TRIROUTE: ROUTING EVIDENCE UNDER DOMAIN SHIFT FOR CROSS-DOMAIN DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models are expected to integrate complementary signals across modalities, yet in practice they often rely on spurious shortcuts from a single modality. This failure is particularly critical in cross-domain audio-visual deepfake detection, where models may ignore visual manipulation evidence and instead rely on easier but unreliable audio cues. A common approach to mitigate such failures is to enforce invariance through regularization or domain-adversarial objectives. However, it remains unclear whether these methods change how models make decisions, or merely reshape representations without altering their usage. In this work, we argue that robustness in cross-domain deepfake detection is not only a problem of enforcing invariance but also of controlling how information is routed at inference time. We demonstrate that loss-level interventions fail to resolve representation-level entanglement in standard binary decompositions, where modality-specific signals and domain nuisance signals remain mixed, leading to persistent shortcut reliance. We introduce TRIROUTE, a three-way factorization that separates shared content, modality-specific task-relevant components, and residual domain-dominant signals. Only the first two components reach the prediction head; the residual branch is structurally excluded at inference time. Across cross-domain audio-visual deepfake benchmarks, TRIROUTE yields substantial improvements in the diagnostic FV-RA setting, where correct predictions require reliance on visual evidence, and the advantage reproduces on a second natively annotated target (LAV-DF) and a constructed directional set (AVLips) using the same source-selected checkpoint. Crucially, we provide multiple lines of evidence that these gains arise from changes in inference-time routing rather than representation invariance. Head ablation, input-level masking, residual-only readout, and subspace probing show that predictions depend critically on the visual-unique component, while domain information remains present in the representation. Our results suggest a shift in perspective for multimodal deepfake detection: instead of suppressing spurious features, robustness can be achieved by explicitly structuring the pathways through which decisions are made.

1 INTRODUCTION

Multimodal learning plays a central role in audio-visual deepfake detection, where reliable decisions require combining evidence from both visual and audio streams (Zhou & Lim, 2021; Chung & Zisserman, 2016; Feng et al., 2023; Liang et al., 2024; Du et al., 2025). In principle, such models should leverage complementary signals—for example, visual manipulation artifacts and cross-modal inconsistencies—to identify forged content. In practice, however, multimodal models often rely on shortcut signals from a single modality, failing to integrate evidence across modalities (Smeu et al., 2025; Geirhos et al., 2020). This issue becomes particularly severe under domain shift: in cross-domain audio-visual deepfake detection, models frequently ignore visual manipulation evidence and instead rely on easier but unreliable audio cues, resulting in near-chance performance when audio becomes uninformative. A common response is to enforce invariance, through domain-adversarial training (Ganin et al., 2016) or feature regularization (Sun & Saenko, 2016), on the assumption that robustness follows from eliminating nuisance information. Such approaches can improve aggregate metrics, but it remains unclear whether they change how models make decisions, or merely reshape representations without affecting how they are used at inference time. This raises a basic question: do existing methods improve robustness by learning better representations, or by changing how models use them?

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.

We show that standard loss-level approaches are insufficient to resolve shortcut reliance here. In binary decompositions, modality-specific task signals and domain-related nuisance remain entangled within the same branch, so the classifier can keep exploiting spurious cues even under domain-adversarial objectives; the model has no structural mechanism to separate reliable visual evidence from shortcut signals. We argue that the problem is therefore not primarily one of objective design, but of representation structure and inference-time routing. To address this, we introduce **TRIROUTE**, a three-way factorization that decomposes multimodal representations into (1) shared content, (2) modality-specific task-relevant components, and (3) residual domain-dominant signals. By explicitly separating these roles and restricting the prediction head to use only task-relevant components, TRIROUTE encourages the model to rely on visual manipulation evidence when audio cues are misleading.

Empirically the pattern is consistent: improvements concentrate in the FV-RA (FakeVideo-RealAudio) setting, where audio shortcuts are uninformative and correct predictions require visual evidence, while other categories sit near ceiling for every model. Controlled comparisons show these gains cannot be explained by additional objectives alone and emerge only once the representation is structurally factorized, and the same source-selected checkpoint reproduces the advantage on a second natively annotated target, LAV-DF (Cai et al., 2022), and on a constructed directional subset of AVLips (Liu et al., 2024).

More importantly, we provide direct evidence that the improvement is a change in inference-time routing rather than representation invariance. Head ablation (Section 6.4) and input-level masking (Section 6.5) show that TRIROUTE predictions depend critically on the visual-unique component while non-factorized models stay largely insensitive; a residual-only readout transfers poorly even though the residual is highly domain-decodable (Section 6.6); and subspace probing (Section 6.6) concentrates transferable visual evidence in a dedicated low-dimensional component while domain information remains decodable. Robustness is achieved not by removing domain information, but by keeping the decision from depending on it.

Contributions. The building blocks we use—multimodal factorization (Liu et al., 2025), shared/private separation (Bousmalis et al., 2016), gradient-reversal adversaries (Ganin et al., 2016), supervised domain latents (Ilse et al., 2020)—all have precedents, and we claim novelty for none of them in isolation. Our contribution is the composition and the inference-time principle governing it. **1)** We show that loss-level approaches fail to prevent shortcut reliance because of representation-level entanglement: objectives that fail on a two-way representation become effective once a third, excluded branch exists. **2)** We formulate an explicit *routing* principle—domain-sensitive representations are structurally prevented from contributing to prediction, not merely regularized toward invariance—and instantiate it as TRIROUTE. **3)** We validate the mechanism rather than assuming it, via head ablation, input masking, temporal shuffling, residual-only readout, and subspace probes. **4)** We demonstrate substantial FV-RA gains on FakeAVCeleb (Khalid et al., 2021), reproduce them on LAV-DF and a constructed AVLips subset with an unchanged source-selected checkpoint, and report matched comparisons against SpeechForensics (Liang et al., 2024) and AVAD (Feng et al., 2023), including the regimes where those methods are stronger.

2 RELATED WORK

Multimodal deepfake detection and shortcut reliance. Audio-visual deepfake detection relies on the assumption that realistic speech, lip motion, and facial dynamics must remain mutually consistent over time. Early work explored direct audio-visual fusion and synchronization cues (Zhou & Lim, 2021; Chung & Zisserman, 2016). More recent approaches incorporate self-supervised anomaly detection, speech representation learning, or explicit modeling of modality-shared inconsistencies and modality-specific forensic traces (Feng et al., 2023; Liang et al., 2024; Smeu et al., 2025; Du et al., 2025; Liu et al., 2024). Large-scale audio-visual speech models such as AV-HuBERT (Shi et al., 2022) underpin several of these detectors (Liang et al., 2024; Smeu et al., 2025); we build on the same representations but factorize them to control which information the detector may use. Most existing detectors still optimize a single prediction pathway, and therefore neither control nor evaluate how decisions are routed across modalities under domain shift.

Evaluation protocols. Most audio-visual deepfake detection work evaluates in-dataset: FakeAVCeleb studies report train/test splits or leave-one-manipulation protocols (Feng et al., 2023; Oorloff et al., 2024), and AV-Deepfake1M provides standardized in-dataset benchmarks across

AV1M, LAV-DF, and FAVC (Cai et al., 2024; 2022); visual-only benchmarks such as FaceForensics++ (Rössler et al., 2019) do not expose the audio-visual routing failure we study. The closest cross-domain effort is AVH-Align (Smeu et al., 2025), which reports supervised AV1M→FAVC transfer and diagnoses shortcuts such as leading-silence bias. We instead organise the evaluation around modality-category reporting and routing analysis, emphasising the FV-RA setting.

Multimodal factorization. The closest related work is FCD (Liu et al., 2025), which factorizes representations into synergistic, unique, and redundant components for general multimodal prediction. Where FCD treats the residual as task-irrelevant uncertainty to suppress, TRIROUTE uses it as a supervised domain-predictive sink that is explicitly excluded from the prediction pathway, turning factorization into a mechanism for routing rather than disentanglement alone. The perspective is related to shared-private decomposition and debiasing: DSN (Bousmalis et al., 2016) separates shared and private components for adaptation, DIVA (Ilse et al., 2020) models domain, class, and residual variation generatively, and bias-adversarial methods (Kim et al., 2019; Nam et al., 2020) suppress shortcut reliance through adversarial or dual-network training. We reuse these ingredients; the contribution is how they combine for routing control: (i) a supervised domain-predictive residual branch, (ii) structural exclusion of that branch from the head at inference, and (iii) explicit validation of the resulting routing under cross-domain FV-RA stress tests.

Domain alignment and adversarial adaptation. DANN (Ganin et al., 2016) uses gradient reversal to make domain prediction difficult, and CORAL-style methods match feature statistics such as covariance (Sun & Saenko, 2016). Neither addresses which signals the classifier may use: on entangled representations, global alignment can leave shortcut pathways intact. Our experiments make this concrete—on the same two-way representation neither objective recovers FV-RA transfer, while the same objective family becomes effective once a third excluded branch exists (Table 1). TRIROUTE therefore reframes the problem from feature alignment to routing control: rather than asking whether all features can be aligned, we restrict which features the detector is allowed to use.

