	3/3 positive
attention (360) − attention (4, matched)	$+0.0216 \pm 0.0279$	2/3 positive
attention (36) − attention (36, matched)	$+0.0881 \pm 0.0253$	3/3 positive
attention (4) − attention (4, matched)	$+0.0850 \pm 0.0229$	3/3 positive

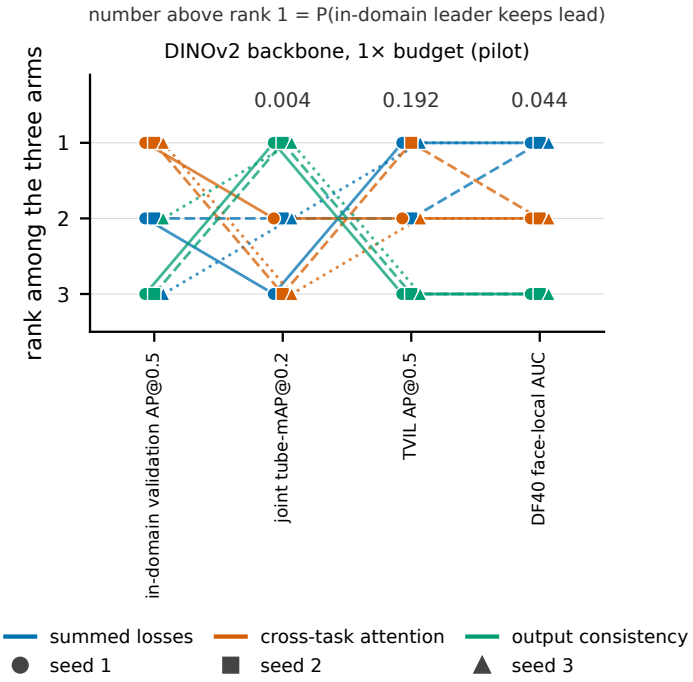


Figure 3: DINOv2-backbone mirror of Figure 10(a) at the pilot budget: the three arms ranked by in-domain validation AP@0.5 and by each out-of-domain readout, three seeds; the numbers above the transfer columns are the inversion statistic of Section 4 (0.004 joint readout, 0.192 inpainting transfer, 0.044 generator shift). Orderings only; levels are never compared across backbones.

the same recipe scores 0.2447 / 0.2414 / 0.2526, $+0.0147$ / -0.0173 / $+0.0384$ against naive with full attention's sign pattern, and full attention minus this label-free arm is -0.0082 / $+0.0013$ / -0.0013 (one of three positive; same signs on the visual subset, AR@10 differences at most 0.0041); no interval or retained fraction is computed, since attention minus naive is not positive at every seed. The official same-seed lift keeps the internal pattern, attention lifting least at every seed: $+0.09$ for attention against $+0.13$ to $+0.18$ for naive and $+0.15$ to $+0.16$ for the regularizer. On AR@10 (full split) every unfrozen difference from naive is at most 0.0033 in magnitude, against $+0.017$ to $+0.019$ for attention on frozen features; AP@0.5 on the visual-manipulation subset gives the same signs, with larger differences (up to 0.057). Levels are not compared across regimes. Every unfrozen model scores below 0.5 in cross-generator pooled face-local AUC, the regularizer's (0.005–0.026) far below the others' (0.13–0.50), so the block is reported as an observation, not a ranking. Unfrozen rows never share a ranked column with frozen rows. **Second fine-tuning recipe** (Table 30; post hoc, Appendix A.8, item (xiv)). The four arms were retrained with the backbone at one hundredth of the heads' learning rate, every other setting unchanged, three seeds, none interrupted, under a reading rule fixed before any run existed. On the official evaluation code (full split) attention minus naive is $+0.0337$ / -0.0073 / $+0.0082$ (two of three seeds positive), the regularizer's difference $+0.0228$ / $+0.0012$ / -0.0028 (two of three), label-free attention minus naive $+0.0339$ / -0.0037 / $+0.0045$ (two of three), and full attention minus the label-free arm -0.0003 / -0.0037 / $+0.0037$ (one of three); the visual subset gives the same signs, and AR@10 differences are at most 0.0023 in magnitude. Attention minus naive does not keep one sign, so, as at one tenth, there is no consistent ordering between naive and either attention arm, and no retained fraction or interval is computed. In the order frozen, one hundredth, one tenth, attention minus naive is positive at three, two and two of three seeds. The same-seed lift over the frozen arm is again smallest for the two attention arms at every seed ($+0.09$ to $+0.10$ for attention and $+0.09$ to $+0.12$ without mask labels, against $+0.13$ to $+0.17$ for naive and $+0.15$ to $+0.16$ for the regularizer), and the same-seed difference between the two recipes is mixed in sign for every arm (at most 0.028 in magnitude). On the internal metric attention minus naive is $+0.0121$ / -0.0044 / $+0.0098$. **Uninterrupted seed-1 reruns** (Table 28; post hoc, item (xx)). The two resumed seed-1 runs of the one-tenth recipe were retrained from scratch without interruption under unchanged configurations; nothing is replaced. On the official full split the rerun scores 0.2216 for attention (resumed 0.2365) and 0.2302 for the regularizer (resumed 0.2496), so the seed-1 differences from naive are -0.0083 (resumed $+0.0065$) and $+0.0003$ (resumed $+0.0196$): with the reruns attention minus naive is positive at one of three seeds instead of two, and the regularizer's difference at two of three as before. Rerun and resumed values differ by 0.0149 and 0.0193, more than the seed-2/3 spread of each arm (0.0086; 0.0017), so the resumption is a plausible contributor to the seed-1 values. It cannot be isolated: each rerun already departs from its interrupted original during the first epoch, before any interruption (internal validation AP 0.620 against 0.783 for attention after one epoch), so two runs of one configuration and seed already differ under this recipe, a second contributor whose source (nondeterministic kernels, or code revised between the runs) was not separated. The reading is unchanged: no consistent ordering under either recipe. **Ten-fold budget** (Table 29; post hoc, item (xxiv)). The one-tenth recipe was run at the ten-fold batch budget of Table 2, every other setting unchanged, for the runs the schedule allowed: naive and label-free attention at seeds 1 and 2 and full attention at seed 1; no seed was added after reading. On the official evaluation code (full split) the arms score 0.2138 / 0.2159 (naive), 0.1884 / 0.1900 (attention without mask labels) and 0.1886 (attention, seed 1): label-free attention minus naive is -0.0254 / -0.0259, negative at both landed seeds; attention minus naive is -0.0253 at its one seed, and attention minus the label-free arm $+0.0002$. The visual subset gives the same signs (-0.0381 / -0.0387; -0.0376), and AR@10 differences are at most 0.0029 in magnitude; with one attention seed, no retained fraction or interval is computed. Under this one untuned end-to-end recipe the ten-fold level is 0.19–0.22 for our arms, beside 0.523 for the OmniFD recipe retrained on our pair and 75.9% for the fully supervised anchor of Table 1, each under its own protocol; the levels are described, not compared. The same-seed lift over the frozen ten-fold arm is $+0.0763$ / $+0.0640$ for naive against $+0.0176$ for attention and $+0.0173$ / $+0.0102$ for the label-free arm, attention lifting least as at the pilot budget; the same-arm difference between the ten-fold and the pilot unfrozen runs is negative for every run (-0.0162 / -0.0428 naive, -0.0479 attention, -0.0563 / -0.0515 label-free). On the internal metric the best-epoch values are 0.9546 / 0.9550, 0.9511 and 0.9489 / 0.9493, each attention arm below naive by 0.0035 to 0.0057. Neither lift is a ranking, and levels are not compared across regimes or budgets.

Table 27: Second dataset pair, pilot budget, frozen OpenCLIP: the three arms trained on FaceForensics++ masks with TVIL training segments (3,471 clips), read on TVIL validation (474 clips; internal AP@0.5) and zero-shot on AV-Deepfake1M validation (43,054 clips; near-floor, orderings only). Trained on TVIL, these models are absent from the zero-shot TVIL table. A face-local spatial task is coupled to a largely non-facial temporal task. Never ranked with first-pair rows; mean ± sample sd; differences from unrounded AP. The label-free attention rows are post hoc (Appendix A.8, item (xi)), no interval.

	TVIL validation AP@0.5			Δ vs naive sum			
Arm	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3	mean ± sd
naive sum	0.5653	0.5418	0.5229	—	—	—	—
cross-task attention	0.6193	0.6395	0.6113	+0.0540	+0.0976	+0.0885	+0.0800 ± 0.0230
attention, no mask labels (post hoc)	0.5931	0.6336	0.6167	+0.0277	+0.0918	+0.0938	+0.0711 ± 0.0376
output-space consistency (w=0.1)	0.5427	0.5636	0.5671	−0.0226	+0.0217	+0.0442	+0.0144 ± 0.0340

	AV-Deepfake1M validation AP@0.5, zero-shot		
Arm	seed 1	seed 2	seed 3
naive sum	0.00167	0.00155	0.00138
cross-task attention	0.00100	0.00101	0.00070
attention, no mask labels (post hoc)	0.00130	0.00142	0.00133
output-space consistency (w=0.1)	0.00081	0.00033	0.00043

Table 28: Unfrozen OpenCLIP ViT-B/16 backbone (one tenth of the heads' learning rate, pilot schedule), three seeds; one recipe, fixed before any run, never tuned. Internal AP and official evaluation-code AP (scored post hoc with the split, driver and decode of Table 10; Appendix A.8) are separate metric families, never ranked against each other; differences are computed from unrounded scores and the internal-metric lifts from the tabulated four-decimal levels. Unfrozen rows are never ranked against frozen-backbone rows; the same-seed lift is the one quantity crossing that boundary. DF40 AUC is an observation, not ranked. †: resumed from its own checkpoint after interruption, optimizer state rebuilt. The label-free attention rows are post hoc (Appendix A.8, item (x)), trained without interruption; no interval. ‡: seed 1 retrained from scratch without interruption (post hoc, item (xx)); these rows replace nothing and enter no mean. The same recipe at the ten-fold batch budget is Table 29.

Arm	seed 1	seed 2	seed 3
Internal AP@0.5, AV-Deepfake1M validation, best epoch			
naive sum	0.9143	0.9191	0.9035
cross-task attention	0.9114†	0.9097	0.9154
output-space consistency (w=0.1)	0.9214†	0.9173	0.9124
attention, uninterrupted rerun‡	0.9195	—	—
consistency, uninterrupted rerun‡	0.9174	—	—
Internal: paired same-seed difference from naive sum			
cross-task attention	−0.0028	−0.0094	+0.0119
output-space consistency (w=0.1)	+0.0072	−0.0017	+0.0089
attention, uninterrupted rerun‡	+0.0052	—	—
consistency, uninterrupted rerun‡	+0.0031	—	—
Internal: same-seed lift, unfrozen minus frozen, same arm			
naive sum	+0.4740	+0.5142	+0.4765
cross-task attention	+0.2195	+0.2113	+0.2195
output-space consistency (w=0.1)	+0.4540	+0.5101	+0.4776
Official evaluation code (post hoc), AP@0.5, full split			
naive sum	0.2300	0.2587	0.2143
cross-task attention	0.2365†	0.2427	0.2514
attention, no mask labels (post hoc)	0.2447	0.2414	0.2526
output-space consistency (w=0.1)	0.2496†	0.2372	0.2355
attention, uninterrupted rerun‡	0.2216	—	—
consistency, uninterrupted rerun‡	0.2302	—	—
Official: paired same-seed difference from naive sum			
cross-task attention	+0.0065	−0.0160	+0.0371
output-space consistency (w=0.1)	+0.0196	−0.0215	+0.0213
attention, no mask labels (post hoc)	+0.0147	−0.0173	+0.0384
attention, uninterrupted rerun‡	−0.0083	—	—
consistency, uninterrupted rerun‡	+0.0003	—	—
Official: cross-task attention minus attention without mask labels (post hoc)			
paired same-seed difference	−0.0082	+0.0013	−0.0013
Official: same-seed lift, unfrozen minus frozen, same arm			
naive sum	+0.1491	+0.1779	+0.1316
cross-task attention	+0.0910	+0.0947	+0.0909
output-space consistency (w=0.1)	+0.1612	+0.1520	+0.1530
attention, uninterrupted rerun‡	+0.0761	—	—
consistency, uninterrupted rerun‡	+0.1418	—	—
DF40 pooled face-local AUC (observation)			
naive sum	0.1290	0.3635	0.2785
cross-task attention	0.1985	0.1715	0.4951
output-space consistency (w=0.1)	0.0047	0.0264	0.0062

Table 29: Unfrozen OpenCLIP ViT-B/16 backbone at the ten-fold batch budget of Table 2: the one-tenth recipe of Table 28, every other setting unchanged, for the runs made (post hoc, Appendix A.8, item (xxiv)): naive and label-free attention at seeds 1 and 2, full attention at seed 1; one recipe fixed before any run, never tuned; no seed added after reading. Official evaluation-code AP@0.5 (split, driver and decode of Table 10) and internal AP are separate metric families, never ranked against each other; differences are computed from unrounded scores. Ten-fold rows are never ranked against pilot rows, and unfrozen rows never against frozen rows; the two lifts, over the frozen ten-fold arm of the same seed (Table 2) and over the pilot unfrozen run of the same arm and seed (Table 28), are descriptive. —: not run.

Arm	seed 1	seed 2
Official evaluation code, AP@0.5, full split		
naive sum	0.2138	0.2159
cross-task attention	0.1886	—
attention, no mask labels	0.1884	0.1900
Official evaluation code, AP@0.5, visual-manipulation subset		
naive sum	0.3187	0.3215
cross-task attention	0.2811	—
attention, no mask labels	0.2805	0.2828
Official: paired same-seed difference from naive sum, full split		
attention, no mask labels (0 of 2 positive)	-0.0254	-0.0259
cross-task attention (0 of 1 positive)	-0.0253	—
Official: cross-task attention minus attention without mask labels		
paired same-seed difference (one seed)	$+0.0002$	—
Official: same-seed lift, unfrozen minus frozen ten-fold arm (Table 2)		
naive sum	$+0.0763$	$+0.0640$
cross-task attention	$+0.0176$	—
attention, no mask labels	$+0.0173$	$+0.0102$
Official: same-arm difference, ten-fold minus pilot budget (both unfrozen)		
naive sum	-0.0162	-0.0428
cross-task attention	-0.0479	—
attention, no mask labels	-0.0563	-0.0515
Internal AP@0.5, AV-Deepfake1M validation, best epoch (diagnostic)		
naive sum	0.9546	0.9550
cross-task attention	0.9511	—
attention, no mask labels	0.9489	0.9493

Table 30: Second fine-tuning recipe (post hoc; Appendix A.8, item (xiv)): unfrozen OpenCLIP ViT-B/16 backbone at one hundredth of the heads' learning rate, every other setting as in Table 28; pilot schedule, three seeds, no run interrupted; reading rule fixed before any run existed, recipe never tuned. Official evaluation-code AP@0.5, full split (split, driver and decode of Table 10); differences are computed from unrounded scores; no interval or retained fraction, since attention minus naive is not positive at every seed. The block is separate from the one-tenth recipe of Table 28 and never ranked against frozen-backbone rows; the same-seed lift is the one quantity crossing that boundary. †: the one-tenth value is the resumed run of Table 28. Internal AP is a diagnostic in a separate metric family.

Arm	seed 1	seed 2	seed 3
Official evaluation code (post hoc), AP@0.5, full split			
naive sum	0.2149	0.2467	0.2420
cross-task attention	0.2485	0.2394	0.2501
attention, no mask labels	0.2488	0.2430	0.2465
output-space consistency (w=0.1)	0.2377	0.2479	0.2391
Official: paired same-seed difference from naive sum			
cross-task attention	$+0.0337$	-0.0073	$+0.0082$
output-space consistency (w=0.1)	$+0.0228$	$+0.0012$	-0.0028
attention, no mask labels	$+0.0339$	-0.0037	$+0.0045$
Official: cross-task attention minus attention without mask labels			
paired same-seed difference	-0.0003	-0.0037	$+0.0037$
Official: same-seed lift, unfrozen minus frozen, same arm			
naive sum	$+0.1340$	$+0.1659$	$+0.1593$
cross-task attention	$+0.1030$	$+0.0914$	$+0.0896$
attention, no mask labels	$+0.1205$	$+0.0874$	$+0.1053$
output-space consistency (w=0.1)	$+0.1493$	$+0.1627$	$+0.1566$
Official: same-seed difference, one hundredth minus one tenth, same arm			
naive sum	-0.0151	-0.0120	$+0.0277$
cross-task attention	$+0.0120^{\dagger}$	-0.0034	-0.0012
attention, no mask labels	$+0.0041$	$+0.0016$	-0.0062
output-space consistency (w=0.1)	$-0.0118^{\dagger}$	$+0.0107$	$+0.0036$
Internal AP@0.5, AV-Deepfake1M validation, best epoch (diagnostic)			
naive sum	0.9159	0.9273	0.9147
cross-task attention	0.9280	0.9229	0.9245
attention, no mask labels	0.9267	0.9237	0.9205
output-space consistency (w=0.1)	0.9195	0.9256	0.9224

A.4 TRANSFER AND COMPANION READOUTS IN FULL

Supports Section 5.3. Table 31 gives the joint where-and-when readout of Section 5.3 per seed and budget (moved here from the main text; its pilot rows also appear with mask IoU, TP counts and stability in Table 41). Table 32 gives the TVIL zero-shot readout of Section 5.3 per seed and budget (also moved here from the main text).

Table 31: Joint real-manipulation readout: tube-mAP@0.2 per seed on 286 manipulated + 100 real AV-Deepfake1M clips with landmark-derived mouth-region masks (weak references that may bias both levels and rankings), and the paired same-seed delta against naive (mean; sign $n + /m-$). Mask IoU inside true-positive segments, TP counts and stability: Table 41. Footers: inversion statistic (Section 4), per-seed fractions in brackets.

Arm	tube-mAP@0.2, s1/s2/s3	Δ vs naive (sign)
Pilot budget, seeds 1–3		
naive sum	0.0319/0.0085/0.0366	—
cross-task attention	0.0081/0.0030/0.0509	-0.0050 (1+/2−)
output-space consistency	0.0989/0.0617/0.0088	$+0.0308$ (2+/1−)
inversion statistic 0.294 (per seed 0.000/0.000/0.883)		
Ten-fold budget, seeds 1–3		
naive sum	0.0156/0.0451/0.0181	—
cross-task attention	0.0063/0.0637/0.0959	$+0.0290$ (2+/1−)
output-space consistency	0.0137/0.0136/0.0066	-0.0150 (0+/3−)
inversion statistic 0.634 (per seed 0.014/0.887/1.000)		

Table 32: Zero-shot temporal transfer to TVIL (video inpainting): no target-set tuning, calibration or selection; internal AP@0.5. AR@10 column: designated secondary, min–max over seeds; pilot-budget rows give the mean only; published in-domain bars omitted. Deltas: paired same-seed, mean and sign $n + /m-$; footers: inversion statistic. Moved here from the main text; DF40 and OpenForensics per-seed readouts (Section 5.3) follow in this appendix.