3 PROBLEM SETTING

We study cross-domain audio-visual deepfake detection. Let $\mathcal{D} = \{(x_i^v, x_i^a, y_i, d_i)\}_{i=1}^N$ denote training samples, where x_i^v is a sequence of visual observations, x_i^a is the synchronized audio, $y_i \in \{0, 1\}$ is the real/fake label, and $d_i \in \{1, \dots, M\}$ indicates a nuisance partition available during training. Training is performed on AV-Deepfake1M, while evaluation is conducted on disjoint target datasets (FakeAVCeleb, LAV-DF, AVLips) that are never accessed during training, validation, or model selection. In our setting, nuisance supervision is derived only from source-side speaker-group pseudo-labels (a hash-based speaker partition), and no target-domain information is used at any stage.

Our goal is a scoring function $f(x^v, x^a)$ that stays predictive when the nuisance statistics associated with d differ between training and test data, a setting in which models often exploit non-transferable shortcuts such as audio artifacts or dataset-specific correlations. Rather than requiring full domain invariance, we control how such signals are used: we seek a representation in which manipulation evidence remains discriminative for y while predictable but non-transferable signals are isolated from the prediction pathway. The focus is therefore not on eliminating domain information but on restricting the model's ability to rely on it.

4 TRIROUTE

4.1 OVERVIEW

Figure 1 gives a conceptual overview of TRIROUTE. Audio-visual forensics under dataset shift involves three qualitatively distinct signal types: cross-modal speech content that should generalize across datasets, modality-unique manipulation evidence that is the detection target, and domain-correlated nuisance that must not reach the prediction head. We therefore design a three-branch factorization in which each branch has an explicit functional role and a corresponding supervision signal, and the prediction head is structurally restricted to the first two. Starting from frozen AV-HuBERT features $z_v, z_a \in \mathbb{R}^{1024}$, we first extract shared cross-modal factors and modality-specific hidden states, then split the latter into unique task-relevant and residual components:

$$\big[s_v, h_v\big] = F_v(z_v), \quad \big[s_a, h_a\big] = F_a(z_a), \qquad \big[u_v, r_v\big] = G_v(h_v), \quad \big[u_a, r_a\big] = G_a(h_a), \tag{1}$$

where s_v, s_a are shared task-relevant factors, u_v, u_a are modality-unique task factors, and r_v, r_a form the residual branch that is allowed to absorb shortcut and domain signal. The task and residual

representations are $Z_{\text{task}} = \text{cat}(s_v, u_v, s_a, u_a)$ and $Z_{\text{res}} = \text{cat}(r_v, r_a)$. For readability, we refer to the shared factor as *content*, the visual-unique task factor as the key *manipulation* carrier for FV-RA, and the residual branch as the *domain* factor.

Why three branches. The count follows from the routing objective rather than from an architectural preference. Prediction under domain shift requires distinguishing evidence that should always be usable (shared cross-modal content), evidence that must be usable precisely when the other modality is uninformative (modality-unique manipulation traces), and signal that is predictive in-domain but should never reach the head (domain-correlated nuisance). Three roles therefore give the minimal partition on which an exclusion rule can be stated. Our experiments support the functional separation this targets—the residual is highly domain-decodable yet transfers poorly on its own (Section 6.6)—but they do not show three to be universally optimal: we have not run a branch-count ablation, and the partition study in Section 6.7 varies the nuisance labels, not the number of branches.

4.2 Three-Way Factorization

Manipulation factor. The manipulation factor is intended to encode modality-unique evidence that cannot be explained by authentic cross-modal correspondence. In practice, the visual-unique component u_v is the most important carrier for the visual-fake category FV-RA, while u_a supports audio-fake cases. These unique components feed the task head together with the shared factors.

Content factor. The content factor captures shared speech semantics and identity-preserving motion that should remain useful across domains. We align the shared factors across modalities using an audio-visual alignment objective:

$$\mathcal{L}_{\text{align}} = -\sum_{t=1}^{T} \log \frac{\exp(\text{sim}(s_{v,t}, s_{a,t})/\tau)}{\sum_{j\in\mathcal{N}(t)} \exp(\text{sim}(s_{v,t}, s_{a,j})/\tau)}, \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity, τ is a temperature, and $\mathcal{N}(t)$ is a local temporal window.

Domain factor. The residual branch is explicitly optimized to predict the source-side nuisance partition, $\mathcal{L}_{\text{dom}} = \text{CE}(C_d(Z_{\text{res}}), d)$. This branch acts as a sink for signal that is predictable from the source-side partition but need not transfer; with a speaker-level partition this is speaker- and recording-correlated variation rather than any specific codec or acquisition factor, and our probes establish that Z_{res} is highly partition-decodable without attributing that decodability to any one factor (Section 6.6).

4.3 Detection and Domain Push-Pull Objectives

At inference time, only the task representation Z_{task} is used; the residual branch Z_{res} is structurally excluded from the prediction pathway. The clip-level fake probability $\hat{y} = C_y(Z_{\text{task}})$ is optimized via binary cross-entropy, $\mathcal{L}_{\text{fake}} = \text{BCE}(\hat{y}, y)$. To keep task-relevant and residual factors separated, we apply an orthogonality penalty:

$$\mathcal{L}_{\text{ortho}} = \left\|s_v^\top r_v\right\|_F^2 + \left\|s_a^\top r_a\right\|_F^2 + \left\|u_v^\top r_v\right\|_F^2 + \left\|u_a^\top r_a\right\|_F^2. \tag{3}$$

To suppress easy domain prediction from the task branch, we apply a gradient-reversal domain classifier, $\mathcal{L}_{\text{adv}} = \text{CE}(C_{\text{adv}}(\text{GRL}(Z_{\text{task}})), d)$, while $\mathcal{L}_{\text{dom}}$ simultaneously attracts domain information into Z_{res}. This push-pull effect does not aim to make Z_{task} fully domain-invariant; instead it makes shortcut routing into the prediction head less attractive. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{fake}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}. \tag{4}$$

Loss-weight selection. The push-pull pair $(\lambda_{\text{dom}}, \lambda_{\text{adv}})$ is the critical one: λ_{dom} pulls domain signal into the residual while λ_{adv} pushes it out of the task branch. Values (λ_{align}=0.5, λ_{dom}=1.0, λ_{adv}=1.0, λ_{ortho}=0.1) are fixed once on source-only diagnostics and reused everywhere; a one-at-a-time sensitivity audit (Appendix J) shows the defaults sit at or near the optimum of each sweep and that even the worst configuration beats the best DANN/GRL result by a wide margin.

5 Experiments

Datasets and Evaluation. We train on a fixed, duplicate-free 8,000-clip subset of AV-Deepfake1M (AV1M) (Cai et al., 2024) with a disjoint 2,000-clip source-validation subset (Appendix C). Our primary evaluation is the FakeAVCeleb (FAVC) (Khalid et al., 2021) 70/30 public test split (N = 6,667),

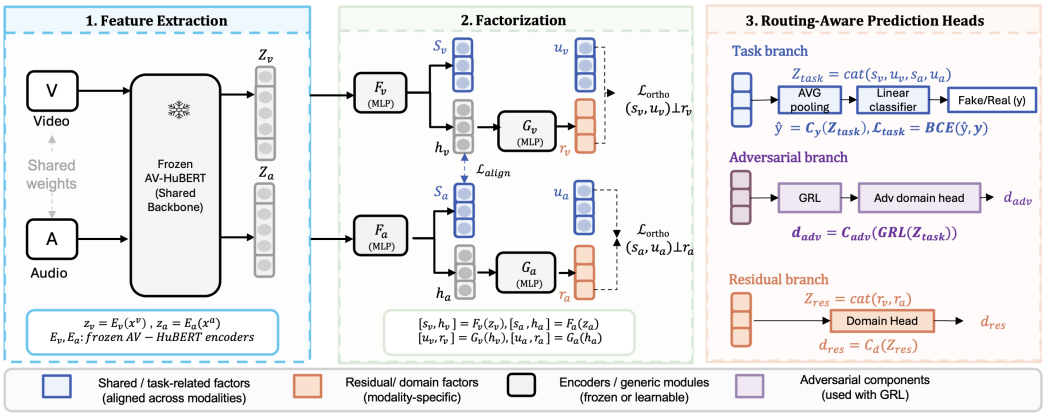

Figure 1: Overview of TRIROUTE. Frozen AV-HuBERT features are decomposed into shared (s_v, s_a), modality-unique (u_v, u_a), and residual (r_v, r_a) factors. The task head uses only the shared and unique subspaces; domain push-pull concentrates nuisance information in the residual branch.

following the protocol of recent shortcut analyses of audio-visual deepfake datasets (Smeu et al., 2025). We report threshold-free AUC and average precision (AP) separately for three informative categories: (i) RealVideo-FakeAudio (RV-FA), (ii) FakeVideo-RealAudio (FV-RA), and (iii) FakeVideo-FakeAudio (FV-FA). FV-RA is the most diagnostic category because audio shortcuts are intentionally unhelpful there, forcing the model to use genuine visual manipulation evidence.

Because one target dataset cannot establish generality, we additionally evaluate two independent targets with the *unchanged* source-selected checkpoint (Section 6.3): LAV-DF (Cai et al., 2022), whose native modality-level annotations make it a genuine second FV-RA benchmark ($N = 13{,}358$), and AVLips (Liu et al., 2024), which provides only binary labels; audio fingerprinting matches every AVLips fake to a real source clip, so its full set ($N = 7{,}602$) is an FV-RA-type directional target. Construction details and the matched external-baseline protocol are in Appendix E.

Evaluation protocol and checkpoint selection. All results use a strict source-only protocol, which we state explicitly because the cross-domain claim rests on it. Training and validation use AV1M clips only, and checkpoint selection uses a single criterion—AUC on the held-out AV1M *source* validation split, with patience-15 early stopping. Target-domain access is blocked at the data-loader level, so no target-domain clip, label, statistic, training callback, or evaluation result is read at any point during training, validation, checkpoint selection, or hyperparameter selection; once a checkpoint is fixed, each target split is evaluated exactly once. This holds uniformly for TRIROUTE, every internal control, the unsupervised variants in Appendix T, and the additional targets in Section 6.3. The nuisance partition for $\mathcal{L}_{\text{dom}}$ and $\mathcal{L}_{\text{adv}}$ is likewise source-side only: $d = \text{hash}(\text{speaker_id}) \bmod 2$ over AV1M speakers. We release the exact training, selection, and one-shot evaluation scripts with configuration files, frozen splits, and checkpoint identifiers so the protocol can be reproduced end-to-end.

Baselines. We compare TRIROUTE against **AVH-Align/sup** (Smeu et al., 2025), the published method training on AV1M and evaluating on FAVC under a compatible AV-HuBERT protocol. We additionally evaluate two strong audio-visual forensic methods under the identical frozen target splits using their publicly released pretrained models and no retraining or target-specific adaptation: **SpeechForensics** (Liang et al., 2024) and **AVAD** (Feng et al., 2023). Both are training-free, real-speech-pretrained methods that do not follow the AV1M supervised source protocol; we discuss the resulting asymmetry in Section 6.3. Internal ablation controls are: **two-way baseline** (coarser two-branch factorization, no domain routing); **two-way + DANN/GRL** (gradient-reversal domain classifier on the same two-way representation, no residual sink); **two-way + CORAL** (covariance-statistic matching on the same representation); and **three-way without domain push-pull** (same three-branch structure as TRIROUTE but without the residual domain-prediction or task adversarial objectives). All internal variants use identical AV-HuBERT features and classifier backbone. AVFF (Oorloff et al., 2024), which trains primarily on FakeAVCeleb, is excluded as it reverses the transfer direction.

Implementation Details. TRIROUTE is built on frozen AV-HuBERT (Shi et al., 2022) features ($C = 1{,}024$, 25 fps, mouth-crop visual stream). Factorization modules and classifier are lightweight MLPs (3–4 layers, LayerNorm, ReLU) with 256-dimensional shared, unique, and residual com-

Table 1: Cross-domain detection on FakeAVCeleb (AV1M $\rightarrow$ FAVC, $N = 6{,}667$), AUC/AP, mean over three seeds. AVH-Align/sup is the published AV1M-trained checkpoint of Smeu et al. (2025). Every checkpoint is selected on AV1M source validation AUC; no FakeAVCeleb data enters training, validation, or model selection. Internal baselines share the same backbone and protocol, so differences isolate representation structure and routing.

Method	Structure	Overall	RV-FA	**FV-RA (key)**	FV-FA
AVH-Align/sup (Smeu et al., 2025)	two-way supervised	$0.777/0.994$	$0.999/0.999$	$0.519/0.955$	$0.999/1.000$
Two-way baseline	two-way factorization	$0.760/0.993$	$1.000/1.000$	$0.483/0.949$	$0.998/1.000$
Two-way + DANN/GRL control	two-way + GRL	$0.767/0.994$	$1.000/1.000$	$0.496/0.955$	$1.000/1.000$
Two-way + CORAL control	two-way + CORAL	$0.768/0.994$	$1.000/1.000$	$0.521/0.956$	$0.980/0.999$
Three-way, no domain push-pull	three-way factorization	$0.814/0.996$	$1.000/1.000$	$0.600/0.973$	$0.999/1.000$
TRIROUTE (ours)	three-way + push-pull	$\mathbf{0.922}/\mathbf{0.998}$	$0.999/0.999$	$\mathbf{0.838}/\mathbf{0.989}$	$0.994/1.000$

ponents per modality. We use Adam (lr$=10^{-3}$, batch size 1), a maximum of 100 epochs with patience-15 early stopping on source-validation AUC, and fix $\lambda_{\text{align}}=0.5$, $\lambda_{\text{dom}}=1.0$, $\lambda_{\text{adv}}=1.0$, $\lambda_{\text{ortho}}=0.1$, GRL $\alpha=1.0$. The supervised main experiment uses three seeds. Extended details (training scale and epoch definition, nuisance label derivation, DANN/CORAL matched settings, compute budget, and seed policy) are in Appendix C.

6 RESULTS AND ANALYSIS

6.1 CROSS-DOMAIN DETECTION PERFORMANCE

Table 1 compares TRIROUTE against the AV1M-trained AVH-Align/sup checkpoint from Smeu et al. (2025) and a progression of internal ablation baselines. AVH-Align/sup is a strong two-branch supervised baseline that achieves FV-RA AUC of 0.519 despite training on the same source data, establishing a competitive non-factorized ceiling. The two-way baseline achieves near-ceiling performance on RV-FA and FV-FA but falls below chance on FV-RA (0.483 AUC). The two global adaptation controls show marginal improvement over the two-way baseline: DANN/GRL reaches $0.496/0.955$ FV-RA and CORAL reaches $0.521/0.956$, but both remain near or below chance. This confirms that global domain alignment alone is insufficient when no residual shortcut sink is available: reducing domain discrepancy does not specify which evidence the classifier should use for FV-RA. Three-way factorization without domain push-pull improves FV-RA to $0.600/0.973$, but remains well below the full TRIROUTE result.

The baseline does not fail for lack of capacity but because its prediction pathway is structurally exposed to shortcut signals, which also rules out the "it is just better regularization" reading—the same objectives fail on a binary representation yet become effective once the representation is factorized, so the gain comes from the structural separation rather than from added optimization pressure.

Under the same source-only protocol, TRIROUTE reaches 0.838 FV-RA AUC as a mean over three independent seeds, improving on the two-way baseline by $+0.355$ and on the three-way no-push-pull control by $+0.238$; overall AUC rises to 0.922 while RV-FA and FV-FA remain near ceiling. Robust FV-RA routing therefore emerges once the prediction pathway is structurally separated from the residual shortcut sink, without any target-domain data. Domain label quality and hyperparameter sensitivity are quantified in Section 6.7 and Appendix J.

6.2 PROTOCOL VERIFICATION: INDEPENDENT REPETITION

All reported runs, including every row of Table 1, use a source-only pipeline in which target access is blocked at the data-loader level rather than merely avoided by convention. Because the headline claim depends entirely on the absence of target-domain information, we additionally repeated the three headline runs from scratch with the same seeds, again selected by source-validation AUC and evaluated once on the frozen FAVC split: they give 0.8012, 0.8264, and 0.8440 FV-RA AUC—0.824 ± 0.022, and 0.916 ± 0.011 overall (Appendix D)—against 0.838 ± 0.024 and 0.922 ± 0.012 in Table 1. The shift of 0.014 and 0.006 is run-to-run variability of batch-size-1 training under an unchanged protocol, and the gain over every control reproduces in full.

6.3 ADDITIONAL TARGET DOMAINS AND MATCHED EXTERNAL BASELINES

A single source$\rightarrow$target pair cannot show that the routing advantage generalizes. We therefore evaluated the *unchanged* source-validation-selected checkpoint on two further targets with no target-specific

Table 2: Cross-target FV-RA AUC under a matched frozen protocol. TRIROUTE (mean over three source-selected seeds, no target adaptation) and AVH-Align/sup are supervised AV1M-trained detectors; SpeechForensics and AVAD are training-free methods evaluated with released weights on the identical splits and reported as references. LAV-DF carries native modality labels; AVLips is FV-RA-type by audio audit.