TVIL, zero-shot temporal (AP@0.5)				
Arm	AP@0.5, s1/s2/s3	mean (± sd)	Δ vs naive (sign)	AR@10, range
pilot budget				
naive sum	0.0490/0.0487/0.0513	0.0497	—	0.3169–0.3326
cross-task attention	0.0337/0.0281/0.0177	0.0265	-0.0232 (0+/3−)	0.2225–0.2539
output-space consistency	0.0296/0.0323/0.0412	0.0344	-0.0153 (0+/3−)	0.2764–0.2831
inversion statistic 0.029				
ten-fold budget				
naive sum	0.0612/0.0521/0.0455	0.0529 ± 0.0079	—	0.3191–0.3551
cross-task attention	0.0719/0.0221/0.0189	0.0376 ± 0.0297	-0.0153 (1+/2−)	0.2382–0.3461
output-space consistency	0.0344/0.0175/0.0247	0.0255 ± 0.0085	-0.0274 (0+/3−)	0.2809–0.3326
inversion statistic 0.260				

- **DF40 per-paradigm tables.** Table 34 gives the AUC per paradigm, arm and seed with the in-distribution evidence AUC behind each flag, and Table 35 the generator coverage (the release defines 40 generators; 39 are in our evaluation pool and 37 were readable); Table 33 gives the pooled face-local AUC per arm and seed at both budgets. The pooled face-local AUC scores 35,688 manipulated videos against the 140 FaceForensics++ test originals. Rows whose frozen score mapping barely separates real from fake on FaceForensics++ validation (evidence AUC ≤ 0.55) carry a flag and are read only as sign/ordering evidence; the main text's sign-inversion finding is stated under that flag. The $10\times$ naive-vs-attention ordering behind the Section 5.3 "seed-unstable" sentence: attention ahead at seed 1 (0.551 vs 0.525), naive ahead at seeds 2 and 3 (0.499 vs 0.377; 0.602 vs 0.397) — no conclusion is drawn from it. Behind the Section 5.3 sign-inversion sentence: pooled face-local AUC 0.001 for each $10\times$ regularizer checkpoint (seed 2: 0.0007) against naive and attention at 0.377–0.602; the DF40 inversion statistic 0.307 at $10\times$ splits per seed as 0.92 / 0.00 / 0.00.
- **TVIL per-seed results, AR@10 and per-method slice.** Table 36 gives AP@0.5 and the designated secondary AR@10 per seed at both budgets on the 508-clip test split, and Table 37 the per-method breakdown at the pilot budget. Published in-domain TVIL results are

quoted only with their protocol (in-domain training vs our zero-shot; Appendix A.7). At the pilot budget attention trails naive at all three seeds (the inversion statistic 0.029 means the in-domain leader keeps its lead in 2.9% of resample $\times$ seed draws), and the secondary AR@10 agrees at every seed. At the ten-fold budget, seed 1 has attention 0.0719 against naive 0.0612 and regularizer 0.0344, with paired clip-level bootstrap intervals ($B = 1000$, percentile) of $+0.0106$ $[-0.0194, +0.0408]$ for attention $-$ naive (compatible with zero) and -0.0268 $[-0.0410, -0.0122]$ for regularizer $-$ naive; seeds 2 and 3 keep the pilot ordering. The $10\times$ inversion statistic 0.260 splits per seed as 0.77 / 0.01 / 0.00.

- **OpenForensics.** Tables 38 and 39 give the cross-dataset mask-transfer results per seed on Test-Dev (18,895 images) and Test-Challenge (45,000 images) under the whole-image union-mask protocol, behind the Section 5.3 retention paragraph. The regularizer's paired Test-Dev deficit of -0.120 ± 0.058 reproduces on Test-Challenge (-0.128 ± 0.058); the OpenForensics inversion statistic 0.561 splits per seed as 0.68 / 1.00 / 0.00; the fraction of in-domain mask quality retained out of domain is 23–44% for the baseline, 15–48% for attention and 4–8% for the regularizer; scoring the in-domain side on manipulated images alone would make each a third lower, with the ordering of arms unchanged, since the validation reals, a third of the images, score zero there (Appendix A.10) while almost every OpenForensics image contains a forged face. Attention against the baseline on Test-Dev is -0.004 ± 0.026, mixed in sign. The regularizer's cross-dataset value-agreement fraction (`ioudiff@0.01`) of 0.882 / 0.182 / 0.488 on Test-Dev tracks the same checkpoints' in-domain FaceForensics++ values of 0.924 / 0.291 / 0.572 in both ordering and magnitude — the regularizer's characteristic value sharpening travels across the domain shift, the seed spread reflecting how early the selected epoch fell rather than any effect of the shift (seeds 2 and 3 were selected at an epoch before the sharpening completed). In the ten-fold block every Test-Dev IoU@0.5 lies below 0.05 (Test-Challenge: at most marginally above it), so that block is read for ordering only. Two caveats attach to every OpenForensics number: FaceForensics++ mask geometry differs by manipulation family (bounding box, face region, or inner face) from OpenForensics' swapped-face outlines, and images are resized without preserving aspect ratio, so these IoUs are conservative throughout.
- **Mask-dose companions on both temporal probes.** Reported here only; pilot budget, three seeds (Table 22). TVIL AP@0.5 arm means: 0.050 (baseline), 0.037 (10% of the mask pairs), 0.034 (1%), 0.027 (full dose); paired same-seed deltas against the baseline -0.013 ± 0.006 at 10% and -0.016 ± 0.006 at 1%, sign 3/3 in both cases. DF40 pooled face-local AUC for the same six checkpoints: 0.473 / 0.455 / 0.409 at the 10% dose, within 0.03 of the matched full-dose attention rows at every seed, and 0.647 / 0.465 / 0.365 at the 1% dose (seed-unstable); all six sit in the near-chance band of the other non-regularizer arms.
- **Mask stability descriptives.** Table 40 gives adjacent-frame flicker and stability of predicted masks per arm and seed, reported as descriptive companions: adjacency is in the sampled-clip sequence, and without motion compensation these numbers cannot separate model instability from content change — they inform the discussion and select nothing.
- **Facial-kinematics descriptives and the audio-laundering probe.** The landmark-kinematics analysis of the evaluation clips (tracker-level measurements, not physiological claims). Behind the kinematics observation of Section 6.3: on the training dataset's four manipulation families — face swapping (autoencoder), expression reenactment, neural rendering, and a graphics-pipeline face swap — each manipulated video is compared against its own source video, so the design is between-video rather than within-clip and answers the same question in a weaker form. The paired contrast of high-percentile lip-jerk is negative at both smoothing windows in three of the four families (the autoencoder face swap, the expression reenactment and the neural rendering), and in those three it holds on all four landmark features (lip speed, acceleration, jerk and mouth-opening rate) — the breadth the cross-generator lip-sync replication lacked, where the effect was carried by lip jerk alone (the frozen-set population shares the sign on all four features at both windows); the graphics-pipeline family is the exception, recorded as a hypothesis-generating observation. Families are reported separately, never pooled, and only the sign is read across corpora. Two further generator families could not be measured at all: one publishes only sparse sampled frames, on which per-frame motion derivatives are undefined, and the other

manipulates non-facial regions, so no lip track exists to compare. Beside it, the audio-perturbation probe on the one audio-visual checkpoint we trained (with the label-cache defect of Appendix A.7): replacing the audio track with silence, matched noise, or another clip's audio (video streams bit-identical) moves its re-ranked AP@0.5 by under 1% relative on all three conditions on a 6,000-record subsample (Figure 4) — this reproduction leans on its visual stream, so the vision-only threat-model argument rests on the published audio-visual systems' own audio-dominance ablations, an extrapolation we label as such rather than a measurement we can make on their checkpoints.

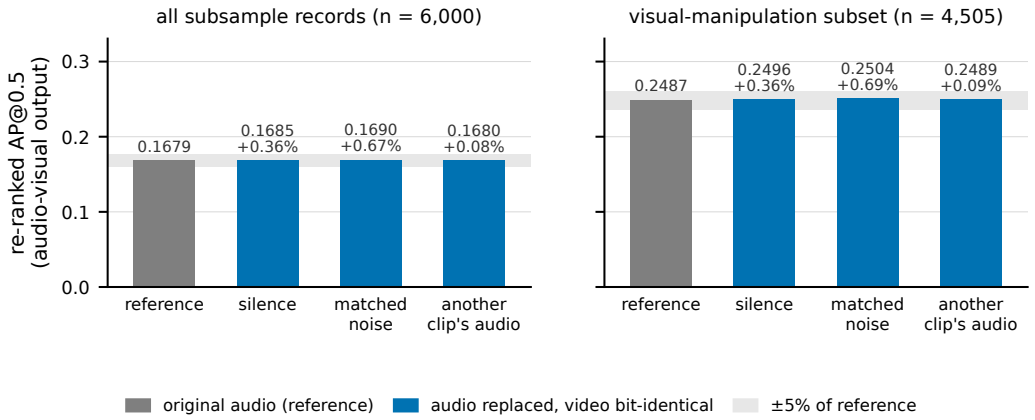


Figure 4: Audio-perturbation probe on the one audio-visual checkpoint we trained, with the label-cache defect of Appendix A.7: re-ranked AP@0.5 with the original audio (grey) and with the audio replaced by silence, matched noise or another clip's audio (blue; video streams bit-identical), on every record of the 6,000-record subsample (left) and on its visual-manipulation subset (right). The band marks $\pm 5\%$ of the reference, the pre-registered near-invariance bracket (a classification, not a significance test); raw-decode AP@0.5 stays at 0.014–0.016 in every condition. One checkpoint, one pass: the near-invariance describes this reproduction, and any link to published audio-visual systems is an extrapolation.

The audio-perturbation probe and the kinematics distributions are shown in Figures 4 and 9.

Table 33: DF40 zero-shot detection, pooled face-local AUC per seed (the per-seed detail behind Section 5.3). Reals are FaceForensics++ test originals; flag d: the frozen score mapping barely separates real from fake in distribution (evidence AUC ≤ 0.55), so the row is reported but not ranked. Budgets are never ranked against each other.

DF40, zero-shot detection (pooled face-local AUC)		
Arm	AUC, s1/s2/s3	flag, s1/s2/s3
pilot budget		
naive sum	0.547/0.529/0.503	d/d/d
cross-task attention	0.481/0.482/0.404	–/–/d
output-space consistency	0.004/0.008/0.003	d/d/d
consistency, no real-mass inversion statistic 0.001	0.007/0.002/0.003	d/d/d
ten-fold budget		
naive sum	0.525/0.499/0.602	d/d/d
cross-task attention	0.551/0.377/0.397	d/–/–
output-space consistency inversion statistic 0.307	0.001/0.001/0.001	d/d/d

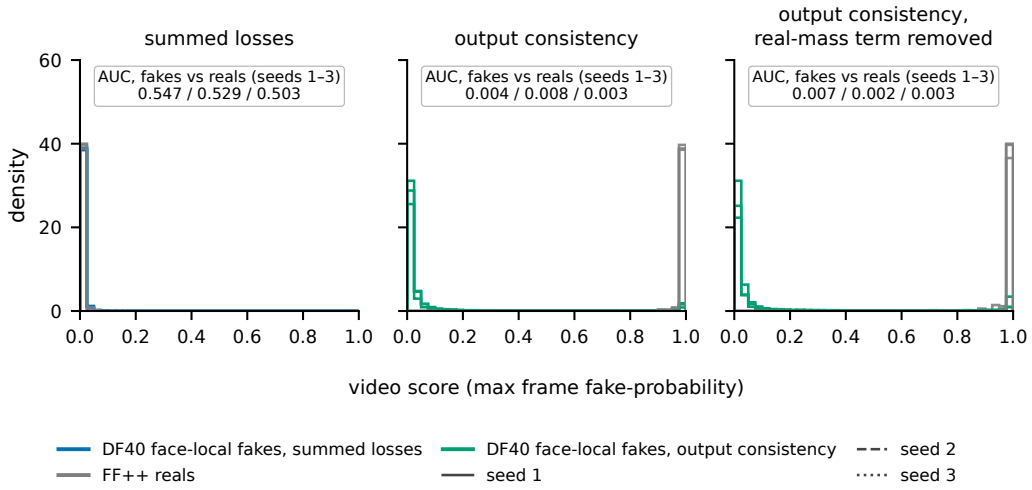


Figure 5: Why the cross-generator AUC is near zero: a source effect rather than a reversed response to generators (Section 6.2). Frozen video-score densities, pilot budget, for FaceForensics++ test originals (grey; Section 4) and DF40 face-local fakes (colour = arm, style = seed): (a) summed losses, (b) the consistency regularizer, (c) without its real-mass term; pooled AUC per seed in each panel. The regularizer family scores FaceForensics++ input near 1 whether it is real or fake, so the reals sit at one end of the axis and the DF40 fakes at the other.

Table 34: DF40 zero-shot detection AUC: each paradigm's fake videos against 140 FaceForensics++ test originals (a training-source real set, so absolute AUC partly measures source). Face-local pools swap, reenactment and editing; synthesis stands apart. Evid.: the same frozen score's AUC on FaceForensics++ validation; d: the frozen score mapping barely separates real from fake in distribution (Evid. ≤ 0.55), so the row is reported but not ranked. Coverage: Table 35.

Arm	Seed	Evid.	Flag	Swap	Reenact	Edit	Face-local	Synthesis
Pilot budget								
naive sum	1	0.5041	d	0.604	0.600	0.353	0.547	0.616
	2	0.5157	d	0.595	0.578	0.331	0.529	0.614
	3	0.5102	d	0.561	0.558	0.303	0.503	0.572
cross-task attention	1	0.5572		0.544	0.545	0.252	0.481	0.643
	2	0.5576		0.518	0.539	0.306	0.482	0.671
	3	0.5486	d	0.467	0.469	0.173	0.404	0.547
output-space consistency (w=0.1)	1	0.5003	d	0.005	0.001	0.011	0.004	0.018
	2	0.5017	d	0.009	0.002	0.023	0.008	0.037
	3	0.5022	d	0.004	0.001	0.009	0.003	0.010
Ten-fold budget (never ranked with the pilot block)								
naive sum	1	0.5121	d	0.534	0.623	0.295	0.525	0.577
	2	0.5120	d	0.548	0.594	0.222	0.499	0.529
	3	0.5071	d	0.673	0.712	0.262	0.602	0.612
cross-task attention	1	0.5163	d	0.694	0.571	0.311	0.551	0.560
	2	0.5532		0.457	0.399	0.219	0.377	0.421
	3	0.5710		0.447	0.456	0.197	0.397	0.547
output-space consistency (w=0.1)	1	0.4973	d	0.001	0.000	0.002	0.001	0.008
	2	0.5266	d	0.001	0.000	0.002	0.001	0.020
	3	0.5065	d	0.001	0.000	0.002	0.001	0.009

Table 35: DF40 generator coverage of the scoring pass behind Table 34. The release defines 40 generators; our evaluation pool holds 39 (WhichisReal was not obtained) and 37 were readable. Every fake is scored against the same 140 FaceForensics++ test originals. Editing videos hold one frame each, so their score is a single-frame score.

Paradigm	Release	Scored	Fake videos	Not scored (reason)
Face swap	10	9	10,528	mobileswap (nested archives, not readable)
Face reenactment	13	13	17,354	—
Face editing	5	4	7,806	e4e (unsupported archive compression)
Face synthesis	12	11	17,945	WhichisReal (not obtained)
Face-local pool			35,688	swap + reenactment + editing
All	40	37	53,633	

Table 36: TVIL zero-shot temporal transfer (508 test clips: 389 inpainted, 119 real; no tuning on TVIL): AP@0.5 and the designated secondary AR@10 per seed, internal metric, fused decode, at most 16 proposals per clip. Paired rows are same-seed AP@0.5 differences, mean ± sample sd over seeds with the number of negative seeds. Pilot and ten-fold blocks are never ranked together.

	AP@0.5				AR@10		
Arm	seed 1	seed 2	seed 3	paired mean ± sd	seed 1	seed 2	seed 3
Pilot budget							
naive sum	0.0490	0.0487	0.0513		0.3326	0.3213	0.3169
cross-task attention	0.0337	0.0281	0.0177		0.2539	0.2539	0.2225
output-space consistency (w=0.1)	0.0296	0.0323	0.0412		0.2831	0.2764	0.2764
paired: attention − naive	−0.0153	−0.0206	−0.0336	−0.0232 ± 0.0094 (3/3)			
paired: consistency − naive	−0.0194	−0.0164	−0.0102	−0.0153 ± 0.0047 (3/3)			
Ten-fold budget							
naive sum	0.0612	0.0521	0.0455		0.3551	0.3461	0.3191
cross-task attention	0.0719	0.0221	0.0189		0.3461	0.2382	0.2539
output-space consistency (w=0.1)	0.0344	0.0175	0.0247		0.3326	0.2989	0.2809
paired: attention − naive	+0.0106	−0.0300	−0.0266	−0.0153 ± 0.0225 (2/3)			
paired: consistency − naive	−0.0268	−0.0346	−0.0208	−0.0274 ± 0.0069 (3/3)			

Table 37: TVIL AP@0.5 per inpainting method, pilot budget, per seed: each method's inpainted clips scored together with all 119 real clips. Lower part: mean same-seed difference against naive sum, with the number of negative seeds. Ten-fold per-method values are not tabulated; at that budget (seed 1 only) the consistency arm trails naive sum on all four methods and cross-task attention is mixed in sign across methods.

	Fake	naive sum			cross-task attention			consistency (w=0.1)		
Method	clips	s1	s2	s3	s1	s2	s3	s1	s2	s3
FGT	92	0.0368	0.0295	0.0380	0.0219	0.0220	0.0135	0.0240	0.0188	0.0233
STTN	91	0.0526	0.0472	0.0590	0.0583	0.0416	0.0051	0.0207	0.0358	0.0486
FuseFormer	96	0.0354	0.0535	0.0443	0.0168	0.0202	0.0206	0.0433	0.0401	0.0434
E2FGVI	110	0.0627	0.0582	0.0703	0.0224	0.0232	0.0157	0.0319	0.0352	0.0413

Method	attention − naive (neg. seeds)	consistency − naive (neg. seeds)
FGT	−0.0156 (3/3)	−0.0127 (3/3)
STTN	−0.0179 (2/3)	−0.0179 (3/3)
FuseFormer	−0.0252 (3/3)	−0.0021 (2/3)
E2FGVI	−0.0433 (3/3)	−0.0276 (3/3)

Table 38: OpenForensics zero-shot mask transfer, Test-Dev (18,895 images; 33 without a fake face): whole-image union-mask IoU at 0.1 and 0.5, value agreement (fraction of pixels whose soft value lies within 0.01 of ground truth) and L1. Δ: same-seed IoU@0.5 difference against naive sum, mean $\pm$ sample sd over seeds. Ten-fold IoU@0.5 values sit below 0.05 and are read as orderings; blocks are never ranked together; ten-fold deltas from the 4-dp values. Inversion statistic 0.561 (pilot).

Arm	Seed	IoU@0.1	IoU@0.5	Value agr. @0.01	L1	Δ IoU@0.5
Pilot budget						
naive sum	1	0.3813	0.1151	0.0721	0.0648	
	2	0.3864	0.1947	0.0017	0.0731	
	3	0.3834	0.0918	0.0839	0.0659	
cross-task attention	1	0.3963	0.1155	0.0496	0.0628	$+0.0004$
	2	0.3765	0.2140	0.0008	0.0771	$+0.0193$
	3	0.3198	0.0604	0.0829	0.0660	-0.0314
	mean					-0.0039 ± 0.0256
output-space consistency (w=0.1)	1	0.1824	0.0202	0.8816	0.0589	-0.0949
	2	0.1805	0.0083	0.1823	0.0678	-0.1864
	3	0.1507	0.0119	0.4875	0.0648	-0.0798
	mean					-0.1204 ± 0.0577
Ten-fold budget						
naive sum	1	0.1522	0.0242	0.8973	0.0565	
	2	0.1754	0.0395	0.9064	0.0557	
	3	0.1463	0.0420	0.9156	0.0548	
cross-task attention	1	0.0940	0.0112	0.9213	0.0569	-0.0130
	2	0.1100	0.0170	0.9185	0.0563	-0.0225
	3	0.1295	0.0335	0.9224	0.0550	-0.0085
output-space consistency (w=0.1)	1	0.0015	0.0002	0.9419	0.0579	-0.0240
	2	0.0009	0.0001	0.9418	0.0579	-0.0394
	3	0.0041	0.0004	0.9412	0.0579	-0.0416

Table 39: OpenForensics zero-shot mask transfer, Test-Challenge (45,000 perturbed images; 73 without a fake face), same checkpoints and columns as Table 38. Every checkpoint scores slightly higher IoU@0.5 here than on Test-Dev; the arm orderings are unchanged except at pilot seed 1, where attention falls below naive; ten-fold deltas from the 4-dp values except seed 2, computed at full precision.

Arm	Seed	IoU@0.1	IoU@0.5	Value agr. @0.01	L1	Δ IoU@0.5
Pilot budget						
naive sum	1	0.3821	0.1272	0.0766	0.0649	
	2	0.3843	0.2059	0.0018	0.0732	
	3	0.3784	0.1036	0.0898	0.0660	
cross-task attention	1	0.3912	0.1223	0.0540	0.0632	-0.0050
	2	0.3665	0.2421	0.0010	0.0786	$+0.0362$
	3	0.3293	0.0721	0.0742	0.0665	-0.0315
	mean					-0.0001 ± 0.0341
output-space consistency (w=0.1)	1	0.2008	0.0248	0.8782	0.0593	-0.1025
	2	0.1959	0.0111	0.1842	0.0682	-0.1948
	3	0.1728	0.0161	0.4866	0.0650	-0.0875
	mean					-0.1282 ± 0.0581
Ten-fold budget						
naive sum	1	0.1644	0.0282	0.8982	0.0568	
	2	0.1731	0.0425	0.9080	0.0561	
	3	0.1628	0.0504	0.9133	0.0552	
cross-task attention	1	0.1133	0.0181	0.9198	0.0570	-0.0101
	2	0.1437	0.0274	0.9153	0.0564	-0.0150
	3	0.1553	0.0447	0.9202	0.0550	-0.0057
output-space consistency (w=0.1)	1	0.0020	0.0003	0.9411	0.0585	-0.0279
	2	0.0019	0.0003	0.9410	0.0585	-0.0422
	3	0.0054	0.0006	0.9403	0.0585	-0.0498

Table 40: Adjacent-frame mask stability on the evaluation clips (286 manipulated, 100 real; 11,966 adjacent pairs): flicker, the mean absolute change of the soft mask, and stability@0.5, the IoU of consecutive masks binarised at 0.5 (two empty masks count as 1). Adjacency is in the sampled-clip sequence and without motion compensation these cannot separate model instability from content change. The ten-fold regularizer's near-zero flicker and high stability follow from its near-empty masks (validation IoU@0.5 of 0.0032–0.0182, Table 9): consecutive empty masks score 1, so those cells measure mask extinction, not coherent detected regions.

	Flicker			Stability@0.5		
Arm	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
Pilot budget						
naive sum	0.0355	0.0342	0.0343	0.5740	0.6707	0.5572
cross-task attention	0.0310	0.0305	0.0289	0.7623	0.8573	0.5958
output-space consistency (w=0.1)	0.0106	0.0080	0.0171	0.5641	0.7812	0.5442
Ten-fold budget (never ranked with the pilot block)						
naive sum	0.0309	0.0331	0.0348	0.4936	0.4607	0.4910
cross-task attention	0.0290	0.0275	0.0294	0.5307	0.6173	0.5674
output-space consistency (w=0.1)	0.0007	0.0008	0.0007	0.8742	0.8683	0.8390

A.5 PROBES

Supports Section 6. One paragraph per exploratory probe: question, population, finding, status. All probes are exploratory analyses of already-trained checkpoints on the frozen weak-GT evaluation set (286 manipulated + 100 real clips unless stated); none feeds a selection or a stopping rule.

(i) Oracle-swap decomposition. Which axis limits the joint metric? Swapping ground-truth masks under each arm's own predicted segments lifts joint tube-mAP@0.2 to arm means of 0.50 (naive), 0.75 (attention), 0.51 (consistency), so the attention arm's temporal recall is real joint-task headroom and mask quality is the binding constraint. The reverse swap (ground-truth segments, predicted masks) collapses every arm at strict tube IoU (≤ 0.0002 at @0.5), while the consistency arm leads the loose @0.2 column at 2 of 3 seeds.

(ii) Mass-matched shrink control. Is the consistency arm's joint-metric lead placement skill or an area prior? Transferring the consistency arm's per-frame mask-area budget onto naive's own soft-mass ranking (no retraining) lifts shrunk-naive to at least the same-seed consistency value at 3 of 3 seeds ($+0.003$ / $+0.018$ / $+0.150$ tube-mAP@0.2) — the pre-registered "area prior" outcome, with no residual placement advantage on this set. The same budget lifts the attention arm even harder at 2 of 3 seeds (to 0.18 and 0.30 on this set).

(iii) Binarisation-threshold sweep. The consistency-over-naive lead at the frozen 0.5 threshold is not a calibration artifact of that threshold at seeds 1–2: it persists at every threshold up to 0.6 and under each arm's own best threshold; at seed 3 naive leads from 0.5 upward (consistency leads at 0.1–0.4). The arms occupy opposite calibration regimes — consistency curves peak at low thresholds (small, value-sharp masks) while naive/attention curves rise toward 0.7–0.9 (diffuse mass that becomes accurate only when cut to its core) — and every curve is smooth, so the frozen rankings are not knife-edge accidents.

(iv) Per-clip correlation. The per-clip coupling between mask-area change and joint-metric change is weak everywhere ($|r| \leq 0.26$ on $n = 286$ paired clips, every seed); what signal exists points in the area-prior direction (smaller relative area, slightly larger tube-IoU gain; negative Pearson at 3 of 3 seeds). The regularizer's joint-metric advantage is a set-level calibration effect, not a clip-by-clip proportional one.

(v) Ranking concordance. Over 16 trained checkpoints, the in-domain validation-AP ranking is *negatively* associated with the joint-metric ranking (Kendall $\tau = -0.29$) and shows little association with in-segment mask quality ($\tau = +0.02$); three of the four readout families would crown a different winner, and the joint metric's winner sits 8th of 16 on validation AP. (One concordance cell was corrected when an independent re-derivation superseded a hand count; the corrected values are quoted.)

(vi) Separate-benchmark emulation. Reading our own checkpoints under an OmniFD-style protocol — each head scored on its own benchmark, never jointly — crowns the attention arm on both axes and ranks the five consistency-carrying checkpoints *last* on spatial IoU, while the same-clip joint readout crowns a consistency checkpoint on both joint columns; no checkpoint wins under both protocols. This is the concrete exhibit that separate-benchmark practice is structurally unable to see the dissociation the paper measures.

(vii) Combinability. A single-seed run combining both mechanisms (attention + consistency at half weight) reaches internal validation AP 0.714 — above the attention arm's seed band — while the consistency arm's characteristic late-epoch mask sharpening still fires on the same schedule, with a shallower version of its extent cost. The two mechanisms co-exist in one model without cancelling; whether the combination helps at scale was not tested and is not claimed. Its three-seed confirmation at the same weight (Table 7) stays within ± 0.02 of attention alone at every seed, mixed in sign, and its joint-readout margin is strongly seed-dependent.

(viii) Pseudo-ground-truth validation. The frozen set itself cannot be validated against manipulated-pixel residuals: only 11 of its 286 clips have a released, length-matched real counter-

part (a structural property of the dataset release, not a processing failure), below the pre-set $n \geq 100$ bar. On a differently-drawn 116-clip sample restricted to pairable clips and built with the byte-identical mask recipe, the pseudo-masks recover a median 51% of the manipulated-pixel residual at the primary residual threshold (55% under a small spatial tolerance), while a median 45% of the residual mass escapes even the tolerant mask. This supports retaining the "weak spatial reference, with potentially biased rankings" label rather than replacing it: the masks find the manipulation region at rough placement, and are far from pixel-accurate.

(ix) Cross-task attention. In the attention arm the spatial-to-temporal direction is concentrated and places more mass on the mouth region than uniform attention would, at all three seeds (head means 9.17 / 6.52 / 7.44), while the temporal-to-spatial direction is near-uniform (normalized entropy 0.999) with no stable preference for manipulated frames (lift 0.968 / 1.008 / 1.039). A single run on a 25%-stitched mix concentrates that direction (lift 4.11). The pattern is consistent with the spatial-to-temporal readout over each frame's projected patch tokens informing temporal decisions; attention weights are not causal importance (Figure 6).

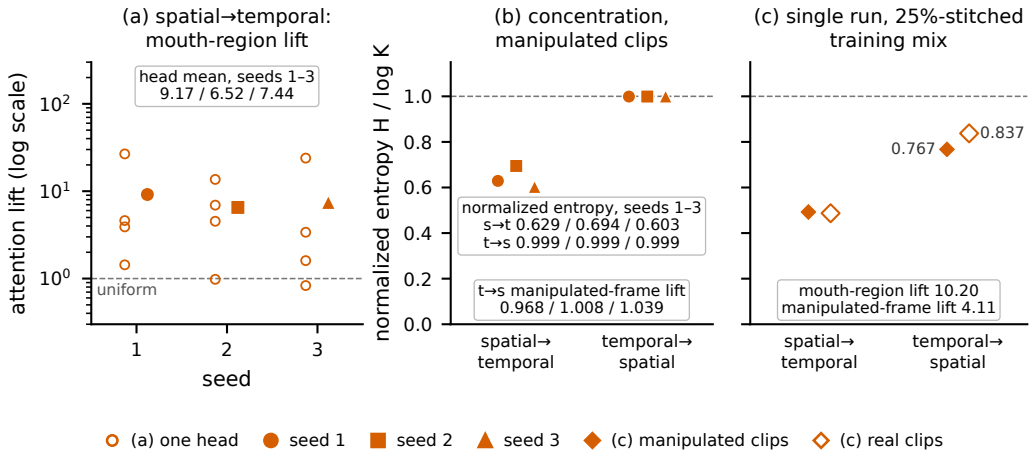


Figure 6: Cross-task attention statistics (pilot budget, three seeds; dashed lines mark uniform attention). (a) Spatial-to-temporal mass on the mouth region, a landmark-derived weak proxy, per head and head mean; one head dominates the mean at each seed (the same head at seeds 1 and 2). (b) Normalized entropy by direction, manipulated clips. (c) One run on a 25%-stitched training mix, an observation only. Attention weights are not causal feature importance.

A.5.1 Mask signature of the output-space regularizer

Supports Section 6.2. Companions to the value-agreement contrast of Section 6.2. Pilot budget, three seeds, at the selected epoch (Table 4; Figure 7): the regularizer's value-agreement fractions are 0.92 / 0.29 / 0.57 against the baseline's 0.15 / 0.007 / 0.17 (the fraction of pixels whose predicted soft value falls within 0.01 of ground truth; seeds 2 and 3 reach 0.91 and 0.92 only by the end of the run, after their selected epochs), mask IoU@0.5 falls to 0.21–0.26 against the baseline's 0.40–0.44, and mask L1 is lower than the same-seed baseline at every seed (Section 6.2 keeps the seed-1 value-agreement pair). At the ten-fold budget (a separate block, never ranked with the pilot rows) the split is starker still: the regularized arm's mask IoU@0.5 at its selected epoch is 0.003–0.018 across seeds, against 0.28–0.38 for the baseline (Table 9).

A.5.2 Same-clip supervision couples slowly

Supports Section 6.2. The stitched construction puts both label types on one clip and so removes the cross-dataset problem by fiat; it is a reference point, not a method. Its informative property is how slowly it trains: at pilot scale, with a quarter of the training batches stitched, the naive and regularized arms reach 0.2938 and 0.3260 internal AP@0.5 (Table 6). Over ten epochs at matched

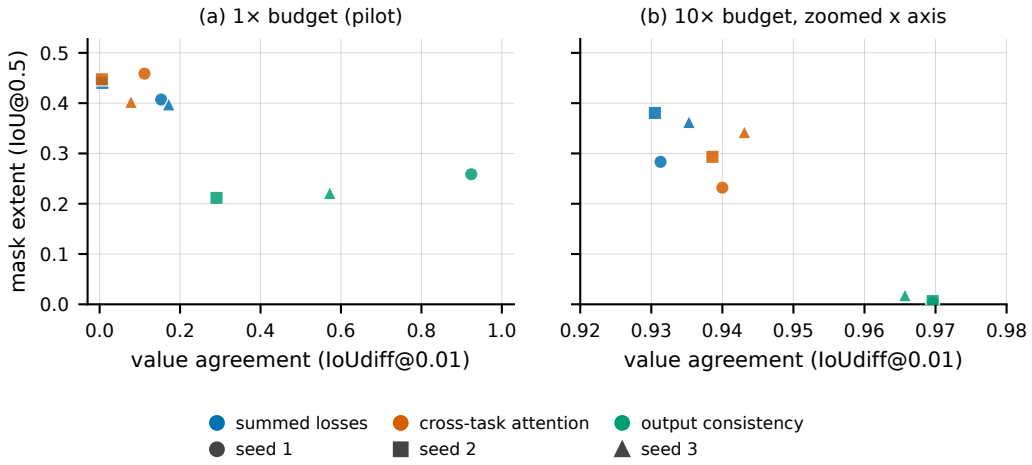


Figure 7: Mask value agreement against extent. Each point is one model at its selected epoch: IoUdiff@0.01 (fraction of pixels whose predicted value lies within 0.01 of ground truth) against mask IoU@0.5; colours mark arms, markers seeds. (a) Pilot budget. (b) Ten-fold budget, a separate block with a zoomed horizontal axis: every arm agrees in value there, and only extent separates the output-space regularizer. Coincident points overlap.

budget the stitched arm's temporal training loss falls 7.6% against 41% for the mixed cross-dataset arm, and its scores are still rising when the budget ends (Table 42; Figure 8). Scored zero-shot on the joint evaluation set, the stitched-only model matches 55 manipulated lip-sync segments at temporal IoU 0.5, against 268–301 for each pilot model trained on both datasets, and reaches a tube-mAP@0.2 of 1.3×10^{-5} (Table 41); with weak masks and a single run we read this only as an ordering: same-clip supervision on constructed FaceForensics++ clips did not carry over to real lip-sync edits. Two measurements bound what the construction can teach: a frame-level probe separates stitched spliced frames from real frames at AUC 0.591, against 0.754 for TalkLip fakes, and real AV-Deepfake1M segments have a median length of 0.30 s, about one sampled frame at 32 frames per clip, far shorter than the stitched segments. On the frozen stitched evaluation set the dose curve rises from 0.017 (baseline, no same-clip supervision) to 0.061 at a quarter dose and 0.062 at full dose. We report the 25%-to-100% flatness as an observation rather than a saturation claim: the two upper points come from runs with different training budgets, and the evaluation set carries about forty-three ground-truth segments, which is too few to separate them.

A.5.3 A MECHANISM HINT FROM FACIAL KINEMATICS

Supports Section 6.3; the four-family detail is in Appendix A.4. On a landmark-tracked sample of the manipulated clips, the manipulations are kinematically *over-smoothed* rather than irregular: within the same clip, the high-percentile lip-jerk of a manipulated segment is below that of the real footage around it in 70% of clips (median difference -715.5 at the primary smoothing window, the direction surviving the second at -97.6; Figure 9), while typical frame-to-frame jerk is mildly higher. Per-clip localization quality is weakly negatively associated with retained extreme motion in all nine seed runs (Spearman -0.13 to -0.29, same sign throughout, uniformly weaker under Pearson): clips whose manipulated segments keep more natural motion extremes are localized worse, consistent with the detectors already using part of this smoothness evidence, so the cue reads as partly redundant with what the models exploit rather than complementary to it. The signature is not specific to the training-time generator: on a second, architecturally unrelated lip-sync generator (the diffusion-based Diff2Lip subset of the extended AV-Deepfake1M++ release (Cai et al., 2025), rather than the GAN-based generator of training) the same within-clip paired contrast on 179 fake/real pairs is negative at both smoothing windows (45–47% of pairs positive) — a single-generator replication on sign and direction only, since the two samples come from different corpora and resolutions, and a narrower one, carried by lip jerk alone while the other three landmark features show no comparable truncation. Across four further manipulation families the same comparison against each clip's own

Table 41: Joint where-and-when readout on the frozen evaluation set (286 AV-Deepfake1M lip-sync clips with weak landmark-derived masks, 100 real); exploratory rankings that may depend on the weak reference. Top: pilot models trained on both datasets. Bottom: the stitched-only model, trained with both labels on constructed FaceForensics++ clips and evaluated zero-shot. TP: segments matched at temporal IoU 0.5; mask IoU is averaged within them. Flicker and stability as in Table 40. Tube-mAP at 0.75 is 0.0000 for every model and tube-mAP at 0.5 at most 0.0002: tube IoU is temporal IoU times mean per-frame mask IoU, and with in-TP mask IoU below 0.2 against the weak references a tube IoU of 0.5 is out of reach, so those columns say nothing about the models.

		tube-mAP			Mask IoU			Stability
Model	Seed	@0.2	@0.5	@0.75	in TP	TP	Flicker	@0.5
Trained on both datasets (pilot budget)								
naive sum	1	0.0319	0.0000	0.0000	0.1683	275	0.0355	0.5740
	2	0.0085	0.0000	0.0000	0.1506	268	0.0342	0.6707
	3	0.0366	0.0000	0.0000	0.1698	271	0.0343	0.5572
cross-task attention	1	0.0081	0.0000	0.0000	0.1458	296	0.0310	0.7623
	2	0.0030	0.0000	0.0000	0.1381	301	0.0305	0.8573
	3	0.0509	0.0001	0.0000	0.1748	299	0.0289	0.5958
output-space consistency (w=0.1)	1	0.0989	0.0002	0.0000	0.1946	280	0.0106	0.5641
	2	0.0617	0.0000	0.0000	0.1545	276	0.0080	0.7812
	3	0.0088	0.0000	0.0000	0.1063	274	0.0171	0.5442
Trained on stitched FaceForensics++ clips only (zero-shot on this set)								
stitched-only, 30 epochs	–	1.3×10^{-5}	0.0000	0.0000	0.1185	55	0.0454	0.4906

Table 42: Same-clip dose curve on the frozen stitched FaceForensics++ validation set (560 clips, about 43 ground-truth segments; boundary-masked AP@0.5 primary). The 0% rows are zero-shot on this set; means are $\pm$ sample sd over seeds. The two upper points come from runs with different training budgets (20 and 30 epochs against 10 for the 0% rows, the 20-epoch runs with a fresh optimizer at epoch 10), so the plateau between them is not a dose effect isolated from compute. Bottom: relative drop of the training temporal loss, a coupling-speed indicator (the two losses are computed on different data).