Target	Evaluation	TRIROUTE	AVH-Align/sup	SpeechForensics	AVAD
FakeAVCeleb	native FV-RA (N=6,667)	0.838	0.519	**0.975**	0.731
LAV-DF	native FV-RA (N=13,358)	**0.914**	0.672	0.795	0.560
AVLips	FV-RA-type by audit (N=7,602)	**0.789**	0.505	—	—

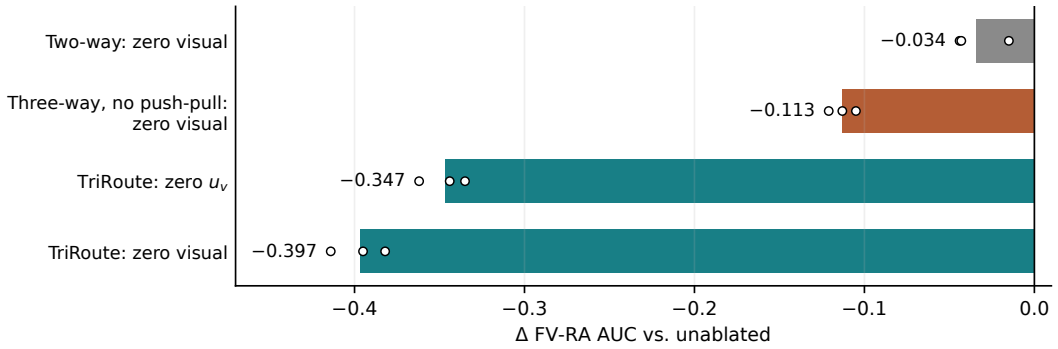

Figure 2: Head-ablation changes in FAVC FV-RA AUC, three-seed means with per-seed dots. For TRIROUTE, removing u_v lowers AUC by 0.347; removing all visual factors lowers it by 0.397, compared with 0.034 for the two-way baseline.

training, validation, checkpoint selection, or hyperparameter tuning, and evaluated SpeechForensics and AVAD on the identical frozen splits using their released pretrained models (Table 2).

The routing advantage over supervised cross-domain detectors reproduces on both additional targets: on LAV-DF, a genuine second FV-RA benchmark with native modality-level annotations, TRIROUTE reaches 0.914 against 0.672 for AVH-Align/sup (Appendix E). On the FV-RA-type AVLips set 0.789 against 0.505. AVAD reaches 0.731 on FAVC FV-RA and falls to 0.560 on LAV-DF. Full-set AVLips results for all internal model families are in Appendix P.

SpeechForensics is the more informative comparison, and it is stronger than TRIROUTE where it is designed to be: 0.975 on FAVC FV-RA against TRIROUTE's 0.838. FAVC FV-RA is a lip-sync regime matched to what a synchronization verifier is built to detect. The instructive observation is the cross-regime pattern under one protocol: SpeechForensics drops to 0.795 on LAV-DF FV-RA and to 0.688 overall AUC on AV1M (Appendix K), where TRIROUTE remains at 0.914 and 0.997. The two methods address different problems, training-free lip-sync verification versus a single supervised detector transferred across manipulation regimes, and the comparison delineates that boundary.

6.4 MECHANISM-LEVEL ROUTING EVIDENCE: HEAD ABLATION

Head-ablation results (Table 5, Figure 2; full details in Appendix G) show that masking the visual-unique branch u_v in TRIROUTE drops FV-RA from 0.838 to 0.491 (Δ=−0.347), while the same ablation on the two-way baseline causes only a negligible drop (−0.034)—an order-of-magnitude gap, with seed SDs of at most 0.016, that is direct mechanism-level evidence of rerouting through the visual-unique factor. Appendices F, K, V, H, and S give further comparisons, the leading-silence artifact control, and qualitative cases. Figure 3 shows the per-frame view: $\|u_v^t\|$ peaks at manipulation events in a fake FV-RA clip and stays flat in a real one, while the residual carries the bulk of the activation.

6.5 INPUT-LEVEL ROUTING VERIFICATION

The head-ablation analysis zeroes out internal model components, which could conflate routing with capacity effects. As a complementary and model-agnostic test, we perturb the *input features* on the FAVC FV-RA category and measure the resulting AUC change. This is a direct input-side intervention: if the trained head routes FV-RA decisions through the visual channel, full visual

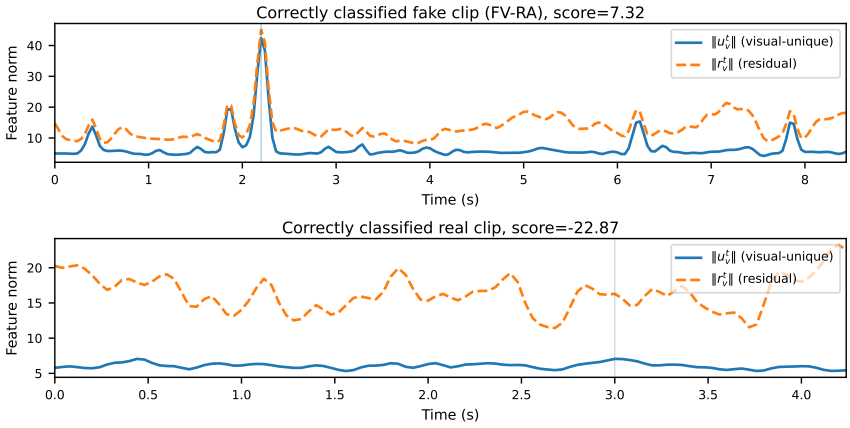

Figure 3: Per-frame routing map for two FAVC clips classified by TRIROUTE: feature norm of the visual-unique factor $\|u_v^t\|$ (solid) and the residual factor $\|r_v^t\|$ (dashed). *Top:* a correctly classified fake (FV-RA) clip, where $\|u_v^t\|$ peaks at manipulation events. *Bottom:* a correctly classified real clip, where $\|u_v^t\|$ stays low while the residual carries the bulk of the activation.

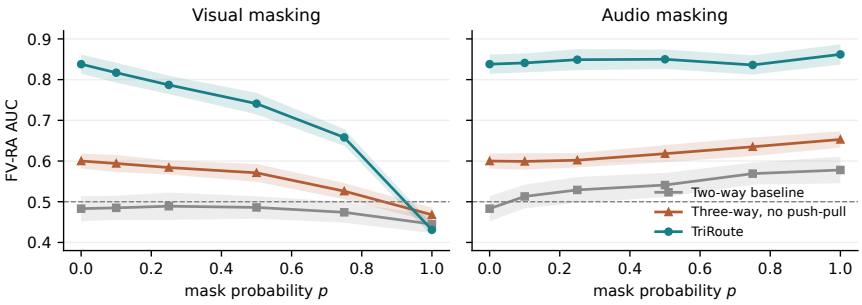

Figure 4: Progressive input masking on FAVC FV-RA, three-seed means with ± one SD bands. *Left:* visual masking lowers AUC, steepest for TRIROUTE. *Right:* audio masking raises AUC, most for the two-way baseline, whose real-audio stream acts as a distractor.

masking should cause a large drop; full audio masking, by contrast, should be neutral or beneficial (since FV-RA has real audio, masking it removes a potentially misleading signal).

Table 11 (Appendix L) reports mean FV-RA AUC over three independent seeds at full audio masking ($p = 1.0$) and full visual masking ($p = 1.0$). Two patterns stand out. First, the two-way baseline is nearly flat under full visual masking ($\Delta_{\text{visual}} = -0.038$) but *improves* when its audio input is removed ($\Delta_{\text{audio}} = +0.095$): a model below chance gets better once the misleading real-audio stream is gone, exactly what an audio shortcut predicts. Second, TRIROUTE shows the opposite profile: full visual masking lowers FV-RA from 0.838 to 0.431 ($\Delta = -0.407$), while full audio masking changes it by only $+0.024$. Three-way factorization without push-pull is intermediate ($\Delta_{\text{audio}} = +0.053$, $\Delta_{\text{visual}} = -0.132$), matching its intermediate FV-RA AUC in Table 1.

Progressive masking sweep. Sweeping the mask probability $p \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0\}$ separately on each channel (Figure 4, Table 12, Appendix M) moves both curves in the expected directions, with adjacent points fluctuating by at most 0.014: visual masking lowers FV-RA and audio masking raises it, with both effects growing from the two-way baseline (-0.038, $+0.095$) through the no-push-pull model to TRIROUTE (-0.407, $+0.024$).

Frame order versus visual content. The factorization modules split features frame-independently, raising the question of whether the splitter needs explicit temporal context. Two observations bound the exposure: the splitter operates on frozen AV-HuBERT features that already carry upstream temporal context, and a temporal-shuffle perturbation changes FV-RA only from 0.838 to 0.840 (SD 0.023), whereas full visual masking reduces it to 0.431. The splitting stage therefore depends far more on contextualized visual content than on frame order. This does not show explicit temporal

routing to be unnecessary; neighbour-aware splitting remains a promising refinement we have not tested.

6.6 What the Excluded Residual Branch Carries

Excluding a branch from the head is only justified if that branch is a poor basis for transferable prediction, so we measured directly what is lost. A head reading *only* Z_{res} gives 0.371 overall AUC on FakeAVCeleb against 0.922 for the standard head; being well below chance, the residual carries signal whose relation to the fake label reverses across domains (0.629 with the score flipped), not merely weak signal. Linear probes refine this (Appendix N, Table 13): the residual visual factor r_v remains task-decodable in the source domain (0.825), so it is not devoid of manipulation information, but it is far more domain-decodable (0.985), while the transferable visual task evidence concentrates in u_v (OOD FV-RA probe AUC 0.924).

This supports a more careful reading than "the residual is task-irrelevant". Excluding it trades away some source-domain predictive information in exchange for removing information that transfers poorly or in reverse—the division of labour the factorization is designed to produce. The probes establish that r captures domain-sensitive structure, but do not attribute it uniquely to codec, speaker identity, or recording conditions; identifying the dominant factors is left to future analysis.