Same-clip share	Model	Seed	AP@0.5 (b.-m.)	AP@0.5	AR@10 (b.-m.)
0% (zero-shot)	naive sum	1	0.01597	0.01323	0.4935
		2	0.01666	0.01463	0.5090
		3	0.01710	0.01420	0.5039
		mean	0.01658 ± 0.00057	0.01402 ± 0.00072	
	consistency (w=0.1)	1	0.02249	0.01992	0.6279
		2	0.01620	0.01631	0.5349
		3	0.02416	0.02162	0.6279
		mean	0.02095 ± 0.00420	0.01928 ± 0.00271	
25%	naive sum, 20 epochs	–	0.06103	0.06010	0.7313
	consistency, 20 epochs	–	0.06142	0.05979	0.7623
100%	stitched-only, 30 epochs	–	0.06215	0.06226	0.7313

Run	Epochs	Training temporal loss	Change
stitched-only	10	$1.7362 \to 1.6048$	-7.6%
stitched-only	30	$1.7362 \to 1.4048$	-19.1%
naive sum, mixed cross-dataset	10	$0.807 \to 0.478$	-41%

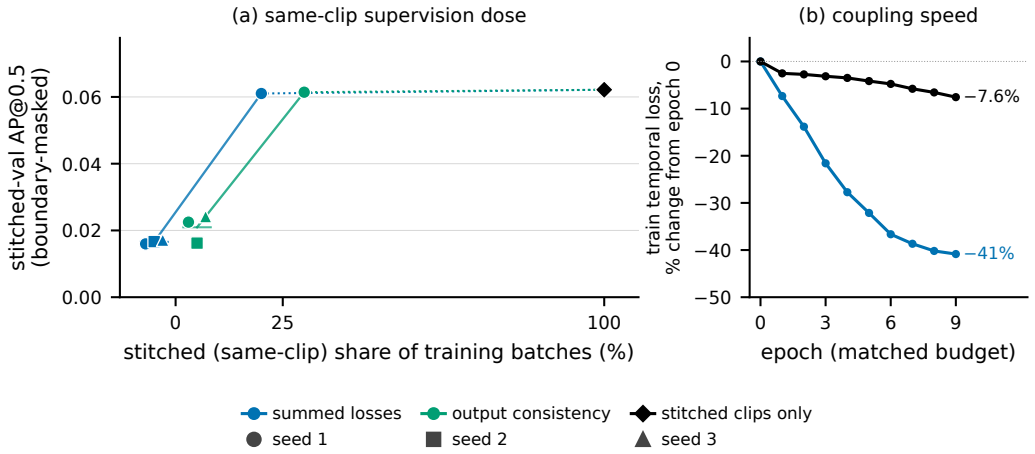


Figure 8: Same-clip supervision. (a) Boundary-masked AP@0.5 on the constructed stitched validation set against the stitched share of training batches (Table 42); 0% models never saw stitched clips. The 25% and 100% points come from runs with different training budgets (20 epochs with a warm restart; 30 epochs), and the set has about 43 ground-truth segments. (b) Temporal training-loss change over ten matched epochs; the two runs train on different data.

source video — a between-video design, weaker than the within-clip one — is negative in three of four, and in those three on all four landmark features (Appendix A.4); the exception is FaceSwap, one of the two graphics-pipeline families (a hypothesis-generating observation). The correlations are modest and correlational; we draw no causal conclusion.

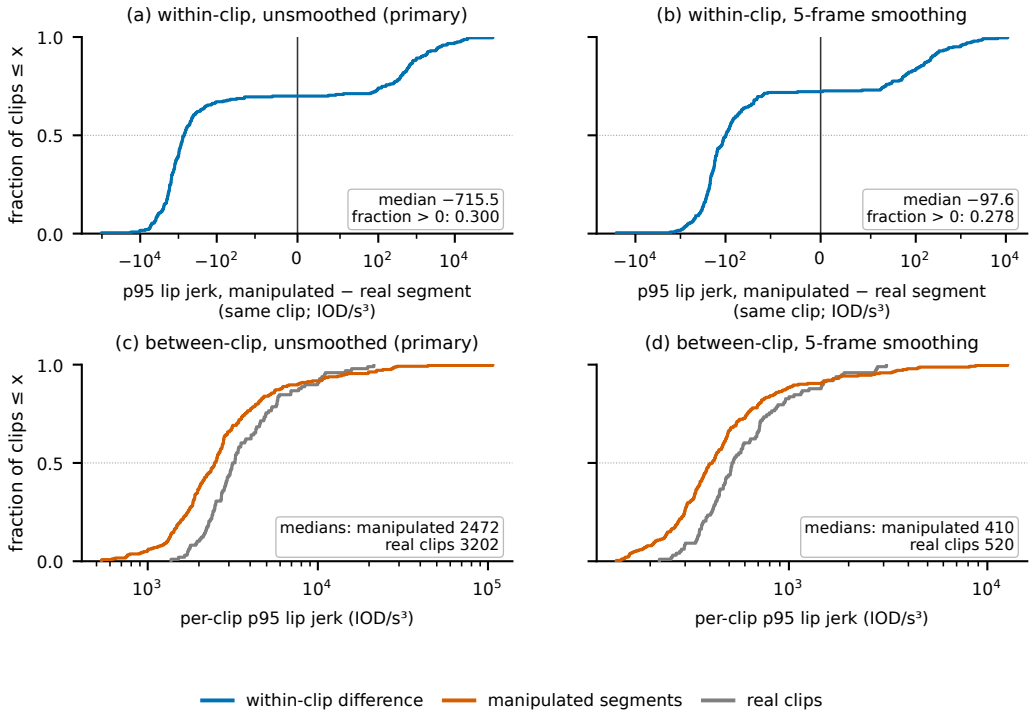


Figure 9: Facial-kinematics descriptives on the landmark-tracked sample of manipulated evaluation clips: cumulative distributions of per-clip 95th-percentile lip jerk (tracker-observed MediaPipe landmarks; IOD = inter-ocular distance), unsmoothed (left) and with 5-frame smoothing (right). (a), (b): within-clip difference, manipulated minus real segment of the same clip; (c), (d): manipulated segments (orange) against real clips (grey). High-percentile lip jerk is suppressed in manipulated segments at both windows. Exploratory and descriptive: tracker-level kinematics, never physiology, with no causal reading; like any frame-differencing feature it is exposed to motion confounds, which the same-clip real segments and the real clips anchor.

A.6 TERM ABLATIONS OF THE CONSISTENCY OBJECTIVE

The objective as implemented (added after external feedback). Every arm minimizes, on each batch, the loss of that batch's own source (`stdt/losses/`; weights from the released configurations). On a FaceForensics++ batch with soft target masks y and predicted masks $\hat{m}$ (pixel means throughout),

$$L_{\mathrm{sp}} \;=\; -\,\mathrm{mean}\big[\,5\,y\log\hat{m} + (1-y)\log(1-\hat{m})\,\big] \;+\; \mathrm{mean}\,(\hat{m}-y)^2.$$

On an AV-Deepfake1M batch of T=32-frame clips with frame labels $f_t \in \{0,1\}$, the BMN-style head emits frame probabilities p_t, start and end probabilities s_t, e_t and a boundary-matching map $\hat{q}_{ij}$ over candidate segments (i,j), and

$$\begin{aligned} L_{\mathrm{tp}} \;=\;& \mathrm{BCE}(p,f) + \mathrm{BCE}(s,s^*) + \mathrm{BCE}(e,e^*) + \mathrm{BCE}\big(\hat{q}, \mathbf{1}[q>0.6]\big) \\ & + 10\,\frac{\sum_{ij}\mathbf{1}[q_{ij}>0]\,(\hat{q}_{ij}-q_{ij})^2}{\sum_{ij}\mathbf{1}[q_{ij}>0]}, \end{aligned}$$

where q_{ij} is the temporal IoU of candidate (i,j) with its best ground-truth segment, s^*, e^* are boundary indicators equal to 1 at a ground-truth boundary frame and 0.5 one frame away, and BCE is the mean binary cross-entropy. Naive sum and cross-task attention use exactly these two terms; the arms differ only in the network. The output-space consistency arm adds $w\,L_{\mathrm{cons}}$ with w=0.1. On FaceForensics++ batches,

$$L^{\mathrm{FF}}_{\mathrm{cons}} \;=\; \mathrm{BCE}\big(\bar{p},\; \mathbf{1}[\pi(\hat{m})>\tau]\big), \qquad \tau{=}0.3,$$

with $\pi(\cdot)$ the mean of the top 2% of a mask's pixel values (peak mass) and $\bar{p}$ the temporal head's fake probability for the image entered as a static 8-frame clip (its frame probabilities averaged); the target is detached, so this term trains the temporal head only. On AV-Deepfake1M batches, with $r_t = 1 - f_t$ the ground-truth real indicator and $\pi_t = \pi(\hat{m}_t)$ the peak mass of the frame-t mask,

$$L^{\mathrm{AV}}_{\mathrm{cons}} \;=\; L_{\mathrm{smooth}} + L_{\mathrm{mass}} + L_{\mathrm{rank}}, \qquad L_{\mathrm{smooth}} \;=\; \frac{\sum_{b,t} r_t r_{t-1}\,\mathrm{mean}\,|\hat{m}_t - \hat{m}_{t-1}|}{\sum_{b,t} r_t r_{t-1}},$$

$$L_{\mathrm{mass}} \;=\; \frac{\sum_{b,t} r_t\,\pi_t}{\sum_{b,t} r_t}, \qquad L_{\mathrm{rank}} \;=\; \underset{b\,:\,\text{both classes}}{\mathrm{mean}}\; \max\big(0,\; 0.2 + \bar{\pi}^{\mathrm{real}}_b - \bar{\pi}^{\mathrm{fake}}_b\big),$$

where $\bar{\pi}^{\mathrm{real}}_b$ and $\bar{\pi}^{\mathrm{fake}}_b$ are clip b's mean peak mass over its real and fake frames, the ranking term is averaged over clips that contain both classes, and the three sub-terms (smoothness, real mass, ranking) carry weight 1; the frame labels are detached, so this term trains the spatial head only. The term ablations below set one sub-term's weight to zero.

Supports Section 6.2. The real-mass ablation (Appendix A.2) met its pre-registered refutation criterion: the transfer sign inversion persists at every seed with the real-mass term removed (pooled face-local AUC 0.007/0.002/0.003 at seeds 1–3), so that term alone does not cause it. Follow-up single-term ablations of the two remaining terms were run under the same pre-registered rules with the same outcome: with the ranking term removed the out-of-domain detection score is 0.016/0.003/0.005 at seeds 1–3, and with the smoothness term removed 0.002/0.017/0.001 at seeds 1–3, against parent values of 0.004/0.008/0.003 — the inversion is essentially unmoved in every case (every ablated value below 0.02, against an intact objective at 0.003–0.008). None of these mask-side term removals eliminates it; this does not test every term of the objective. Removing the ranking term does have a pre-registered side effect: without its fake-over-real constraint the mask head drifts to near-empty masks (in-TP mask overlap of order 10^{-4} or below at all three seeds), an outcome the ablation's caveat predicted before any run started. The follow-up removing ranking and smoothness together reached the same outcome (0.015/0.005/0.002 at seeds 1–3), generalizing the negative to the pair jointly; the mask-drift side effect reappears there too, at two of three seeds.

What the coupling channel taught (post hoc, not pre-registered). None of those ablations touches the regularizer's only path to the temporal head: the FaceForensics++ term, which trains frame-level fakeness toward the predicted mask's own verdict (peak mask mass above τ); the real-mass, ranking and smoothness terms act on the mask side. A probe added 2026-09-12 ran every regularizer-family checkpoint, the pseudo-label arm and the summed-loss and attention arms over

FaceForensics++ training and validation images (four evenly spaced frames per video, the frame rule of the cross-generator stage), recording that verdict beside the temporal head's own output (Table 43). Three readings, each stated over the 25 frozen regularizer checkpoints whose teacher is their own predicted mask (the three unfrozen-backbone checkpoints and the stitched-supervision arm, whose teacher is the ground-truth mask on stitched batches, are the two stated exceptions and are reported separately below). (i) Every one of the 25 answers near 1 on FaceForensics++ input whether real or fake (median 0.92–1.00 on both classes, frame-level AUC 0.504–0.530), while the summed-loss and attention arms, which have no such channel, answer 0.001–0.003 on the same images. (ii) The teacher barely separates the classes it is built from: the fraction of images it calls manipulated differs between real and fake frames by at most 0.25 across those 25, and it agrees with the same statistic computed from the *ground-truth* mask on only 23–79% of images. (iii) Both arms whose teacher is instead the *ground-truth* mask fail to saturate: the stitched-supervision arm sits at median 0.34 real / 0.36 fake (AUC 0.595, teacher gap 0.34), and the pseudo-label arm does not saturate either (median 0.71 real / 0.72 fake, AUC 0.604) and has a small temporal difference — -0.006 at the exploratory seed and $+0.0035 \pm 0.0076$ internal AP over three matched seeds trained post hoc on the same budget, every seed inside the ± 0.02 band — no consistent benefit under this pseudo-label recipe. This is not a clean teacher-quality comparison: loss form and weight also differ. Scored zero-shot on the generator probe that arm also inverts (pooled face-local AUC 0.010 for the exploratory checkpoint and 0.0057 / 0.0012 / 0.0033 for the three matched seeds), and, like the stitched-trained and unfrozen regularizer rows, it *clears* the in-distribution evidence check flagging every other regularizer row (0.580, and 0.577 / 0.564 / 0.574 for the matched seeds, against the pre-registered 0.55 bar), so the inversion does not require a near-chance in-distribution mapping. As a check on the probe itself, each checkpoint's video-level validation score reproduces the in-distribution evidence AUC computed independently for the cross-generator stage.

The direction of this negative has a single-dataset analogue: Oh et al. (2026) argue that partially-labelled multi-task methods relying on predictions from the unlabelled task establish unreliable task associations, and replace that coupling with feature-level retrieval. Our result is the cross-dataset instance — two separately collected sources, a pixel task and a temporal task — and suggests a source-associated response. Probes of trained checkpoints cannot determine the training dynamics responsible for it.

Table 43: Teacher-target probe: what the FaceForensics++ coupling channel taught the temporal head (post hoc, not pre-registered; Appendix A.8). *Teacher* = fraction of FaceForensics++ training images the arm's own teacher calls manipulated, on real / fake images; *head* = median frame-level fakeness the temporal head outputs on them; *AUC* = the head's frame-level AUC over those images; *val* = the arm's video-level validation AUC, which reproduces the evidence AUC of Table 34. The pseudo-label row's teacher is the ground-truth mask (0.000 / 0.978); every other row's is its own predicted mask. Unfrozen rows are never ranked with frozen ones.

Arm (pilot budget unless noted)	Teacher real	Teacher fake	Head real	Head fake	AUC / val
consistency w=0.1, seed 1	0.730	0.871	1.0000	1.0000	0.510 / 0.500
consistency w=0.1, seed 2	0.699	0.830	0.9994	0.9995	0.514 / 0.502
consistency w=0.1, seed 3	0.632	0.806	0.9998	0.9998	0.510 / 0.502
consistency, 10×, seed 1	0.002	0.019	0.9921	0.9922	0.505 / 0.497
consistency, 10×, seed 2	0.003	0.047	0.9350	0.9399	0.521 / 0.527
consistency, 10×, seed 3	0.036	0.104	0.9828	0.9833	0.507 / 0.507
no real-mass, seed 1	0.898	0.946	0.9992	0.9993	0.505 / 0.495
no ranking, seed 1	0.166	0.400	0.9959	0.9961	0.510 / 0.500
no smoothness, seed 1	0.650	0.814	1.0000	1.0000	0.509 / 0.498
no ranking + smoothness, seed 1	0.510	0.734	0.9900	0.9904	0.508 / 0.503
soft coupling	0.714	0.859	0.9818	0.9831	0.530 / 0.517
consistency, DINOv2, seed 1	0.283	0.507	1.0000	1.0000	0.510 / 0.505
attention + consistency, seed 1	0.097	0.246	0.9995	0.9996	0.519 / 0.509
naive sum, seed 1	0.960	0.974	0.0022	0.0023	0.520 / 0.504
attention, seed 1	0.989	0.990	0.0008	0.0010	0.550 / 0.557
pseudo-label (GT-mask teacher)	0.000	0.978	0.7118	0.7232	0.604 / 0.580
consistency, unfrozen, seed 1	0.925	0.994	0.9995	0.9999	0.734 / 0.636

A.7 SELECTIVE TABULATION IN THE RECENT LITERATURE

Supports Section 2, Table 1, the post-hoc OmniFD mask-loss contrast and the audit of what the cited multi-objective systems ablate (Table 47). The main text touches the vision-only "capability boundary" foil in one clause; the full record is here, stated as observations about published tables. On the AV-Deepfake1M benchmark, the dataset paper (Cai et al., 2024) reports vision-only temporal baselines of 36.08 (ActionFormer + InternVideo) and 20.24 (ActionFormer + VideoMAEv2) AP@0.5 on its test set ("fullset"). DeformTrace's Table 2 (Zhu et al., 2026a) additionally tabulates two far stronger vision-only rows — TE-TAD (Kim et al., 2024) at 71.4 and TadTR (Liu et al., 2022) at 60.5 — carrying venue citation markers only, with neither their provenance nor the visual feature backbone stated in the paper (neither the TE-TAD nor the TadTR paper reports an AV-Deepfake1M result). MG-RWKV (Ni et al., 2026) tabulates only the 20.24 row among vision-only baselines, describes it as the strongest single-modality baseline, and infers a "fundamental capability boundary between visual-only and audio-visual approaches"; UniSkip-Mamba (Yu et al., 2026) likewise tabulates no vision-only row above 20.24 (its other one, MesoInception4, is at 8.50) and reports no vision-only ablation of its own model. These statements are about what the tables contain, verified against the papers' text. Our own measurement — the same ActionFormer head (Zhang et al., 2022) scoring 64.6 AP@0.5 with InternVideo2-1B features (Wang et al., 2024) on a 10% training subset — sits on DeformTrace's side of the discrepancy and indicates the apparent boundary is largely a feature-backbone artifact. This row is a side result, printed as a context row of Table 1, and is never ranked against the joint arms. We do not report the vision-only BA-TFD (Cai et al., 2022) we trained through its reference implementation: that code caches each clip's boundary-map training target under the clip's file name, and AV-Deepfake1M's clip paths end in only four file names, so every training clip received one of four targets (the audio-visual checkpoint of the Appendix A.4 probe was trained the same way). Selective tabulation of vision-only baselines in the literature is the pattern these observations record. The AV-Deepfake1M validation split every row of ours is scored on (the upstream test labels are withheld) partitions manipulation variants within clip directories rather than speakers: all 1,656 of its speakers and all 22,531 of its source recordings also occur in the training split, and 40,008 of its 57,340 records name a training-split real clip as their source; no clip file repeats across the two splits. The published validation-split bars (WAFL; LipForensics on the ++ validation split) share this protocol, so it bounds what any validation-split number says about unseen speakers; the dataset paper's and DeformTrace's bars are test-set bars on 411 held-out subjects (Cai et al., 2024); only the cross-dataset probes are free of it. A speaker-exposure readout on this split (post hoc, item (xxiii); Table 44) finds AP@0.5 lowest on the most-exposed quartile of speakers and highest on the least-exposed one, at every seed and arm, with attention minus naive positive at every seed in every quartile: the in-domain number does not rise with the training footage of a speaker. Exposure is confounded with per-speaker clip count and recording conditions and the quartiles differ in size, so this is an ordering across slices, not an attribution.