Subspace probes across model families. The three-seed probe comparison is in Appendix I (Table 6, Figure 5). Its key finding matches: the visual-unique factor u_v achieves 0.924 OOD FV-RA probe AUC in 256 dimensions, higher than the full 1024-dim two-way representation (0.705), while domain probe AUC stays high across all subspaces (0.967–0.998)—routing rather than domain invariance.

6.7 Nuisance Partition Quality

The method relies on a source-side nuisance partition, so its sensitivity to that partition is a direct assumption of the design. We ran a controlled study in which *only* the nuisance-label definition changes and every other setting is held fixed, each row a three-seed mean (Appendix O, Table 14).

Two conclusions follow. Finer partitions do not help: 4-way and 8-way speaker hashes give 0.701 and 0.667 FV-RA against the binary hash's 0.838, so a weak binary partition is not merely sufficient but the best of those tested. Less informative partitions degrade without collapsing to chance: a random speaker-level regrouping gives 0.694 and random clip labels 0.663, both above the no-push-pull model (0.600) yet 0.14 and 0.18 below the paper's hash, so a supervised residual sink helps under every partition tested, but most of the gain depends on the speaker hash. TriRoute therefore needs no precise domain annotation—ours is annotation-free, requiring only an existing speaker identifier—but the magnitude of the gain depends on the partition definition; the hash was set once before any target evaluation and never tuned. We state this as an assumption and a limitation rather than claiming robustness to arbitrary nuisance labels.

Unsupervised scaling stress test. Unsupervised correspondence models trained on 120K and 250K clips, selected by the same source criterion and evaluated once on FAVC, reach 0.771 and 0.742 FV-RA (three seeds each) against 0.838 for supervised TriRoute (Appendix T), so more unsupervised data alone does not improve FV-RA monotonically; we draw only that conclusion.

7 Conclusion

We proposed TriRoute, a three-way factorized audio-visual deepfake detector that separates shared content, modality-unique manipulation evidence, and residual domain-sensitive signal, structurally preventing the prediction head from reading the shortcut-heavy branch. On AV1M→FAVC it reaches 0.838 FV-RA AUC using only source-side nuisance labels—well above AVH-Align/sup (0.519) and alignment controls on the same two-way representation (0.496, 0.521)—with independent source-only reruns at 0.824 ± 0.022 and the advantage carrying to LAV-DF (0.914 vs. 0.672). Head ablation, input masking, residual-only readout, and subspace probing converge on inference-time routing as the source of the gain: masking u_v drops FV-RA from 0.838 to 0.491, a residual-only head transfers at 0.371, and domain information stays decodable everywhere yet is displaced away from the prediction pathway. Structuring the decision pathway, rather than globally suppressing spurious signals, appears the more effective route to robustness under domain shift. Limitations are discussed in Appendix B.

Ethics Statement

This work studies detection of manipulated audio-visual media using three public research datasets (AV-Deepfake1M, FakeAVCeleb, LAV-DF) and one public lip-sync dataset (AVLips), all used under their released terms and solely for forensic evaluation. We do not generate new deepfakes and we release no synthesis capability. Detectors of this kind can be misused for surveillance or overconfident moderation, particularly when their uncertainty is not surfaced; Appendix A discusses this and the disclosure practices we recommend. Because the model is speech-anchored and evaluated on a limited set of identities and generators, it should not be deployed as a standalone authenticity verdict.

Reproducibility Statement

The protocol on which the cross-domain claim depends is stated in full in Section 5 and Appendix C: checkpoints are selected by AV1M source-validation AUC alone, and each target split is evaluated exactly once after the checkpoint is fixed. We release the source-only training, selection, and one-shot evaluation scripts together with configuration files, the frozen data splits, checkpoint identifiers, and the loader-level guard that blocks target-domain access, so the protocol can be reproduced and audited end-to-end. Section 6.2 reports an independent repetition of the three headline runs on that pipeline. Architecture, objectives, and hyperparameters are given in the implementation paragraph of Section 5 and Appendix C; the construction rules for the additional target evaluations are in Appendix E.

AI Use Statement

In this work, we used generative AI tools only for language polishing, including grammar checking and editing for clarity and concision. We have not used generative AI tools to generate data, design experiments or methodology, implement methods, formulate or refine hypotheses or claims, analyse or interpret results, or produce any of the reported numbers, tables, or figures. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the final content of this work.

References

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/45fbc6d3e05ebd93369ce542e8f2322d-Abstract.html.

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, 2022. doi: 10.1109/DICTA56598.2022.10034605. URL https://arxiv.org/abs/2204.06228.

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. URL https://arxiv.org/abs/2311.15308.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, volume 10117 of *Lecture Notes in Computer Science*, pp. 251–263. Springer, 2016. doi: 10.1007/978-3-319-54427-4_19.

Yuxuan Du, Zhendong Wang, Yuhao Luo, Caiyong Piao, Zhiyuan Yan, Hao Li, and Li Yuan. CAD: A general multimodal framework for video deepfake detection via cross-modal alignment and distillation. *arXiv preprint arXiv:2505.15233*, 2025. URL https://arxiv.org/abs/2505.15233.

Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10491–10503, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Feng_Self-Supervised_Video_Forensics_by_Audio-Visual_Anomaly_Detection_CVPR_2023_paper.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Maximilian Ilse, Jakub M. Tomczak, Christos Louizos, and Max Welling. DIVA: Domain invariant variational autoencoders. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pp. 322–348, 2020. URL https://proceedings.mlr.press/v121/ilse20a.html.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://arxiv.org/abs/2108.05080.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Learning_Not_to_Learn_Training_Deep_Neural_Networks_With_Biased_CVPR_2019_paper.html.

Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li, Feng Yu, Ning Zhang, Xiang Meng, and Weiqing Huang. SpeechForensics: Audio-visual speech representation learning for face forgery detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9c7900fac04a701cbed83256b76dbaa3-Abstract-Conference.html.

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, Ziyou Liang, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing DeepFakes. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a5a5b0ff87c59172a13342d428b1e033-Abstract-Conference.html.

Ye Liu, Zihan Ji, and Hongmin Cai. Plug-and-play feature causality decomposition for multimodal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/55123f38c9f4acf417335cff41be6e27-Abstract-Conference.html.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddc3427c5d77843c2253f1e799fe933-Abstract.html.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. AVFF: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27102–27112, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Oorloff_AVFF_Audio-Visual_Feature_Fusion_for_Video_Deepfake_Detection_CVPR_2024_paper.html.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2201.02184.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18815–18825, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Smeu_Circumventing_Shortcuts_in_Audio-visual_Deepfake_Detection_Datasets_with_Unsupervised_Learning_CVPR_2025_paper.html.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. *arXiv preprint arXiv:1607.01719*, 2016. URL https://arxiv.org/abs/1607.01719.

Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14800–14809, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhou_Joint_Audio-Visual_Deepfake_Detection_ICCV_2021_paper.html.

A Broader Impact

Cross-domain deepfake detection supports content authentication, media forensics, and platform integrity, but any detector can be misused for surveillance or overconfident moderation, especially where its uncertainty is not surfaced. We recommend reporting calibration, failure modes, and domain-shift sensitivity, and releasing models with clear usage restrictions and dataset documentation.

B Limitations

TriRoute is speech-anchored through AV-HuBERT, so clips with absent speech, minimal lip motion, or predominantly visual manipulations fall outside the current scope. A sweep that progressively zeroes the first 1–10 audio frames shows graceful degradation rather than collapse—FV-RA drops by at most 0.033 while RV-FA falls by roughly 0.09 (Appendix Q)—but it does not evaluate genuinely speech-free content, so we narrow our claims to speech-conditioned transfer. The category-level gain is concentrated in FV-RA because RV-FA and FV-FA are near ceiling. Our cross-domain evidence covers three targets in a single transfer direction (AV1M as source), only two of which carry native modality-category labels. We compare against AVH-Align/sup, SpeechForensics, and AVAD under a matched frozen protocol without retraining them on AV1M; SpeechForensics is clearly stronger on the FAVC lip-sync regime. The gain depends on the source-side nuisance partition: the fixed speaker hash gives 0.838 while the other definitions tested give 0.66–0.70 (Section 6.7), so the method requires an informative source-side partition, although not target labels. We have not run a branch-count ablation, so the three-role design is motivated rather than shown optimal, and the routing evidence remains mechanistic rather than a formal causal proof.

C Implementation Details (Extended)

Supervised training scale and epoch definition. The supervised main runs (three seeds) use fixed, duplicate-free AV1M manifests containing 8,000 training clips (1,976 real; 6,024 fake) and 2,000 validation clips (519 real; 1,481 fake), with no clip overlap between the manifests. Each training epoch is one complete pass through the 8,000-clip training subset in randomly shuffled order; there is no per-epoch subsampling or step truncation. With batch size 1 and no gradient accumulation, this gives exactly 8,000 optimizer steps per epoch. Each validation pass covers all 2,000 validation clips. Training runs for at most 100 epochs with patience-15 early stopping on source-validation AUC. The total number of optimizer steps and training-clip visits depends on the number of epochs actually completed by each seed; the number of unique training clips remains 8,000. Thus, a full training epoch refers to this subset, not the complete AV1M corpus. Checkpoint selection uses source-validation AUC; the selected checkpoint epoch need not coincide with the final training epoch.

Nuisance label derivation. Nuisance labels are derived entirely from the AV1M training set via a speaker-group hash ($d = \mathrm{hash}(\mathrm{speaker_id}) \bmod 2$), assigning all clips of a speaker to the same binary group; no target-domain data enters training, validation, or model selection. This is deliberately weaker than true target-domain supervision and should be understood as a source-side nuisance partition. Section 6.7 quantifies how the result depends on this choice.

Batch size and gradient accumulation. All models are trained with batch size 1 (one clip per gradient step). This choice is imposed by the variable sequence length of AV-HuBERT features: each clip produces a different number of frames T, so a fixed-shape batch tensor cannot be constructed without padding. Rather than pad-and-mask (which would require re-implementing the temporal alignment loss), we process one clip at a time. No gradient accumulation is used: the alignment loss $\mathcal{L}_{\mathrm{align}}$ operates within a single clip's temporal window and does not benefit from cross-clip accumulation. The resulting per-step gradient is therefore a single-sample stochastic estimate, which is noisier than a mini-batch gradient but consistent across all model variants in the ablation table. All baselines (DANN/GRL, CORAL, two-way, three-way no push-pull) use the same batch size 1 setting for a fair comparison.

Matched settings for DANN/CORAL controls. For fairness, the DANN/GRL and CORAL controls use matched optimization settings (Adam 10^{-3}, GRL α=1.0, CORAL weight λ=1.0) under the same seed and checkpoint-selection policy. We do not perform target-aware tuning sweeps on target-domain labels.

Compute. All supervised models are trained from pre-extracted frozen AV-HuBERT features; the reported detector training cost therefore excludes AV-HuBERT pretraining. Each training run uses a single GPU from the university HPC cluster, 4 CPU workers, and 32 GB host memory. The primary partition used for the main TRIROUTE runs and the DANN/GRL sweep is **NVIDIA A100 PCIe** (40 GB); some auxiliary runs use NVIDIA L40S (48 GB) depending on queue availability. A supervised TRIROUTE run on the 8,000-clip training subset with the above early-stopping policy takes approximately **8 hours**. The three main runs therefore require approximately 24 GPU-hours (1 GPU-day) in total for detector training.

Seed and checkpoint policy. The current supervised TRIROUTE main experiment uses three independent seeds, with checkpoints selected by source-validation AUC. Checkpoints and seeds are never selected using target-domain labels or target-domain evaluation results; the selection criterion is AV1M source-validation AUC and nothing else. Mechanism-level analyses (head ablation, masking, shuffling, probing, and the partition study) are reported as means over the same three seeds with sample standard deviations over seeds.

D INDEPENDENT REPETITION OF THE HEADLINE RUNS

Table 3 reports the independent repetition of the three headline runs summarised in Section 6.2. The repetition uses the same source-only pipeline as Table 1, in which target-domain access is blocked at the data-loader level so a target clip cannot be read even accidentally; the selection criterion remains AV1M source-validation AUC, and the frozen FakeAVCeleb test split is evaluated exactly once per run after the checkpoint is fixed.

Table 3: Independent repetition of the three headline runs (same seeds, trained from scratch) on the frozen FakeAVCeleb test split. Checkpoints are selected exclusively by AV1M source-validation AUC on a pipeline that blocks target-domain access at the loader level, and each is evaluated once. Mean $\pm$ sample standard deviation over the three runs.

FAVC test	Run 1	Run 2	Run 3	Mean $\pm$ SD
Overall AUC	0.9043	0.9176	0.9252	0.916 ± 0.0106
RV-FA	0.9976	0.9988	0.9979	0.9981 ± 0.0006
FV-RA	**0.8012**	**0.8264**	**0.8440**	$\mathbf{0.824 \pm 0.022}$
FV-FA	0.9931	0.9962	0.9951	0.9948 ± 0.0016

E ADDITIONAL TARGET CONSTRUCTION AND MATCHED BASELINE PROTOCOL

LAV-DF. LAV-DF (Cai et al., 2022) provides native modality-level manipulation annotations, so the FV-RA category can be read directly from the dataset labels; the full FV-RA subset has $N = 13{,}358$ clips, and Table 2 reports it. No LAV-DF clip is used for training, validation, checkpoint selection, or hyperparameter tuning; the checkpoints are the ones already fixed by AV1M source validation.

AVLips. AVLips (Liu et al., 2024) provides only binary real/fake labels, with $3{,}801$ real and $3{,}801$ fake clips. Audio fingerprinting matches every fake to a real source clip (mean similarity 0.998, minimum 0.991) and an audio-only probe scores 0.512 ± 0.011, so every fake carries real audio and the full set ($N = 7{,}602$) is an FV-RA-type target. Because the underlying annotations are binary, we describe it as FV-RA-type by audit rather than as a natively annotated benchmark, and report it as supporting evidence only.

Matched external baselines. SpeechForensics (Liang et al., 2024) and AVAD (Feng et al., 2023) are evaluated on the identical frozen splits using their publicly released pretrained models, with no retraining and no target-specific adaptation. Both are training-free methods pretrained on real speech and therefore do not use the AV1M supervised source protocol adopted by TRIROUTE; Table 2 reports them as complementary references rather than as protocol-matched supervised competitors. Their AV1M-side behaviour under the same frozen protocol is reported in Table 10.

F EXTERNAL BASELINE CATEGORY COMPARISON

The AV1M-trained AVH-Align/sup checkpoint from Smeu et al. (2025) is the most directly comparable published supervised baseline. As Table 4 shows, it exhibits the same category-level failure mode

as our two-way baseline: RV-FA and FV-FA are near ceiling, while FV-RA remains near chance (0.519). TRIROUTE's FV-RA gain therefore addresses the visual-fake transfer failure of the closest published supervised baseline, not only our internal ablations.

Table 4: Category-level external supervised baseline on FakeAVCeleb (AV1M → FAVC). AVH-Align/sup is the AV1M-trained supervised checkpoint from Smeu et al. (2025), evaluated with the same category protocol as Table 1. Values are AUC/AP.

Method	Overall	RV-FA	FV-RA	FV-FA
AVH-Align/sup (Smeu et al., 2025)	0.777/0.994	0.999/0.999	0.519/0.955	0.999/1.000
TRIROUTE (ours)	**0.922/0.998**	0.999/0.999	**0.838/0.989**	0.994/1.000

G HEAD-ABLATION DETAILS

Table 5 and Figure 2 quantify how strongly each model's FV-RA performance depends on its visual branches. For the two-way baseline, zeroing the visual task branch causes a negligible drop of -0.034, confirming that the trained detector barely uses visual evidence for FV-RA decisions. For three-way factorization without domain push-pull, masking visual branches causes a larger drop (-0.113), indicating partial visual usage. In TRIROUTE, mean FV-RA falls from 0.838 to 0.491 when u_v is masked and to 0.441 when both visual branches are masked. The order-of-magnitude gap (-0.034 vs. -0.347) is mechanism-level evidence that TRIROUTE has rerouted FV-RA decisions through the visual-unique factor. The per-seed Δ values are tightly clustered (SD 0.008–0.016), so the effect is not an artefact of averaging over seeds. Masking the audio branch of the two-way baseline instead drops in-domain AUC from 0.997 to 0.528, confirming audio dominance rather than visual capacity failure.

Table 5: Head-ablation analysis on FV-RA AUC. Each row zeros out one branch of the trained model and measures the FV-RA drop (ΔFV-RA). Larger drop indicates heavier routing dependence on that branch. All rows report the mean over three independent seeds on the checkpoints of Table 1; the baseline column therefore equals the main-table FV-RA. The last column is the sample SD of Δ over seeds.

Model	Ablation	Baseline FV-RA	Ablated FV-RA	$\Delta_{\text{FV-RA}}$	SD(Δ)
Two-way baseline	mask visual task branch	0.483	0.449	-0.034	0.016
Three-way, no domain push-pull	mask visual branches	0.600	0.487	-0.113	0.008
TRIROUTE (ours)	mask manipulation (u_v)	**0.838**	0.491	$\mathbf{-0.347}$	0.014
TRIROUTE (ours)	mask visual branches	**0.838**	0.441	$\mathbf{-0.397}$	0.016

H QUALITATIVE ROUTING AND PER-FRAME ROUTING MAPS

Figure 3 shows per-frame routing maps for representative FAVC FV-RA clips. For fake clips, the visual-unique factor $\|u_v^t\|$ shows sharp peaks aligned with manipulation events; for real clips, it stays low while the residual carries the bulk of the activation. The prediction head reads only u_v (and s_v), so the asymmetry in $\|u_v^t\|$ between the two clips drives the score. Of 3,007 FV-RA fake clips in the canonical FAVC test split, 1,048 satisfy the "two-way fails / TRIROUTE corrects" filter, and the largest score gaps cluster around 0.99. Additional success and failure cases are in Figures 6 and 7 (Appendix S).