On the extended AV-Deepfake1M++ release (Cai et al., 2025), a published LipForensics system (Klein et al., 2025) reaches 97.86% AP@0.5 on that release's validation split, scored against visual-manipulation labels, with 2,606 subjects shared between its training and validation sets; no vision-only BA-TFD (Cai et al., 2022) or UMMAFormer (Zhang et al., 2023) ablation is published on either release — the two facts are quoted only together.

Zero-shot scoring on AV-Deepfake1M++ validation (post hoc). The nine ten-fold checkpoints of Table 1, unchanged, were scored zero-shot on the AV-Deepfake1M++ validation split (77,326 records; nothing trained or tuned on that release; decode carried from each row's configuration; official evaluation code; Table 45; post hoc, Appendix A.8, item (xxv)), under a reading rule fixed on 2026-09-22 before any export existed. On the full split the arms score 0.0831 / 0.0924 / 0.0890 (naive), 0.1166 / 0.1175 / 0.1120 (attention) and 0.1152 / 0.1221 / 0.1330 (attention without mask labels): attention minus naive is $+0.0335$ / $+0.0251$ / $+0.0230$ and the label-free arm minus naive $+0.0321$ / $+0.0297$ / $+0.0440$, each positive at every seed, with the same signs on the visual-manipulation subset and on AR@10 ($+0.014$ to $+0.019$); attention minus the label-free arm is $+0.0014$ / -0.0046 / -0.0210 (one of three positive). These are levels under that release's protocol, in their own block; the in-domain ordering of Table 1, both attention arms above naive, holds at every seed on this split, and no interval is computed. The split holds one diffusion generator, Diff2Lip, at 199 fakes, all from the release's third source corpus (labelled `silent_videos`; the dataset paper names its sources as VoxCeleb2, LRS3 and EngageNet), which is absent from AV-Deepfake1M.

Table 44: Speaker-exposure slices of the official evaluation set (post hoc, item (xxiii)). Exposure = number of training-split records of an evaluation speaker; the 1,656 speakers are cut into equal-count quartiles Q1 (least exposed, 29–151 records) to Q4 (521–1,513). AP@0.5 with the official evaluation code on the unchanged ten-fold exports of Table 1; mean ± sample sd over seeds 1–3; differences per seed with the count of positive values. The quartiles differ in size (full split 3,098 / 7,004 / 12,817 / 34,421 records; visual subset 2,334 / 5,273 / 9,602 / 25,845) and exposure is confounded with per-speaker clip count, recording length and conditions, so the rows are an ordering across slices, never an attribution; no slice is ranked against the full-split rows.

Arm	Q1 (least exposed)	Q2	Q3	Q4 (most exposed)
AP@0.5, full split (mean ± sd over seeds)				
Naive sum	0.1883 ± 0.0104	0.1604 ± 0.0086	0.1529 ± 0.0074	0.1389 ± 0.0072
Cross-task attention	0.2276 ± 0.0025	0.1897 ± 0.0012	0.1835 ± 0.0013	0.1602 ± 0.0012
Attention, no mask labels	0.2409 ± 0.0116	0.1996 ± 0.0105	0.1918 ± 0.0102	0.1700 ± 0.0102
AP@0.5, visual subset				
Naive sum	0.2805 ± 0.0151	0.2392 ± 0.0127	0.2285 ± 0.0109	0.2080 ± 0.0106
Cross-task attention	0.3383 ± 0.0038	0.2826 ± 0.0017	0.2744 ± 0.0021	0.2397 ± 0.0017
Attention, no mask labels	0.3583 ± 0.0172	0.2967 ± 0.0151	0.2862 ± 0.0152	0.2540 ± 0.0151

Q4 − Q1 per seed (1 / 2 / 3; count of positive values)		
Arm	full split	visual subset
Naive sum	-0.0457 / -0.0476 / -0.0548 (0/3)	-0.0673 / -0.0695 / -0.0808 (0/3)
Cross-task attention	-0.0697 / -0.0654 / -0.0672 (0/3)	-0.1019 / -0.0958 / -0.0982 (0/3)
Attention, no mask labels	-0.0684 / -0.0735 / -0.0708 (0/3)	-0.1004 / -0.1081 / -0.1043 (0/3)
Paired same-seed differences per quartile, full split (seed 1 / 2 / 3; count of positive values)		
Quartile	attention − naive	no mask labels − naive
Q1	$+0.0529$ / $+0.0322$ / $+0.0327$ (3/3)	$+0.0519$ / $+0.0505$ / $+0.0552$ (3/3)
Q2	$+0.0391$ / $+0.0220$ / $+0.0268$ (3/3)	$+0.0392$ / $+0.0319$ / $+0.0465$ (3/3)
Q3	$+0.0383$ / $+0.0248$ / $+0.0288$ (3/3)	$+0.0383$ / $+0.0317$ / $+0.0466$ (3/3)
Q4	$+0.0289$ / $+0.0145$ / $+0.0203$ (3/3)	$+0.0292$ / $+0.0247$ / $+0.0393$ (3/3)

On that source (its 199 fakes against its 199 real clips) the arms score 0.0094 / 0.0158 / 0.0147, 0.0128 / 0.0207 / 0.0155 and 0.0094 / 0.0138 / 0.0218, and against every record of the split without a visual edit (audio-only edits keep their segments as ground truth), 0.0003 at every seed and arm; the unseen-source TalkLip slice (LRS3, 7,447 records) scores 0.0034 / 0.0031 / 0.0033, 0.0086 / 0.0073 / 0.0045 and 0.0065 / 0.0069 / 0.0082, against 0.0987 / 0.1092 / 0.1060, 0.1340 / 0.1354 / 0.1306 and 0.1331 / 0.1411 / 0.1519 on VoxCeleb2, the source corpus AV-Deepfake1M was built from. Every arm is lower on both unseen-source slices than on VoxCeleb2 at every seed, and both unseen-source slices stay below 0.022 at every seed and arm, so generator shift and source shift are not separated here: the Diff2Lip values license no statement about diffusion-era transfer beyond that slice. Read with the speaker-exposure slices above, as orderings only, the pooled level is carried by the source corpus shared with the training data. The checkpoints were trained on TalkLip-only AV-Deepfake1M; these rows refine that narrowness rather than retire it, and they are never placed in a column with AV-Deepfake1M, TVIL or DF40 values.

Published bars and their protocols. Table 46 lists the published temporal-localization numbers this paper quotes outside Table 1, each with the protocol it was obtained under. Every one is in-domain: the system was trained on the benchmark it is scored on. Ours differ: TVIL is scored zero-shot, and all our arms are vision-only on frozen features trained on FaceForensics++ masks and AV-Deepfake1M segments, so no row of the table is ranked against ours. The audio-visual range quoted in Section 2 mixes splits: WAFL scores the validation split that our in-domain rows use, DeformTrace reports its ablations on the test set, and AuViRe's table carries a "validation performance" footnote that no row is marked with. The TVIL bars also rest on different features: UniCaCLF's table puts UMMAFormer at 89.40 against the 88.68 of UMMAFormer's own paper, and MG-RWKV repeats UMMAFormer's published baseline digits.

Related context: we also evaluated the released ForgeryNet-trained OmniFD model (Liu et al., 2025a) zero-shot on our evaluation split (Table 1 context row: AP@0.5 0.228 full / 0.343 visual-subset under its released decode; 0.227 / 0.343 when capped to 16 proposals for a count-matched column). The row carries its protocol: an end-to-end trained video backbone, different training data, and a released evaluation pipeline that scores only each video's first 256 frames — 26.4% of split records (15,159 of the 57,340) are longer than that window and are scored truncated against full ground truth. No ranked comparison against our frozen-backbone arms is made in either direction; the row exists to place the nearest published relative under an honest protocol column. Table 1

Table 45: Zero-shot scoring of the nine ten-fold checkpoints of Table 1, unchanged, on the AV-Deepfake1M++ validation split (post hoc, Appendix A.8, item (xxv)): nothing trained or tuned on that release; decode carried from each row's configuration; official evaluation code; the split is named test in the harness metadata, as for AV-Deepfake1M. Columns: the full split (77,326 records) and its visual-manipulation subset (58,388; audio-only edits removed); the TalkLip slice (77,127: the 37,969 TalkLip fakes with every record without a visual edit, audio-only edits keeping their audio segments as ground truth as in the full split); the Diff2Lip slice (39,357: the 199 Diff2Lip fakes with the same records); and the three source corpora, VoxCeleb2 (69,481), LRS3 (7,447) and the `silent_videos` source (398: the 199 Diff2Lip fakes and that source's 199 real clips). Diff2Lip is the one diffusion generator in the split and all of its fakes come from the `silent_videos` source, which is absent from AV-Deepfake1M, so generator and source shift are confounded there; LRS3 is the unseen-source slice with the training generator. Mean and sample sd over seeds; differences from unrounded scores. These values form their own block, never placed in a column with AV-Deepfake1M, TVIL or DF40 values; the checkpoints were trained on TalkLip-only AV-Deepfake1M, a narrowness these rows refine, not retire. Row labels: naive = naive sum, attention = cross-task attention, label-free = attention without mask labels; s1–s3 = seeds; sd = sample sd; positive = count of seeds with a positive difference.

Arm	full	visual	TalkLip	Diff2Lip	VoxCeleb2	LRS3	silent
AP@0.5, zero-shot, official evaluation code							
naive, s1	0.0831	0.1281	0.0834	0.0003	0.0987	0.0034	0.0094
naive, s2	0.0924	0.1417	0.0927	0.0003	0.1092	0.0031	0.0158
naive, s3	0.0890	0.1373	0.0893	0.0003	0.1060	0.0033	0.0147
naive, mean	0.0882	0.1357	0.0884	0.0003	0.1046	0.0033	0.0133
naive, sd	0.0047	0.0069	0.0047	0.0000	0.0054	0.0001	0.0035
attention, s1	0.1166	0.1779	0.1170	0.0003	0.1340	0.0086	0.0128
attention, s2	0.1175	0.1790	0.1179	0.0003	0.1354	0.0073	0.0207
attention, s3	0.1120	0.1724	0.1124	0.0003	0.1306	0.0045	0.0155
attention, mean	0.1154	0.1764	0.1157	0.0003	0.1333	0.0068	0.0163
attention, sd	0.0029	0.0036	0.0029	0.0000	0.0025	0.0021	0.0040
label-free, s1	0.1152	0.1759	0.1156	0.0003	0.1331	0.0065	0.0094
label-free, s2	0.1221	0.1868	0.1225	0.0003	0.1411	0.0069	0.0138
label-free, s3	0.1330	0.2025	0.1334	0.0003	0.1519	0.0082	0.0218
label-free, mean	0.1234	0.1884	0.1238	0.0003	0.1420	0.0072	0.0150
label-free, sd	0.0090	0.0134	0.0090	0.0000	0.0094	0.0009	0.0063
AP@0.5: paired same-seed differences; positive = count of seeds with a positive value							
attention − naive, s1	+0.0335	+0.0497	+0.0336	−0.0000	+0.0352	+0.0052	+0.0034
attention − naive, s2	+0.0251	+0.0373	+0.0252	+0.0000	+0.0262	+0.0041	+0.0048
attention − naive, s3	+0.0230	+0.0351	+0.0231	−0.0001	+0.0246	+0.0012	+0.0008
attention − naive, positive	3/3	3/3	3/3	1/3	3/3	3/3	3/3
label-free − naive, s1	+0.0321	+0.0478	+0.0322	−0.0000	+0.0344	+0.0031	+0.0001
label-free − naive, s2	+0.0297	+0.0451	+0.0298	−0.0000	+0.0319	+0.0038	−0.0021
label-free − naive, s3	+0.0440	+0.0652	+0.0441	−0.0000	+0.0459	+0.0049	+0.0071
label-free − naive, positive	3/3	3/3	3/3	0/3	3/3	3/3	2/3
attention − label-free, s1	+0.0014	+0.0019	+0.0014	+0.0000	+0.0008	+0.0020	+0.0033
attention − label-free, s2	−0.0046	−0.0078	−0.0046	+0.0000	−0.0057	+0.0003	+0.0069
attention − label-free, s3	−0.0210	−0.0301	−0.0210	−0.0001	−0.0213	−0.0037	−0.0063
attention − label-free, positive	1/3	1/3	1/3	2/3	1/3	2/3	2/3
AR@10, zero-shot, official evaluation code							
naive, s1	0.1544	0.2200	0.1544	0.0238	0.1679	0.0413	0.1540
naive, s2	0.1560	0.2219	0.1557	0.0256	0.1695	0.0406	0.2260
naive, s3	0.1555	0.2212	0.1553	0.0252	0.1692	0.0395	0.1976
attention, s1	0.1736	0.2496	0.1737	0.0222	0.1878	0.0560	0.1526
attention, s2	0.1751	0.2521	0.1751	0.0219	0.1891	0.0580	0.1827
attention, s3	0.1698	0.2448	0.1699	0.0204	0.1846	0.0469	0.1547
label-free, s1	0.1722	0.2482	0.1724	0.0206	0.1868	0.0522	0.1339
label-free, s2	0.1741	0.2510	0.1741	0.0208	0.1883	0.0559	0.1536
label-free, s3	0.1736	0.2496	0.1734	0.0230	0.1875	0.0556	0.2249

Table 46: Published temporal-localization AP@0.5 quoted in this paper outside Table 1 (sources' percent form) with the protocol each was obtained under. All are in-domain; none is ranked against our rows (last line).

System	Evaluated on	Input	Trained on	AP@0.5
AV-Deepfake1M, audio-visual				
WAFL (Wang et al., 2026)	validation (57,340)	audio-visual (VideoMAE, Wav2Vec 2.0)	AV-Deepfake1M train	97.26
AuViRe (Koutlis & Papadopoulos, 2026)	split not stated	audio-visual (AV-HuBERT)	AV-Deepfake1M	96.5
DeformTrace (Zhu et al., 2026a)	test set (per its ablations)	audio-visual (RAVEn)	AV-Deepfake1M, 30 epochs	92.0
AV-Deepfake1M and AV-Deepfake1M++, vision-only				
LoCC (Datta et al., 2026)	lip-sync videos; split not stated	mouth crops	AV-Deepfake1M lip-sync videos	41.22
LipForensics (Klein et al., 2025)	++ validation, visual labels	mouth crops, frozen VSR front-end	++ train (subjects shared with validation, as on AV-Deepfake1M itself)	97.86
ActionFormer retrained by GEM-TFL (Zhu et al., 2026b)	split not stated; GEM-TFL's own mAP@0.5	vision-only, ResNet-50 features	AV-Deepfake1M	67.0
MesoInception4, as tabulated by UniSkip-Mamba (Yu et al., 2026)	split not stated	vision-only (MesoNet)	AV-Deepfake1M (as tabulated)	8.50
TVIL (video inpainting), visual features				
UMMAFormer (Zhang et al., 2023)	split not restated (test split = 508 clips)	TSN, ActivityNet-pretrained	TVIL train	88.68
UMMAFormer, as re-tabulated by UniCaCLF (Yin et al., 2026)	split not stated	TSN, Kinetics-pretrained	TVIL	89.40
UniCaCLF (Yin et al., 2026)	split not stated	TSN, Kinetics-pretrained	TVIL (split not stated)	87.07
MG-RWKV (Ni et al., 2026)	split not stated	TSN	TVIL, 45 epochs	91.22
Ours (Tables 1, 32)	AV-Deepfake1M validation; TVIL test, zero-shot	vision-only, frozen OpenCLIP ViT-B/16	FaceForensics++ masks + AV-Deepfake1M segments	—

carries a second OmniFD context row under a second protocol: the same released recipe retrained end-to-end on our two training sets (FaceForensics++ masks as its spatial task, AV-Deepfake1M segments as its temporal task) from the public Omnivore Swin-S initialization — never from the released ForgeryNet checkpoint — with OmniFD's own loss composition, optimizer and augmentation unchanged, for 10,000 iterations at the release's 128 + 8 batch composition. The pre-registration fixed that budget to match our pilot arms' temporal exposure, computing the target as $10 \times 1{,}000 \times 8 = 80{,}000$ clip draws; our mixed loader in fact draws each batch from one source or the other, so a pilot arm sees about 5,000 temporal batches ($\approx 40{,}000$ clips) beside about 5,000 spatial batches ($\approx 160{,}000$ images). The retrain therefore saw roughly twice the pilot arms' clip exposure and eight times their image exposure (Table 1), not the matched temporal budget the pre-registration intended — a correction to the pre-registered arithmetic, recorded 2026-09-12 (Appendix A.8), not a change to the run. Scored by the identical chain as the zero-shot row, it reaches AP@0.5 0.523404 full / 0.766654 visual-subset (0.518810 / 0.766499 when capped to 16 proposals). This is OmniFD's recipe on our pair under our exposure budget, not OmniFD at its best: one pre-registered retrain, chosen once and never tuned. The two OmniFD rows are never ranked against each other, and neither is ranked against our arms in either direction — the level gap to the frozen-backbone arms combines an end-to-end video backbone with the stated image over-exposure, and is described, not attributed.