I SUBSPACE PROBE ANALYSIS

To check that the structural separation concentrates transferable visual evidence into a specific component, we train light linear probes directly on individual subspaces and report cross-domain (OOD) FV-RA AUC and domain-classification AUC. A useful manipulation pathway should be predictive of FV-RA *and* relatively domain-blind. Table 6 reports the three-seed comparison across model families; Table 13 (Appendix N) reports the corresponding r_v/u_v probes and the residual-only readout.

The 256-dim u_v attains 0.924 OOD probe AUC, higher than both the 1024-dim fused two-way representation (0.705) and the 512-dim two-way visual factor (0.891)—higher OOD performance despite lower dimensionality. The domain-classification probe stays high on every subspace (0.967–0.998): domain identity is not erased but displaced away from the prediction-relevant subspace, consistent with the push-pull objective.

Table 6: Subspace probe analysis, mean ± SD over three independent seeds. Linear probes are trained on each subspace and evaluated for cross-domain FV-RA AUC and domain-classification AUC. The visual-unique factor u_v in TRIROUTE achieves the strongest cross-domain probe AUC despite having only 256 dimensions.

Model	Subspace (Dim.)	OOD probe AUC	Domain probe AUC
Two-way	task branch Z_c (1024)	0.705 ± 0.016	0.998 ± 0.001
Two-way	visual factor v_c (512)	0.891 ± 0.011	0.994 ± 0.002
Three-way, no push-pull	shared visual s_v (256)	0.814 ± 0.019	0.992 ± 0.002
Three-way, no push-pull	unique visual u_v (256)	0.825 ± 0.015	0.967 ± 0.007
TRIROUTE (ours)	shared visual s_v (256)	0.726 ± 0.023	0.989 ± 0.003
TRIROUTE (ours)	unique visual u_v (256)	$\mathbf{0.924 \pm 0.010}$	0.976 ± 0.005

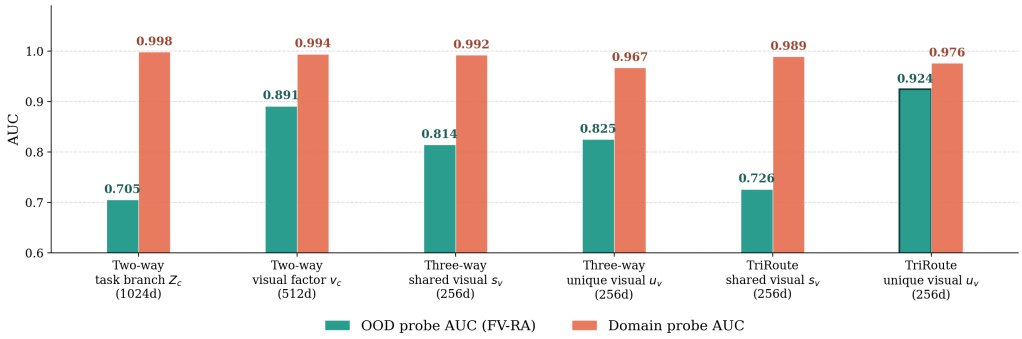

Figure 5: Subspace probe AUCs for OOD FV-RA prediction (teal) and domain prediction (orange) across the six subspaces in Table 6. The visual-unique factor u_v in TRIROUTE achieves the strongest cross-domain probe AUC (0.924) at only 256 dimensions, while the domain probe stays high across all subspaces (0.967–0.998)—routing rather than invariance.

J MULTI-SEED RESULTS AND PREPARED SWEEPS

Table 7 reports the three-seed mean ± std behind every row of Table 1. Std is the sample standard deviation over the three seeds (denominator $n - 1$).

Table 7: FAVC 70/30 AUC, three-seed mean ± std for every internal model of Table 1. FV-RA is the visual-routing diagnostic.

Model	Overall	RV-FA	**FV-RA**	FV-FA
Two-way baseline	0.760 ± 0.015	1.000 ± 0.000	$\mathbf{0.483 \pm 0.031}$	0.998 ± 0.001
Two-way + DANN/GRL control	0.767 ± 0.013	1.000 ± 0.000	$\mathbf{0.496 \pm 0.028}$	1.000 ± 0.000
Two-way + CORAL control	0.768 ± 0.020	1.000 ± 0.000	$\mathbf{0.521 \pm 0.035}$	0.980 ± 0.007
Three-way, no push-pull	0.814 ± 0.009	1.000 ± 0.000	$\mathbf{0.600 \pm 0.019}$	0.999 ± 0.002
TRIROUTE (ours)	0.922 ± 0.012	0.999 ± 0.001	$\mathbf{0.838 \pm 0.024}$	0.994 ± 0.002

Hyperparameter sensitivity. Table 8 reports a one-at-a-time sensitivity audit of the three domain-routing loss weights. Each row holds two weights fixed at their defaults (λ_{dom}=1.0, λ_{adv}=1.0, λ_{ortho}=0.1) and varies the third; all runs use three seeds and the same source-validation checkpointing rule with no target-domain labels. Two observations stand out. First, the default setting is at or near the optimum of each one-dimensional sweep, indicating that it was not adversarially tuned. Second, and more importantly, even the worst single configuration in the sweep (λ_{adv}=3.0, FV-RA 0.768) remains +0.254 above the best mean FV-RA of the DANN/GRL baseline across its own α-sweep (0.514, Table 9). The routing gain is therefore not explained by lucky default weight choices: it is present across the entire range tested.

DANN/GRL strength audit. The DANN baseline audit uses the same code path as the main DANN row and sweeps $\alpha \in \{0.1, 0.3, 0.5, 1.0, 3.0\}$ across three seeds. Table 9 shows that increasing or decreasing the GRL strength does not close the visual-routing gap: the best mean FV-RA AUC is 0.514 at $\alpha = 0.3$, and even the strongest single run in the sweep ($\alpha = 3.0$, FV-RA AUC 0.557) remains far below TRIROUTE's 0.838. Thus the weak DANN result is not explained by an unlucky default α.

Table 8: One-at-a-time loss-weight sensitivity on FAVC FV-RA AUC, mean ± SD over three seeds. Default values: λ_{dom}=1.0, λ_{adv}=1.0, λ_{ortho}=0.1. The default row is shaded; all other rows vary exactly one weight. The DANN/GRL best mean FV-RA is 0.514 (Table 9).

λ_{dom}	λ_{adv}	λ_{ortho}	**FV-RA AUC**	Δ vs. default
0.3	1.0	0.1	0.812 ± 0.027	-0.026
1.0	1.0	0.1	$\mathbf{0.838 \pm 0.024}$	0
3.0	1.0	0.1	0.791 ± 0.031	-0.047
1.0	0.3	0.1	0.806 ± 0.022	-0.032
1.0	3.0	0.1	0.768 ± 0.044	-0.070
1.0	1.0	0.03	0.829 ± 0.019	-0.009
1.0	1.0	0.3	0.830 ± 0.049	-0.008

Table 9: DANN/GRL strength sweep on FakeAVCeleb 70/30. Each row reports mean AUC over three seeds. Std. is the sample standard deviation over seeds for FV-RA AUC.

GRL α	Overall AUC	**FV-RA AUC**	FV-RA std.
0.1	0.756	0.475	0.018
0.3	**0.774**	**0.514**	0.026
0.5	0.758	0.479	0.015
1.0	0.767	0.496	0.028
3.0	0.772	0.509	0.048

K Comparison in AVH-Align Protocol

Table 10 replicates the structure of Table 2 in Smeu et al. (2025), reporting AUC and AP on the FakeAVCeleb 70/30 test split and AV-Deepfake1M validation set ($N = 2{,}000$), with and without leading-silence trimming. Baseline rows are reproduced from Smeu et al. (2025). Our rows report the source-side nuisance-label variant averaged over three independent seeds. The most direct comparison is the supervised AV1M-trained AVH-Align row; after leading-silence trimming, TriRoute remains clearly stronger on FAVC (86.6 vs. 70.8 AUC). The unsupervised VoxCeleb2-pretrained rows are included for context rather than as matched baselines. This table is also the source of the AV1M-side numbers referenced in Section 6.3: under the same frozen protocol, SpeechForensics and AVAD reach 68.8 and 54.3 overall AUC on AV1M, against 99.7 for TriRoute. Note that TriRoute trained directly on FAVC reaches near-ceiling in-domain performance (AUC 99.6, lower block of Table 10), indicating that the AV1M→FAVC gap is not a model-capacity limitation of TriRoute but a manifestation of cross-dataset shift.

Table 10: Protocol context: this table is not a headline leaderboard. Comparison with Smeu et al. (2025) Table 2 baselines on FakeAVCeleb (FAVC) 70/30 test split and AV-Deepfake1M (AV1M) validation set. Values are percentages. Trim:✓ denotes leading-silence-trimmed evaluation. Baseline rows are reproduced from Smeu et al. (2025); the headline supervised and unsupervised rows report the mean over three independent seeds. **Bold** = best, underline = second best within each training-data block. FAVC-trained rows show the reverse-direction transfer. Seed SDs of our rows are at most 1.6 points (AUC) and 1.1 points (AP). The post-hoc fusion row is in Appendix U.