The recipe with and without its mask loss (post hoc). We retrained this recipe at seeds 1, 2 and 3, twice per seed, identical except that the spatial mask loss has weight 0 in one (Appendix A.8). Images and clips pass through the shared trunk and neck in separate forward calls and no auxiliary loss is configured, so at weight 0 the image stream updates no parameter: the variant is temporal-only training of the same network, and the contrast is the recipe's whole effect of mask supervision, not the label-versus-architecture split of Section 5.2. Scored by the same chain on the uncapped export, the variant minus the full recipe is -0.0046 / -0.0009 / -0.0157 AP@0.5 on the full split at seeds 1 / 2 / 3 (mean -0.0071 ± 0.0077, sample standard deviation) and -0.0064 / -0.0004 / -0.0246 on the visual subset (-0.0105 ± 0.0126); at AR@10 it is -0.0012 / -0.0020 / -0.0092 (-0.0041 ± 0.0044) and -0.0009 / -0.0017 / -0.0166 (-0.0064 ± 0.0088). The variant is lower at every seed on all four readouts, a uniform paired sign pattern. Its mean is smaller than the spread of the three full-recipe runs alone (0.517261–0.526106 full / 0.755529–0.769526 visual-subset AP@0.5; the pre-registered retrain, whose configuration differs only in its seed, lies inside it at 0.523404 / 0.766654), but that marginal spread is not the yardstick for a paired contrast; with three seeds and no interval we read the sign pattern as directional only, not as a separated effect. Unlike the frozen testbed, this recipe gives mask gradients a path through the shared trunk. Table 1 keeps the pre-registered retrain, and none of the six runs is ranked against our arms.

What the cited multi-objective systems ablate. Section 2 states that none of the coupling methods it cites reports the comparison of Section 5.2: a second label source's loss removed while its data and the module that source fed are kept, under a frozen backbone that loss cannot reach. Table 47 records, for each of the fourteen multi-objective methods and systems cited in Sections 2.1 and 2.3, for two more cited only in Appendices A.6 and A.7, and for one single-objective localizer checked alongside them, what its ablation tables remove, read from the paper's own text. The statement is about reporting, not about whether those gains would survive the ablation. Three audited systems do remove an auxiliary loss with the architecture kept, but the auxiliary objective is computed from the target task's own labels (a contrastive term) or from none (a reconstruction term), so no second label source is removed. One method controls the architecture prior of a loss-side coupling by keeping its constraint networks with frozen random weights (Zamir et al., 2020); our frozen-module control (Table 2) is the in-architecture analogue.

A.8 Pre-registration timeline

Supports the protocol claim of Section 4. For each pre-registered stage behind the main comparisons and readouts, Table 48 gives the date its reading rules, closed lists or frozen populations were written down and the date its first result was read (the OmniFD retrain of Appendix A.7, a context row, was pre-registered on 2026-09-07, before any of its runs, and first read on 2026-09-10; the exploratory probes of Appendix A.5 are outside the table); Table 49 lists every dated amendment to those rules and lists. Rule dates are the dates written in the pre-registration notes and plan-approval records;

Table 47: Ablations reported by the multi-objective methods and systems cited in Sections 2.1 and 2.3 (fourteen), by two more cited only in the appendix (task-prototype retrieval, UniCaCLF) and by one single-objective localizer (FPN-Transformer), from each paper's own text. "The Section 5.2 comparison" is a second label source's loss removed, its data and the coupling module kept, under a frozen backbone the removed loss cannot reach; no row reports it.

System	Coupling	What its ablations remove	Nearest comparison
Multi-task learning from partial or disjoint supervision			
Cross-task consistency (Zamir et al., 2020)	output-space consistency networks, training-time loss only	the consistency terms (network identical); the constraint networks replaced by frozen random networks	architecture-prior control of a loss-side coupling
MTPSL (Li et al., 2022)	feature mapping into a joint task space inside a cross-task loss	the cross-task loss together with its mapping network; mapping variants with the loss kept	module and loss removed together
HiTTs (Zhang et al., 2025)	task tokens with inter-task attention; self-generated pseudo-label losses	the inter-task attention; the pseudo-label losses with the tokens kept	pseudo-supervision, not a second label source
Knowledge Assembly (Spinola et al., 2023)	shared backbone; twin-network output consistency across datasets	the consistency loss, labels kept; the other task's data; its labels, images kept (reID ssl)	labels removed, shared backbone kept
NexusFlow (Lin et al., 2026)	invertible flow aligning task features	the flow module, or its invertible layers, with both label partitions kept	module removed, labels kept
UniD (Sun et al., 2026)	one backbone distilled from frozen per-task specialists	video training and inference memory only	no label source to remove
REFINE (Xu et al., 2026)	frozen source feature concatenated with a target-trained residual encoder	encoder and adapter capacity only	no supervision ablation
Task-prototype retrieval (Oh et al., 2026)	prototype-based feature retrieval on a shared backbone	the method's own losses; module presence in the baseline rows unstated	own losses only
Joint and unified forgery models			
OmniFD (Liu et al., 2025a)	shared encoder; cross-task interaction module with learnable queries	the unified encoder; the interaction module (all labels kept); tasks with their data, module kept (Table II)	tasks removed with their data, module kept, end-to-end backbone
Venus-DeFakerOne (GuangJian Team, 2026)	detection and segmentation tokens of one language model	training stages; single-domain models; mask supervision (Table 9)	mask supervision removed; module presence unstated
GEM-TFL (Zhu et al., 2026b)	classification stage producing pseudo proposals for a regression stage	modules, added cumulatively	one label source
UVIF (Liu et al., 2026)	shared encoder for video and image inputs; two heads; pseudo-label and alignment losses	the image loss with its data and head ("trained on videos"); the coupling losses with both label sets kept	task removed with its data and head
Audio-visual detection and temporal localization			
Next-frame prediction (Anshul et al., 2025)	detection and localization models trained per task; contrastive auxiliary loss over the same labels	the contrastive losses, architecture kept	auxiliary loss over the same labels
FPN-Transformer (Zheng et al., 2025)	one localization head, no auxiliary task	feature extractors, epochs, learning rate	single objective
AuViRe (Koutlis & Papadopoulos, 2026)	localization plus self-supervised reconstruction of real speech	the reconstruction loss, architecture kept; a video-level detection loss added	auxiliary loss without a second label source
Divide and Conquer (Li et al., 2026)	independently trained modules fused at score level	whole modules	no joint training
UniCaCLF (Yin et al., 2026)	localization plus a supervised contrastive loss over the same labels	the contrastive loss with its module kept; the module	auxiliary loss over the same labels

first-result dates are the dates, written in the results notes, on which each result was first read. The code release contains these notes and a file giving each note's first and last commit time (in UTC; the dates here are local time, and a commit can trail the writing of a note by several hours). The project's research log is not released. Where the order is unfavourable the table says so. The DF40 re-run protocol was written after the first DF40 pass: it repaired the evaluated population and changed no scoring rule. The evaluation-reliability note and the 2026-09-03 additions to the interval lists were first committed after their first values had been computed (artifact file times); that they were written before those values rests on an unreleased working log.

Post-hoc analyses (not pre-registered). Twenty-five analyses were added between 2026-09-12 and 2026-09-22 (the last one scored on 2026-09-25), after every result above had been read. Each re-reads existing checkpoints under a frozen protocol, repeats a pre-registered control at another budget, encoder, dataset pair or recipe, or removes one component of an existing arm; none proposes a method or a new comparison family, and their role is to remove alternative explanations of the pre-registered results. They are: (i) the teacher-target probe of the coupling channel, covering the regularizer family, the pseudo-label arm and the summed-loss and attention arms (Appendix A.6, Table 43); (ii) scoring of the pilot control arms with the official evaluation code, with paired intervals for the two label contrasts (Tables 16 and 20); (iii) the last-epoch re-reading of Appendix A.1 and the fixed-prediction metric audit (Table 3); (iv) the existing pseudo-label configuration re-trained at the three matched seeds on the pilot budget, with the frozen zero-shot protocols applied to its checkpoints (Appendix A.6); (v) the $s{\to}t$-only controls described below, with their interval amendment; (vi), on 2026-09-13, the OmniFD retrain recipe re-run at three pinned seeds with and without its spatial mask loss (Appendix A.7); (vii), on 2026-09-13, scoring of the nine unfrozen-backbone checkpoints with the official evaluation code (Table 28); (viii), launched on 2026-09-13, attention without mask labels trained at the three matched seeds on the ten-fold budget (Tables 1 and 13); (ix), launched on 2026-09-13, two further matched seeds (4 and 5) of the naive, attention and label-free attention arms at pilot budget (Table 17); (x), launched on 2026-09-13, the label-free attention arm on the frozen DINOv2 backbone and under the unfrozen recipe at the three matched seeds (Tables 15 and 28); (xi), launched on 2026-09-13, the label-free attention arm on the second dataset pair at the three matched seeds (Table 27); (xii), launched on 2026-09-13, a spatial-only twin of the temporal-only control (Appendix A.2); and, launched on 2026-09-15 under a reading rule dated 2026-09-15 that was written before any of their runs existed, (xiii) the frozen-module and label-free spatial-to-temporal-only arms at the ten-fold budget (Table 2; the frozen module retains 126% and the single-task readout 85% of the mean attention gain, each positive at every seed) and (xiv) the four unfrozen arms under a second fine-tuning recipe (backbone at one hundredth of the heads' learning rate; Table 30: attention minus naive positive at 2/3 seeds, no consistent ordering), (xv) a linear-CKA comparison of the features the temporal head reads across the existing pilot checkpoints (Table 23), (xvi) the attention readout with one and two heads (Table 16: 56% and 88% of the mean attention gain retained, each positive at every seed; attention above the one-head readout at every seed and above the two-head readout at two of three; its main-text cutoff moved on 2026-09-22 before any value was read), (xvii) the naive arm with one normalization shared by both heads, with and without mask labels (the exact tie breaks at every seed; Table 18), and (xviii) the attention arm trained with its mask targets permuted within each batch (Tables 2 and 16); and, launched on 2026-09-18 after external feedback, under a reading rule written before any run existed, (xix) the reduced-mask-pool arms retrained with the mask pool re-iterated so that spatial and temporal batch counts match the full-dose arm (exposure-matched; Table 26: full-dose minus matched-dose positive at 3/3 seeds for the 10% pool and 2/3 for the 1% pool) and (xx) uninterrupted reruns of the two resumed unfrozen seed-1 runs (Table 28: the rerun values differ from the resumed ones by more than the seed-2/3 spread and replace nothing); and, launched on 2026-09-19 under a reading rule dated 2026-09-16 that was written before any of its runs existed, whose placement clause was amended on 2026-09-22 before any value was read, (xxi) the DINOv2 label ablation at the ten-fold budget, nine runs of the three DINOv2 pilot configurations (Table 15: the label-free arm retains 73% of the mean attention gain, positive at every seed, and full attention is above it at every seed); and, scored on 2026-09-22 under a reading rule written the same day before any export existed, (xxii) the three existing pilot checkpoints of the parameter-matched self-attention control on the official evaluation code (Tables 2 and 16: it carries 51% of the mean attention gain, positive at every seed, and attention exceeds it at every seed); and, scored on 2026-09-22 under a reading rule written the same day before any value existed, (xxiii) the nine ten-fold exports of Table 1 re-scored on four speaker-exposure quartiles of the evaluation set (Table 44: AP@0.5 falls from the least- to the most-exposed quartile at

Table 48: Pre-registration timeline: for each pre-registered stage, the date its rules were fixed and the date its first result was read. Rule dates are written in the pre-registration notes; first-result dates in the results notes. Local dates.

Stage	Rules fixed	First result
Official-harness scoring of the joint arms (round-trip self-check)	2026-08-26 (self-check required before any score is quoted); 2026-08-30 (table layout, before any arm score; the self-check had passed)	2026-08-31 (first two 10× rows, scored 2026-08-30)
OmniFD zero-shot context row (protocol and checkpoint fixed)	2026-08-26 (plan approval); checkpoint/configuration pairing settled 2026-09-01	2026-09-02 (harness-scored row)
Real-mass term ablation (three-outcome rule, Appendix A.9)	2026-08-26 (plan approval)	2026-08-28 (seed-1 rows and transfer readout); rule applied 2026-08-29 with all three seeds
Ranking / smoothness term ablations (same rule)	2026-09-02 (before any run started)	2026-09-02 (in-domain rows); 2026-09-03 (transfer readout; rule applied per term at $\geq$ 2 seeds)
Inversion statistic (estimator and 10 closed blocks)	2026-09-01 (blocks and random streams fixed)	2026-09-02 (all 10 blocks)
Evaluation-reliability analysis (closed 46-pair list)	2026-09-02 (note first committed 20:18, after the first values; written 14:31 according to an unreleased working log)	2026-09-02 (first 64 of 184 cells, 14:47–14:51, artifact file times); 2026-09-03 (all 130 computable cells)
Pilot confirmation seeds (matched-seed recipe, fused decode)	2026-07-27 (recipe fixed before launch)	2026-07-28 (first row); all nine rows by 2026-07-30
Ten-fold budget runs (sizing and reading questions)	2026-08-04 (before any ten-fold number; the 2026-08-11 approval changed no rule)	2026-08-13 (first ten-fold row)
Frozen joint evaluation set (weak-label caveat)	2026-07-29 (set frozen); scoring recipe fixed 2026-07-31 before any score	2026-07-31 (first scores)
TVIL zero-shot recipe	2026-08-18 (before any TVIL number)	2026-08-19 (all 12 rows)
DF40 generator list and score mapping	2026-07-26 (generator list frozen); 2026-08-20 (stage rules, before any DF40 number)	2026-08-21 (first pass; near-chance rule applied to 10 of 12 arms)
DF40 re-run protocol (population repair)	2026-08-21 to 2026-08-23, *after* the first pass (repaired data path and coverage minimum; scoring and mapping unchanged)	2026-08-24 (re-run; 36 of 39 generators)
OpenForensics zero-shot recipe	2026-08-27 (frozen before any number was read)	2026-08-29 (24 summary results)
Bootstrap interval lists (per family)	2026-08-03 (validation-AP paired-bootstrap protocol and closed 13-pair list); 2026-08-28 (scale-and-seeds list); later additions in Table 49	2026-08-04 (first validation-AP intervals). Exceptions: one ten-fold-seed TVIL interval was computed 2026-08-25 under the earlier zero-shot list; the 2026-09-03 additions were first committed after their first intervals
Single-task and label-free controls; data scaling; mask-label efficiency; rank-association closed lists	2026-08-26 (dated approval records: reading rules and closed lists fixed before any of these runs)	controls 2026-08-27 (label-free control rule applied 2026-08-29); mask-label rule applied 2026-08-28; scaling result final 2026-08-30; rank-association matrix 2026-09-01
Second dataset pair (three-outcome rule; zero-shot mirror readout)	2026-09-05 (before any second-pair run; identity check amended before any value; mirror details fixed before the first mirror row)	2026-09-05 (first rows); rule applied 2026-09-06 at three seeds per arm; mirror readout 2026-09-06
Unfrozen-backbone arms (memory-probe fallback, three-outcome rule, cross-generator secondary sentence)	2026-09-05 (before any unfrozen run)	2026-09-06 (first rows); rule applied 2026-09-06 at eight of nine rows (one sign flip decides it, which the ninth row could not undo); secondary sentence 2026-09-07

Table 49: Dated amendments to pre-registered rules and closed lists, in date order. The last column gives the order relative to the first result of the amended stage; the text says whether the change preceded the values it governs.

Date	Stage	What changed	Order vs. first result
2026-08-11	Frozen joint evaluation set	Weak-label caveat calibrated on a 116-clip validation sample; absolute scores quotable only as within-set orderings. Set, masks and metrics unchanged	after
2026-08-21	DF40	Re-run protocol after the first pass exposed an archive-layout defect: same scoring protocol and mapping on a repaired data path, new run identifier, first pass kept; expected coverage 33 of 39 generators	after
2026-08-21 to 2026-08-23	DF40	Archive member-resolution transforms and their admission rules written (08-21, 08-22), with no re-scoring below 33 of 39 generators; entries admitted only where the inventory proved the transform (08-23)	after
2026-08-22	TVIL zero-shot	Per-method readouts added, fixed before any per-method number	after
2026-08-26	DF40	One late-arriving generator admitted by the same rule; separate addendum pass under a new run identifier; the 2026-08-24 coverage result unchanged	after
2026-08-30	OpenForensics	Extension block for ten-fold seeds 2–3, provided for in the pre-registration; same protocol	after
2026-09-01	Bootstrap interval lists	Ten-fold seed-1 TVIL pair left under the earlier zero-shot list, under which its interval had been computed on 2026-08-25	after
2026-09-01	OmniFD zero-shot row	Checkpoint/configuration mismatch settled: the evaluation configuration inherits the released checkpoint's branch	before
2026-09-02	Bootstrap interval lists	Joint tube-mAP readout added for the DINOv2 pairs, before any interval on them	after
2026-09-02	Evaluation reliability	DF40 run identifiers accepted as a closed set; arms aligned on the scored face-local population. Written before any DF40 cell, after 64 tube and TVIL cells	after
2026-09-03	Bootstrap interval lists	Ranking/smoothness ablation pairs added after their point differences were known; first committed after their first intervals	after
2026-09-03	Bootstrap interval lists	DF40 interval code widened from one run identifier to a closed list so cross-run pairs can be computed; pairs aligned on the face-local population; first committed after their first intervals	after
2026-09-05	Second dataset pair	Identity check for five runs launched before the source echo existed replaced by configuration tests and a construction check; before any value	before
2026-09-05	Second dataset pair	Population and scorer of the zero-shot mirror readout fixed before any mirror number, after the first in-domain rows	after

every seed and arm; attention minus naive positive at every seed in every quartile); and, launched on 2026-09-22 under a reading rule written the same day before any run existed, (xxiv) the one-tenth unfrozen recipe of Table 28 at the ten-fold batch budget, five runs, naive and label-free attention at seeds 1 and 2 and full attention at seed 1 (Table 29: label-free attention minus naive negative at both landed seeds and attention minus naive negative at its one seed; levels described under one untuned recipe, never ranked against frozen or pilot rows); and, scored on 2026-09-25 under a reading rule written on 2026-09-22 before any export existed, (xxv) the nine ten-fold exports of Table 1 re-scored zero-shot on the AV-Deepfake1M++ validation split and five of its slices (Table 45: both attention arms above naive at every seed on the full split; the one diffusion slice, 199 Diff2Lip fakes from one unseen source, and the unseen-source TalkLip slice both below 0.022 at every seed and arm, so generator and source shift are not separated). Analyses (i)–(iii) and (vii) re-read checkpoints that already existed; (iv) adds runs of an unchanged configuration and scores them under protocols frozen earlier; (v) adds six runs of two new configurations (one flag, and that flag with mask-loss weights set to 0) scored by the unchanged official chain; (vi) adds six runs of the retrain configuration, three of them with a single loss weight set to 0, scored by the unchanged chain; (viii) adds three runs of the ten-fold attention configuration with its two mask-loss weights set to 0, (ix) six runs of the three pilot configurations at seeds 4 and 5, (x) three runs each of the DINOv2 and unfrozen attention configurations with their two mask-loss weights set to 0, (xi) three runs of the second pair's attention configuration with its two mask-loss weights set to 0 (read on that block's internal readout), and (xii) three runs of the naive configuration with its three temporal loss weights set to 0, all read under one reading rule dated 2026-09-15 that was written after the (viii) runs and five of the six (ix) runs had been scored, before the (x)–(xii) runs had finished, and before any of their values was read. Analysis (xv) re-reads twenty-one existing checkpoints and adds no run; (xvi) adds six runs of the attention configuration with a different head count, (xvii) six runs of the naive configuration with one added normalization, three of them with mask-loss weights set to 0, and (xviii) three runs of the attention configuration with its mask targets permuted, each under a reading rule dated 2026-09-15 that was written before any of its runs existed. Apart from (v), whose own pre-launch reading rule was replaced before any of its runs had finished, been exported or been scored, none changes a pre-registered rule, a frozen population or a metric. One pre-registered conditional branch was declared and not taken: the improved-stitched-grounding ladder of 2026-07-30 was to run only if the 10-epoch regularized stitched arm (seed 1) gained at most $+0.02$ over its naive stitched twin; the observed $+0.0321$ left it dormant, and no run from it exists.

Analysis (v) asks how much of the attention gain remains without its temporal-to-spatial direction, at matched seeds and pilot budget. Its first arm keeps only spatial-to-temporal attention; its three runs were launched on 2026-09-12, before this revision, under a reading rule committed that day. A gradient audit then showed that this arm still passes mask-loss gradients into the shared trainable patch projection, so it removes the temporal-to-spatial direction, not auxiliary supervision. A dated post-hoc protocol, fixed on 2026-09-12 before any of these runs had been exported or scored, replaced that rule (early-epoch internal validation AP, from at most the first four of ten epochs, had already appeared in the training logs) and added matched copies with mask-loss weights set to zero and every other configuration key unchanged; gradient guard tests verify the shared projection path and its removal. That protocol makes official full-split AP primary and the visual subset secondary, reports every seed, and permits no equivalence claim. One no-label run died before its first training epoch and was relaunched unchanged. An amendment dated 2026-09-13, written after five of the six new arm-seed exports had been scored and before any interval for these arms existed, added paired video-bootstrap intervals for a closed list of five contrasts ($B = 1000$, percentile, RNG root 20260913), accepted only after the interval driver reproduced the earlier seed-1 label intervals exactly; every cell is reported (Table 21). The protocol, its amendment and the results note were first committed after all six runs and their intervals, so the order stated here rests on the dates and file times recorded in those files, not on commit history. Spatial-to-temporal attention only, trained without mask labels, retains 76% of the mean attention gain over naive; the remaining gap to full attention is not attributed (Appendix A.2).

The rule committed on 2026-09-12 read internal AP against the ± 0.02 band and rested on the premise that temporal-to-spatial attention, through a shared normalization, is the only route by which the mask loss reaches anything the temporal head reads. The shared trainable patch projection is a second route, so that premise was incomplete. Applied as written, full attention minus spatial-to-temporal attention only is $+0.0058$ / -0.0005 / $+0.0189$ on internal AP, inside the band at every seed (the rule's "established" branch), and $+0.0120$ / -0.0004 / $+0.0274$ on official AP, outside it at

seed 3 ("unresolved"). Neither outcome is adopted: the revision protocol reports paired differences and uncertainty and does not turn a small difference into a proof of no effect.

Corrections to pre-registered arithmetic. One pre-registered quantity was miscomputed and is corrected here rather than silently restated. The OmniFD retrain's iteration budget (Appendix A.7) was set to match our pilot arms' temporal exposure on the assumption that every mixed batch carries both streams; the loader draws each batch from one source or the other, so the retrain in fact saw about twice the clip exposure and eight times the image exposure of our pilot arms. The run, its budget and its score are unchanged — only the exposure ratio reported with the row is, and the row was never ranked against our arms in either direction.

A.9 PRE-REGISTRATION EXCERPTS

Supports Appendix A.8. Excerpts are quoted verbatim from the dated records released with the code (formatting marks removed). The first two rules produced outcomes contrary to the working hypotheses. Those outcomes do not establish when the rules were fixed; the chronology depends on the accompanying dated records.

1. Real-mass ablation reading rule (plan-approval record, 2026-08-26): the inversion counts as removed if the ablated arm's "pooled face-local AUC $\geq$ 0.40 at every landed seed"; if its "pooled AUC $<$ 0.10 at $\geq$2 seeds", "the paper reports the inversion as reproduced under ablation, names the rank and smoothness terms as remaining candidates, and makes NO mechanism claim"; anything between is "partial attribution, reported with the per-seed numbers and no causal verb." Outcome (2026-08-29, all three seeds in): AUC 0.0070 / 0.0023 / 0.0027, below 0.10 at every seed — the real-mass hypothesis was rejected by the study's own pre-registered rule.
2. Rank/smoothness ablation rules (plan amendment, 2026-09-02, before any run started): the same three-outcome structure per term; an arm zeroing both terms would run only if both single-term arms met the rejection condition, and if it also met it "the paper states the inversion is not attributable to any single remaining term", in place of a mechanism claim and never as an upgrade. Outcome: both single-term arms met the rejection condition (2026-09-03); the arm zeroing both terms was then run and also met it (2026-09-04) — the statement in Appendix A.6 is the one this rule prescribed.
3. DF40 near-chance evidence branch (recorded 2026-08-20 in the stage's pre-registration, before any DF40 number existed): "if an arm's evidence AUC (`frame_prob`) $\leq$ 0.55 (near-chance in-distribution), that arm's DF40 per-paradigm row is read as descriptive only and the results note leads with the evidence number. The mapping is not swapped in any branch; a broken mapping is a reported finding, not a knob." Outcome (first pass, 2026-08-21): the rule applied to 10 of 12 arms, and every affected row — including the main text's sign-inversion finding — has carried the descriptive flag since.

A.10 CONFIDENCE-INTERVAL PROTOCOL

Supports every interval in Section 5. All intervals are paired bootstrap over evaluation units (clips, videos or images, stated per family), $B = 1000$, percentile method, with the full set-level metric recomputed per replicate. Manipulated and real units are resampled as separate strata on the joint evaluation set and on DF40; validation AP, TVIL, OpenForensics and the stitched set are resampled as a single stratum. The pair lists were fixed in dated notes, with the exceptions and amendments listed in Appendix A.8, and intervals exist only for listed pairs. Several listed pairs have no computed interval — ten-fold validation AP and most ten-fold transfer pairs, the three-fold and DINOv2 validation-AP pairs, and validation AP for the real-mass ablation, the label-free control and the mask doses — so those comparisons rest on per-seed sign patterns. Within-seed intervals never aggregate across seeds; cross-seed statements use sign patterns over paired per-seed deltas. Tables 50–55, 20 and 21 list every computed interval.

Readout resolution (Tables 56 and 57 give it per pair and seed; Section 5.1 summarizes it): each readout's resolution is measured under pre-registered settings, in the spirit of a minimum-detectable-effect analysis (Card et al., 2020) but without a power model, as the smallest evaluation-set size

at which the full-population sign of a paired difference reproduces in at least 95% of stratified subsamples, and the smallest gap the full set separates from zero under a paired bootstrap. The populations are the official benchmark's 57,340 records, the out-of-domain detection population, the 508-clip inpainting set and the 386-clip joint where-and-when set (286 manipulated + 100 real).

Metric definitions (the scopes and reading rules are in Section 4):

- *In-domain temporal localization.* Segment AP at tIoU 0.5 is the primary readout, with AR@10 as the secondary readout where tabulated (Appendix A.1); one fixed implementation and one fixed decode rule (Section 4) for every arm.
- *In-domain spatial localization.* FaceForensics++ validation frames: four from each manipulated video and two from its original video as reals, whose masks are empty (one third of the images). Unless stated otherwise, metrics are read at the epoch selected on temporal AP@0.5 and averaged over images. Mask IoU@0.5 binarises prediction and ground truth at that threshold; with an empty ground truth the intersection is empty, so every real image scores zero whatever is predicted. Value agreement (IoUdiff in Figure 7) is the fraction of pixels whose predicted soft value lies within 0.01 of the ground truth; L1 is the mean absolute pixel error.
- *Joint where-and-when readout.* A prediction scores through a *tube* IoU — temporal IoU multiplied by mean per-frame mask IoU — reported as tube-mAP at 0.2 (primary) with 0.5/0.75 as secondary thresholds, plus mask IoU within true-positive segments.
- *OpenForensics.* Mask IoU of the whole-image union of forged-face masks, under the same IoU family as the in-domain mask metric, restricted to forged-face annotations.
- *TVIL.* Segment AP@0.5, with AR@10 as the designated secondary readout.
- *DF40.* Video-level detection AUC under the frozen mapping: the maximum frame fake-probability over four evenly spaced frames per video (the freeze on FaceForensics++ validation, the reals asymmetry and the near-chance rule are stated in Section 4).

Table 50: In-domain validation AP@0.5 (internal metric; pilot budget; OpenCLIP backbone): paired differences with within-seed 95% bootstrap intervals, one column per seed. Unit: clip, single stratum over the 8,000 validation clips; $B = 1000$, percentile; AP recomputed on every replicate; seeds are never pooled. Arms: A cross-task attention, C output-space consistency, N naive sum; control arms are defined below the table. Point differences are the interval driver's, from unrounded AP, so a last digit can differ from the 4-dp tables (e.g. $+0.2517$ here, $+0.2516$ in Table 4).

Pair	Seed 1	Seed 2	Seed 3
Main arms			
$\mathrm{A} - \mathrm{N}$	$+0.2517\ [+0.2415, +0.2610]$	$+0.2935\ [+0.2827, +0.3041]$	$+0.2689\ [+0.2580, +0.2792]$
$\mathrm{C} - \mathrm{N}$	$+0.0272\ [+0.0236, +0.0302]$	$+0.0023\ [-0.0030, +0.0076]$	$+0.0078\ [+0.0028, +0.0124]$
Consistency-term ablations			
$\mathrm{C}_{\setminus\mathrm{rank}} - \mathrm{C}$	$+0.0013\ [-0.0004, +0.0032]$	$+0.0074\ [+0.0057, +0.0091]$	$-0.0513\ [-0.0565, -0.0460]$
$\mathrm{C}_{\setminus\mathrm{rank}} - \mathrm{N}$	$+0.0285\ [+0.0253, +0.0314]$	$+0.0097\ [+0.0046, +0.0147]$	$-0.0435\ [-0.0465, -0.0405]$
$\mathrm{C}_{\setminus\mathrm{smooth}} - \mathrm{C}$	$-0.0005\ [-0.0021, +0.0012]$	$+0.0083\ [+0.0028, +0.0139]$	$-0.0152\ [-0.0172, -0.0133]$
$\mathrm{C}_{\setminus\mathrm{smooth}} - \mathrm{N}$	$+0.0266\ [+0.0235, +0.0296]$	$+0.0107\ [+0.0075, +0.0138]$	$-0.0074\ [-0.0124, -0.0024]$
$\mathrm{C}_{\setminus\mathrm{both}} - \mathrm{C}$	$-0.0551\ [-0.0592, -0.0508]$	$+0.0079\ [+0.0030, +0.0132]$	$-0.0412\ [-0.0440, -0.0382]$
$\mathrm{C}_{\setminus\mathrm{both}} - \mathrm{N}$	$-0.0280\ [-0.0311, -0.0251]$	$+0.0102\ [+0.0069, +0.0133]$	$-0.0334\ [-0.0385, -0.0279]$
Attention controls			
$\mathrm{A} - \mathrm{A}_{\mathrm{frozen}}$	$+0.0887\ [+0.0806, +0.0965]$	$+0.1237\ [+0.1166, +0.1315]$	$+0.0821\ [+0.0757, +0.0885]$
$\mathrm{A}_{\mathrm{frozen}} - \mathrm{N}$	$+0.1630\ [+0.1547, +0.1722]$	$+0.1698\ [+0.1617, +0.1788]$	$+0.1868\ [+0.1775, +0.1956]$
Attention plus consistency (weight 0.05)			
A+C − A	$-0.0104\ [-0.0167, -0.0039]$	$-0.0168\ [-0.0218, -0.0121]$	$+0.0015\ [-0.0045, +0.0085]$

$\mathrm{C}_{\setminus\mathrm{rank}}$, $\mathrm{C}_{\setminus\mathrm{smooth}}$, $\mathrm{C}_{\setminus\mathrm{both}}$: consistency with its ranking term, its smoothness term, or both removed. $\mathrm{A}_{\mathrm{frozen}}$: attention module kept at its random initialization. A+C: attention with the consistency term at weight 0.05.

Table 51: Joint where-and-when readout on the 386-clip set (286 manipulated, 100 real; landmark-derived mouth-region pseudo-masks, so rankings may also be biased): paired differences with within-seed 95% intervals for tube-mAP@0.2 (top) and mask IoU@0.5 inside true-positive segments (bottom). Unit: clip, manipulated and real clips resampled as separate strata; $B = 1000$, percentile; metric recomputed per replicate. Arms: A cross-task attention, C output-space consistency, N naive sum.

Tube-mAP@0.2			
Pair	Seed 1	Seed 2	Seed 3
Pilot budget, OpenCLIP			
A − N	-0.0237 $[-0.0444, -0.0046]$	-0.0055 $[-0.0157, +0.0010]$	$+0.0143$ $[-0.0113, +0.0411]$
C − N	$+0.0670$ $[+0.0264, +0.1118]$	$+0.0532$ $[+0.0271, +0.0835]$	-0.0278 $[-0.0569, -0.0080]$
A+C − C	$+0.0111$ $[-0.0227, +0.0467]$	$+0.0189$ $[-0.0169, +0.0558]$	$+0.2352$ $[+0.1852, +0.3029]$
Stitched-supervision mix, OpenCLIP (one run per arm)			
C − N		$+0.0579$ $[+0.0323, +0.0940]$	
Ten-fold budget, OpenCLIP			
A − N	-0.0093 $[-0.0217, +0.0009]$	$+0.0186$ $[-0.0135, +0.0535]$	$+0.0778$ $[+0.0461, +0.1162]$
C − N	-0.0019 $[-0.0173, +0.0153]$	-0.0315 $[-0.0583, -0.0088]$	-0.0115 $[-0.0280, +0.0031]$
Pilot budget, DINOv2			
A − N	$+0.0001$ $[-0.0002, +0.0011]$	-0.0001 $[-0.0010, +0.0000]$	-0.0058 $[-0.0152, -0.0018]$
C − N	$+0.0027$ $[+0.0005, +0.0088]$	$+0.3360$ $[+0.2930, +0.3905]$	$+0.3013$ $[+0.2531, +0.3580]$
Mask IoU@0.5 inside true-positive segments			
Pair	Seed 1	Seed 2	Seed 3
Pilot budget, OpenCLIP			
A − N	-0.0225 $[-0.0326, -0.0117]$	-0.0126 $[-0.0211, -0.0043]$	$+0.0050$ $[-0.0087, +0.0179]$
C − N	$+0.0263$ $[+0.0089, +0.0426]$	$+0.0039$ $[-0.0127, +0.0213]$	-0.0635 $[-0.0780, -0.0485]$
A+C − C	-0.0285 $[-0.0405, -0.0165]$	-0.0151 $[-0.0295, -0.0010]$	$+0.1053$ $[+0.0950, +0.1164]$
Stitched-supervision mix, OpenCLIP (one run per arm)			
C − N		$+0.0466$ $[+0.0318, +0.0621]$	
Ten-fold budget, OpenCLIP			
A − N	-0.0436 $[-0.0563, -0.0313]$	$+0.0201$ $[+0.0028, +0.0378]$	$+0.0351$ $[+0.0179, +0.0525]$
C − N	$+0.0049$ $[-0.0101, +0.0194]$	-0.0266 $[-0.0461, -0.0083]$	-0.0157 $[-0.0295, +0.0004]$
Pilot budget, DINOv2			
A − N	$+0.0006$ $[-0.0045, +0.0055]$	-0.0259 $[-0.0317, -0.0203]$	$+0.0032$ $[-0.0045, +0.0111]$
C − N	-0.0772 $[-0.0868, -0.0677]$	$+0.2705$ $[+0.2521, +0.2919]$	$+0.1562$ $[+0.1381, +0.1734]$

"A+C": attention with the consistency term at weight 0.05, compared with consistency alone. Stitched-supervision mix: both arms trained with 25% of batches stitched from FaceForensics++, same seed. No three-fold intervals exist for this readout.

Table 52: Stitched FaceForensics++ validation (560 clips; internal AP@0.5 with 2-frame boundary masking, "masked", primary, and without it): paired differences with within-seed 95% intervals. Unit: clip, single stratum; $B = 1000$, percentile; AP recomputed per replicate. Pilot checkpoints never saw stitched clips; the stitched-mix pair is one run per arm after 20 epochs with 25% of batches stitched. C output-space consistency, N naive sum.

Pair	AP@0.5	Seed 1	Seed 2	Seed 3
Pilot checkpoints (no stitched training clips)				
C − N	masked	$+0.0065$ $[+0.0030, +0.0105]$	-0.0005 $[-0.0048, +0.0044]$	$+0.0071$ $[+0.0043, +0.0118]$
C − N	unmasked	$+0.0067$ $[+0.0037, +0.0102]$	$+0.0017$ $[-0.0015, +0.0060]$	$+0.0074$ $[+0.0046, +0.0109]$
Stitched-supervision mix (one run per arm, 20 epochs)				
C − N	masked		$+0.0004$ $[-0.0089, +0.0113]$	
C − N	unmasked		-0.0003 $[-0.0083, +0.0078]$	

Table 53: TVIL zero-shot AP@0.5 (508 test clips): paired differences with within-seed 95% bootstrap intervals. Unit: clip, single stratum; $B = 1000$, percentile; AP recomputed per replicate. Budgets are separate blocks; –: no interval computed. Arms: A cross-task attention, C output-space consistency, N naive sum.

Pair	Seed 1	Seed 2	Seed 3
Pilot budget, OpenCLIP			
A − N	-0.0153 $[-0.0367, +0.0034]$	-0.0206 $[-0.0411, -0.0020]$	-0.0336 $[-0.0531, -0.0155]$
C − N	-0.0194 $[-0.0305, -0.0083]$	-0.0164 $[-0.0293, -0.0059]$	-0.0102 $[-0.0243, +0.0004]$
Ten-fold budget, OpenCLIP			
A − N	$+0.0106$ $[-0.0194, +0.0408]$	–	–
C − N	-0.0268 $[-0.0410, -0.0122]$	–	–

Table 54: DF40 pooled face-local detection AUC (frozen mapping; reals are FaceForensics++ test originals), pilot budget, OpenCLIP: paired differences with within-seed 95% intervals. Unit: video; manipulated and real videos resampled as separate strata; $B = 1000$, percentile; AUC recomputed per replicate. Arms near chance on FaceForensics++ validation make these descriptive patterns, not transfer claims. A cross-task attention, C output-space consistency, N naive sum.

Pair	Seed 1	Seed 2	Seed 3
Main arms			
A − N	-0.066 $[-0.096, -0.036]$	-0.047 $[-0.079, -0.009]$	-0.099 $[-0.137, -0.063]$
C − N	-0.542 $[-0.583, -0.499]$	-0.521 $[-0.564, -0.480]$	-0.500 $[-0.547, -0.457]$
Consistency-term ablations			
$\mathrm{C}_{\setminus\mathrm{mass}}$ − C	$+0.0026$ $[+0.0018, +0.0032]$	-0.0061 $[-0.0077, -0.0046]$	-0.0008 $[-0.0012, -0.0005]$
$\mathrm{C}_{\setminus\mathrm{mass}}$ − N	-0.5399 $[-0.5792, -0.5020]$	-0.5267 $[-0.5688, -0.4897]$	-0.5005 $[-0.5484, -0.4503]$
$\mathrm{C}_{\setminus\mathrm{rank}}$ − C	$+0.0119$ $[+0.0101, +0.0138]$	-0.0054 $[-0.0069, -0.0039]$	$+0.0019$ $[-0.0023, +0.0078]$
$\mathrm{C}_{\setminus\mathrm{rank}}$ − N	-0.5306 $[-0.5726, -0.4927]$	-0.5260 $[-0.5664, -0.4878]$	-0.4978 $[-0.5452, -0.4553]$
$\mathrm{C}_{\setminus\mathrm{smooth}}$ − C	-0.0021 $[-0.0030, -0.0012]$	$+0.0085$ $[+0.0071, +0.0099]$	-0.0023 $[-0.0030, -0.0017]$
$\mathrm{C}_{\setminus\mathrm{smooth}}$ − N	-0.5445 $[-0.5838, -0.5059]$	-0.5121 $[-0.5549, -0.4687]$	-0.5020 $[-0.5446, -0.4533]$

$\mathrm{C}_{\setminus\mathrm{mass}}$, $\mathrm{C}_{\setminus\mathrm{rank}}$, $\mathrm{C}_{\setminus\mathrm{smooth}}$: consistency with its real-mass, ranking or smoothness term removed. Main-arm rows resample the full re-scored population (51,765 videos, of which the face-local subset is scored); ablation rows resample the face-local population aligned across scoring passes (35,688 manipulated + 140 real videos).

Table 55: OpenForensics zero-shot whole-image forged-face mask IoU, pilot budget, OpenCLIP: paired differences with within-seed 95% intervals on Test-Dev (18,895 images; IoU@0.5 primary, IoU@0.1 secondary) and Test-Challenge (45,000 images; IoU@0.5). Unit: image, single stratum per split; $B = 1000$, percentile; mean IoU recomputed per replicate. Ten-fold checkpoints carry no intervals. A cross-task attention, C output-space consistency, N naive sum.

Readout	Seed 1	Seed 2	Seed 3
C − N			
Dev IoU@0.5	-0.0949 $[-0.0974, -0.0925]$	-0.1864 $[-0.1897, -0.1831]$	-0.0798 $[-0.0820, -0.0776]$
Chal. IoU@0.5	-0.1025 $[-0.1041, -0.1009]$	-0.1948 $[-0.1969, -0.1927]$	-0.0875 $[-0.0890, -0.0861]$
Dev IoU@0.1	-0.1989 $[-0.2025, -0.1955]$	-0.2059 $[-0.2098, -0.2015]$	-0.2327 $[-0.2364, -0.2291]$
A − N			
Dev IoU@0.5	$+0.0004$ $[-0.0014, +0.0022]$	$+0.0193$ $[+0.0170, +0.0214]$	-0.0314 $[-0.0332, -0.0295]$
Chal. IoU@0.5	-0.0050 $[-0.0062, -0.0039]$	$+0.0362$ $[+0.0347, +0.0377]$	-0.0315 $[-0.0328, -0.0302]$
Dev IoU@0.1	$+0.0150$ $[+0.0132, +0.0167]$	-0.0099 $[-0.0111, -0.0087]$	-0.0636 $[-0.0661, -0.0612]$

Table 56: Evaluation reliability: full-set paired difference (Δ), smallest sign-stable fraction of the evaluation set (%), and resolvable gap, per seed. Readouts are separate tables and budgets and backbones separate blocks; each last row summarises all computed cells of that readout, including control pairs not shown. –: inputs not available. The three-fold budget has no consistency arm. A cross-task attention, C output-space consistency, N naive sum.

(a) Official evaluation code, AP@0.5; 57,340 records, single stratum

	Seed 1			Seed 2			Seed 3		
Pair	Δ	%	gap	Δ	%	gap	Δ	%	gap
Pilot budget, OpenCLIP									
A − N	$+0.0646$	$\leq$5	0.0020	$+0.0672$	$\leq$5	0.0022	$+0.0778$	$\leq$5	0.0022
C − N	$+0.0075$	$\leq$5	0.0005	$+0.0045$	$\leq$5	0.0007	-0.0002	$>$90	0.0006
Three-fold budget, OpenCLIP									
A − N	$+0.0474$	$\leq$5	0.0021	$+0.0490$	$\leq$5	0.0021	$+0.0613$	$\leq$5	0.0021
Ten-fold budget, OpenCLIP									
A − N	$+0.0335$	$\leq$5	0.0020	$+0.0182$	$\leq$5	0.0021	$+0.0234$	$\leq$5	0.0020
C − N	$+0.0018$	15	0.0010	$+0.0011$	40	0.0008	-0.0052	$\leq$5	0.0011
Pilot budget, DINOv2									
A − N	$+0.0677$	$\leq$5	0.0021	$+0.0689$	$\leq$5	0.0022	$+0.0631$	$\leq$5	0.0023
C − N	-0.0047	$\leq$5	0.0003	-0.0029	$\leq$5	0.0005	-0.0046	$\leq$5	0.0006
Median, all 21 cells: 5%; gap 0.0020 (range 0.0003–0.0023).									

(b) Joint where-and-when tube-mAP@0.2; 386 clips (286 manipulated + 100 real strata)

	Seed 1			Seed 2			Seed 3		
Pair	Δ	%	gap	Δ	%	gap	Δ	%	gap
Pilot budget, OpenCLIP									
A − N	-0.0237	40	0.0207	-0.0055	50	0.0080	$+0.0143$	75	0.0271
C − N	$+0.0670$	30	0.0456	$+0.0532$	15	0.0278	-0.0278	30	0.0239
Stitched-supervision mix, OpenCLIP (one run per arm)									
C − N	$+0.0579$	15	0.0308						
Three-fold budget, OpenCLIP									
A − N	–	–	–	–	–	–	–	–	–
Ten-fold budget, OpenCLIP									
A − N	-0.0093	50	0.0110	$+0.0186$	75	0.0343	$+0.0778$	10	0.0350
C − N	-0.0019	$>$90	0.0159	-0.0315	30	0.0259	-0.0115	50	0.0151
Pilot budget, DINOv2									
A − N	$+0.0001$	90	0.0007	-0.0001	60	0.0005	-0.0058	15	0.0066
C − N	$+0.0027$	30	0.0043	$+0.3360$	$\leq$5	0.0499	$+0.3013$	$\leq$5	0.0523
Median, all 19 cells: 30%; gap 0.0239 (range 0.0005–0.0523).									

%: smallest tested fraction (5, 10, 15, 20, 30, 40, 50, 60, 75 or 90%) at which at least 95% of 1,000 stratified subsamples drawn without replacement reproduce the full-set sign; $\leq$5: already at the smallest; $>$90: at none. Gap: half-width of the full-set paired 95% percentile bootstrap interval ($B = 1000$), the smallest $|\Delta|$ the full set separates from zero. Joint-set labels are pseudo-masks (exploratory rankings).

Table 57: Evaluation reliability, zero-shot readouts (continued from Table 56; same columns and conventions). Each last row summarises all computed cells of that readout, including control pairs not shown. DF40 rows for arms near chance in distribution are descriptive. The three-fold budget has no consistency arm. A cross-task attention, C output-space consistency, N naive sum, L attention without mask labels (paired against the temporal-only control, identical to N on this readout).

(c) TVIL zero-shot AP@0.5; 508 clips, single stratum

	Seed 1			Seed 2			Seed 3		
Pair	Δ	%	gap	Δ	%	gap	Δ	%	gap
Pilot budget, OpenCLIP									
A − N	−0.0153	60	0.0197	−0.0206	40	0.0187	−0.0336	20	0.0179
C − N	−0.0194	20	0.0113	−0.0164	20	0.0116	−0.0102	50	0.0122
L − N	−0.0145	60	0.0177	−0.0188	40	0.0180	−0.0237	40	0.0169
Three-fold budget, OpenCLIP									
A − N	−0.0374	30	0.0253	−0.0206	40	0.0197	−0.0547	10	0.0222
Ten-fold budget, OpenCLIP									
A − N	+0.0106	90	0.0310	−0.0300	30	0.0230	−0.0266	20	0.0202
C − N	−0.0268	20	0.0147	−0.0346	10	0.0178	−0.0208	30	0.0162
Pilot budget, DINOv2									
A − N	−0.0423	30	0.0329	+0.0014	>90	0.0282	−0.0239	60	0.0318
C − N	−0.0518	10	0.0193	−0.0536	10	0.0195	−0.0506	10	0.0222
Median, all 45 cells: 40%; gap 0.0178 (range 0.0053–0.0329).									

(d) DF40 pooled face-local AUC; 35,828 videos (35,688 manipulated + 140 real strata)

	Seed 1			Seed 2			Seed 3		
Pair	Δ	%	gap	Δ	%	gap	Δ	%	gap
Pilot budget, OpenCLIP									
A − N	−0.0663	15	0.0311	−0.0472	40	0.0372	−0.0992	10	0.0354
C − N	−0.5424	≤5	0.0390	−0.5206	≤5	0.0425	−0.4997	≤5	0.0473
Three-fold budget, OpenCLIP									
A − N	+0.1181	10	0.0401	+0.0205	75	0.0362	−0.0359	50	0.0349
Ten-fold budget, OpenCLIP									
A − N	+0.0257	60	0.0328	−0.1225	10	0.0342	−0.2052	≤5	0.0388
C − N	−0.5240	≤5	0.0395	−0.4984	≤5	0.0401	−0.6013	≤5	0.0352
Pilot budget, DINOv2									
A − N	−0.0215	75	0.0380	−0.1539	≤5	0.0392	−0.1632	≤5	0.0375
C − N	−0.4141	≤5	0.0372	−0.4775	≤5	0.0395	−0.4895	≤5	0.0322
Median, all 45 cells: 10%; gap 0.0353 (range 0.0004–0.0473).									

A.11 RANK-ASSOCIATION MATRIX

Supports Section 6.3; numeric twin of Figure 10. Table 58 gives the in-domain diagnostics against the out-of-domain targets (Kendall τ) over the sixteen pre-registered pilot-scale checkpoints (nine seed runs of the three arms plus seven further checkpoints: three stitched-mix arms, the parameter-matched control at three seeds and the soft-coupled variant). Diagnostics enter at their raw values; for mask L1 and flicker smaller is better, so a positive τ against those two rows means the out-of-domain score rises as the in-domain diagnostic gets worse. The two joint-set targets cover all sixteen checkpoints for every diagnostic except the in-distribution evidence AUC; that diagnostic and the two transfer targets had been scored for the nine seed runs alone when the matrix was fixed, and the other seven checkpoints, scored on them later for the out-of-sample check below (whose 34 checkpoints include all sixteen), are not added, so the TVIL and DF40 columns and the evidence-AUC row are at $n = 9$ throughout. Top-1 agreement fails in every one of the 28 diagnostic–target cells: the checkpoint that leads an in-domain diagnostic is never the one that leads an out-of-domain target (Section 6.3).

Table 58: In-domain diagnostics against out-of-domain targets, Kendall τ over the sixteen pre-registered pilot-scale checkpoints ($n = 16$ in the two joint-set columns except the evidence-AUC row; $n = 9$ in the TVIL and DF40 columns and in the evidence-AUC row). Columns: joint tube-mAP@0.2 and joint in-segment mask IoU on the where-and-when set; zero-shot inpainting AP@0.5 (TVIL); generator-shift face-local AUC (DF40). Rows marked lower-is-better enter at their raw values.

In-domain diagnostic	tube-mAP@0.2	in-seg. mask IoU	TVIL AP@0.5	DF40 AUC
validation AP@0.5	-0.29 (n=16)	$+0.02$ (n=16)	-0.50 (n=9)	-0.06 (n=9)
mask IoU@0.5	-0.38	$+0.10$	-0.06	$+0.39$
mask value agreement	$+0.41$	-0.03	$+0.11$	-0.44
mask L1 (lower better)	-0.42	$+0.02$	$+0.00$	$+0.44$
flicker (lower better)	-0.49	-0.05	$+0.50$	$+0.72$
mask stability @0.5	-0.14	$+0.00$	-0.39	$+0.17$
in-distribution evidence AUC	-0.61 (n=9)	-0.28 (n=9)	-0.06 (n=9)	$+0.28$ (n=9)

Out-of-sample check of the strongest cells (Section 6.3 gives the reading). Thirty-four checkpoints from fourteen configuration families (pilot scale, same backbone) were split in half under a seeded split fixed in advance, and the diagnostic most rank-associated with each transfer target on one half was read on the other. For inpainting transfer the selected diagnostic (in-distribution evidence AUC, $\tau = -0.33$ on the selection half) came out at $\tau = +0.08$ on the held-out half, its sign flipped; for generator-shift detection the selected diagnostic (mask IoU@0.5, $\tau = +0.39$) kept its sign but weakened to $\tau = +0.26$, below the threshold fixed in advance ($|\tau| \geq 0.30$ with a stable sign, on both targets, by the same diagnostic). Neither instance qualifies; the association is never a selection criterion.

Scope of the rank-transfer claim. Rank instability is a property of a readout and a shift, not of transfer in general. On ActivityForensics (Bao et al., 2026), a vision-only temporal-localization benchmark whose shift is between generator families, the average-AP ordering of four localizers is, by its tables, unchanged across the intra-domain, open-world and both cross-domain directions. Our claim is therefore stated over the readouts and shifts of Section 5.3 (Appendix A.4) — a joint where-and-when criterion, video inpainting, generator shift and in-the-wild swaps, each scored zero-shot from a model trained on one lip-sync family — and not as a general claim that in-domain rankings never transfer.

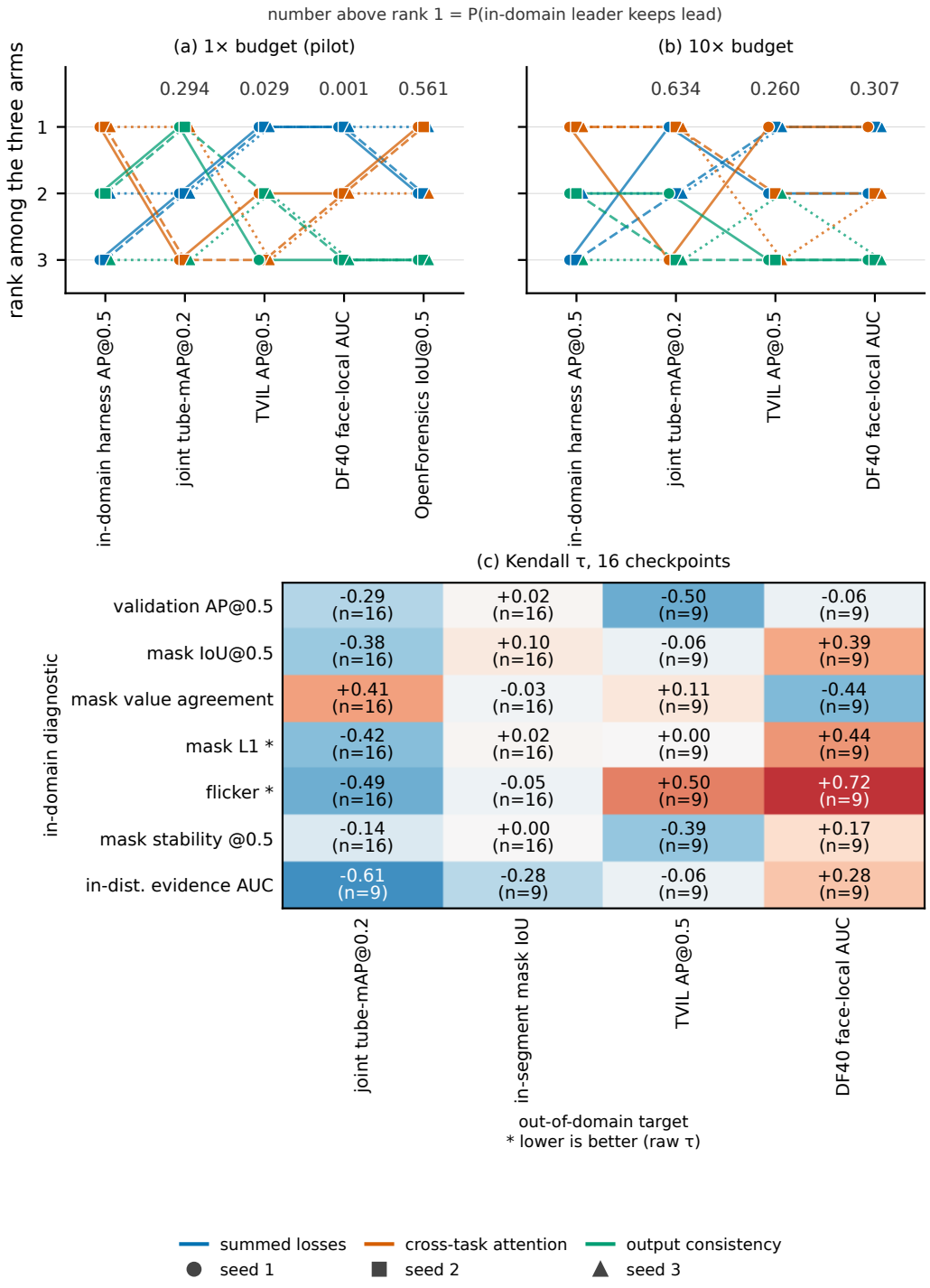


Figure 10: In-domain rankings do not transfer. (a), (b): pilot and ten-fold budgets, never in one panel; arms ranked by in-domain harness AP@0.5 and transfer readouts; colours arms, markers seeds; numbers above transfer columns: inversion statistic (Section 4; probability the in-domain leader also leads; low means inversion; not a p-value). (c): Kendall's τ, up to sixteen checkpoints (per-cell n shown), raw orientation, starred rows lower-is-better. DINOv2 mirror: Figure 3.