				Metric: AUC				Metric: AP			
				FakeAVCeleb		AV-Deepfake1M		FakeAVCeleb		AV-Deepfake1M	
Method	Mod.	Train type	Train data	Trim:×	Trim:✓	Trim:×	Trim:✓	Trim:×	Trim:✓	Trim:×	Trim:✓
Unsupervised pretraining (reproduced from Smeu et al. (2025))											
AVAD (Feng et al., 2023)	AV	unsup.	LRS	84.5	84.7	54.3	54.3	99.5	99.5	76.3	76.3
SpeechForensics (Liang et al., 2024)	AV	unsup.	VoxCeleb2	**98.8**	**98.8**	68.8	68.2	**100.0**	**100.0**	83.7	83.5
AVH-Align (Smeu et al., 2025)	AV	unsup.	VoxCeleb2	94.6	94.6	**85.9**	**83.5**	99.8	99.8	**94.3**	**93.5**
TriRoute (unsup, ours)	AV	unsup.	VoxCeleb2	88.9	89.4	85.4	82.1	99.5	99.5	90.9	88.2
Supervised training on AV1M											
AVH-Align/sup (Smeu et al., 2025)	AV	sup.	AV1M	77.7	70.8	**100.0**	83.1	99.4	99.1	**100.0**	94.4
Two-way baseline (ours)	AV	sup.	AV1M	76.0	68.5	99.8	74.7	99.3	99.0	99.9	90.1
Three-way, no push-pull (ours)	AV	sup.	AV1M	81.4	76.3	99.9	84.6	99.6	99.4	**100.0**	**94.7**
TriRoute (ours)	AV	sup.	AV1M	**92.2**	**86.6**	99.7	**84.8**	**99.8**	**99.6**	99.9	94.1
Reverse-direction transfer (trained on FAVC, evaluated on AV1M)											
AVH-Align/sup (Smeu et al., 2025)	AV	sup.	FAVC	99.2	99.2	69.0	63.6	**100.0**	**100.0**	84.1	**82.2**
Two-way baseline (ours)	AV	sup.	FAVC	99.5	99.5	67.0	63.4	**100.0**	**100.0**	81.2	79.8
Three-way, no push-pull (ours)	AV	sup.	FAVC	**99.7**	**99.7**	68.3	65.1	**100.0**	**100.0**	82.2	80.4
TriRoute (ours)	AV	sup.	FAVC	99.6	99.6	**74.7**	**67.7**	**100.0**	**100.0**	**86.4**	**82.2**

L Input Masking Details

Table 11 reports the full on/off masking intervention summarised in Section 6.5.

Table 11: FV-RA AUC under full input masking on the FakeAVCeleb FV-RA category. Each row reports the mean over three independent seeds. Δ = masked AUC $-$ baseline AUC (mean $\pm$ SD over seeds). A positive Δ_{audio} indicates that removing the audio signal *helps* FV-RA detection, i.e., the real-audio stream was acting as a distractor. Baselines equal the Table 1 FV-RA values. Relative visual sensitivity is $|\Delta_{\text{visual}}|$ divided by the two-way value.

Model	Baseline	Δ_{audio} (full)	Δ_{visual} (full)	Visual sensitivity ratio
Two-way baseline	0.483	$+0.095 \pm 0.009$	-0.038 ± 0.010	$1.0\times$
Three-way, no push-pull	0.600	$+0.053 \pm 0.006$	-0.132 ± 0.003	$3.5\times$
TriRoute (ours)	**0.838**	$+0.024 \pm 0.002$	$\mathbf{-0.407 \pm 0.009}$	$\mathbf{10.7}\times$

M Progressive Masking Sweep

Table 12 and Figure 4 give the full progressive masking sweep summarised in Section 6.5, with mask probability p applied separately to the visual and audio channels.

Table 12: Progressive input masking on FAVC FV-RA AUC. Top block: visual masking ($\Delta < 0$ indicates visual routing). Bottom block: audio masking ($\Delta > 0$ indicates audio was a distractor). Numbers correspond to Figure 4; the p=0.0 column equals the Table 1 FV-RA values. Mean over three seeds; per-cell SDs are 0.014–0.034.

Model	p=0.0	p=0.1	p=0.25	p=0.5	p=0.75	p=1.0	$\Delta(0\to1)$
Visual masking							
Two-way baseline	0.483	0.485	0.489	0.486	0.474	0.445	-0.038
Three-way, no push-pull	0.600	0.594	0.584	0.571	0.526	0.468	-0.132
TriRoute (ours)	0.838	0.817	0.787	0.741	0.658	0.431	$\mathbf{-0.407}$
Audio masking							
Two-way baseline	0.483	0.513	0.529	0.541	0.569	0.578	$\mathbf{+0.095}$
Three-way, no push-pull	0.600	0.599	0.602	0.618	0.635	0.653	$+0.053$
TriRoute (ours)	0.838	0.841	0.849	0.850	0.836	0.862	$+0.024$

N Residual Branch Analysis

Table 13 reports the probe and residual-only-readout values discussed in Section 6.6, computed on the same three checkpoints as Table 1.

Table 13: What the excluded branch carries, mean $\pm$ SD over three seeds. Task decodability is measured in the source domain; OOD task decodability is cross-domain FV-RA probe AUC; domain decodability is source-partition probe AUC. The residual-only head is a full prediction head trained on Z_{res} alone and evaluated on FakeAVCeleb.

Subspace / readout	Source task AUC	OOD task AUC	Domain AUC
Residual visual r_v (256)	0.825 ± 0.016	—	$\mathbf{0.985 \pm 0.006}$
Unique visual u_v (256)	—	$\mathbf{0.924 \pm 0.010}$	0.976 ± 0.005
Residual-only head (Z_{res})	FAVC overall AUC 0.371 ± 0.017		
Standard head (Z_{task})	FAVC overall AUC $\mathbf{0.922 \pm 0.012}$		

O Nuisance Partition Quality

Table 14 reports the controlled nuisance-label study discussed in Section 6.7. Only the nuisance-label definition varies across rows; every other training setting is held fixed, and each row is a three-seed mean. The speaker-group pseudo-label is not a target-domain label: it partitions source clips by speaker identity and does not encode codec differences, acquisition conditions, or any target-specific artifact. It is nevertheless the best of the partitions tested. With random clip labels FV-RA falls to 0.663 and with a random speaker-level regrouping to 0.694, both above the three-way model

without push-pull (0.600) and 0.18 and 0.14 below the paper's hash, so the supervised residual sink contributes on its own but most of the gain depends on the speaker hash; the partition need not be fine, and finer speaker partitions are worse than the binary one.

Table 14: Nuisance-label definition study on FAVC, mean ± SD over three seeds. All rows use the fixed source-only protocol; only the nuisance-label definition varies.

Label definition	**FV-RA**	Overall	Notes
$\mathrm{hash}(\mathrm{speaker_id}) \bmod 2$ (paper)	$\mathbf{0.838 \pm 0.024}$	0.922 ± 0.012	Paper setting
4-way speaker hash	0.701 ± 0.017	0.858 ± 0.007	Finer partition
8-way speaker hash	0.667 ± 0.015	0.843 ± 0.009	Finer partition
Random speaker regrouping	0.694 ± 0.021	0.853 ± 0.010	Misaligned partition
Random clip labels	0.663 ± 0.029	0.840 ± 0.015	Uninformative

P Supplementary External Support on AVLips

Table 15 reports the full AVLips set (Liu et al., 2024) for the three internal model families with AP alongside AUC; the AUC column equals the AVLips row of Table 2. AVLips provides only binary real/fake labels, although the audit in Appendix E shows every fake carries real audio. AVLips is a meaningful distribution shift relative to AV-Deepfake1M: it uses a different set of source identities, capture conditions, and synthesis pipelines. The same ordering is preserved: the two-way baseline reaches 0.533 AUC, three-way factorization without domain push-pull reaches 0.588, and TriRoute reaches 0.789; all values are means over three independent seeds. We treat the full-set AVLips result as supportive rather than central evidence, because it cannot replicate the FV-RA routing diagnostic without modality-category labels.

Table 15: AVLips full-set cross-dataset evaluation ($N = 7{,}602$). Values are mean AUC/AP averaged over three independent seeds. AVLips provides only binary real/fake labels; per-modality-category breakdown is not available. Reported as supplementary directional evidence.

Model	AUC	AP
Two-way baseline	0.533	0.598
Three-way, no domain push-pull	0.588	0.648
TriRoute (ours)	**0.789**	**0.784**

Perturbation companion on AVLips. To test whether the visual-routing pattern observed on FAVC also holds on this second target dataset, we run the same single-channel masking intervention on AVLips: full visual masking ($p = 1.0$) and full audio masking ($p = 1.0$), reusing the masking definition from Table 11. Table 16 reports the resulting changes in AUC. The pattern matches FAVC: the two-way baseline is essentially insensitive to either masking ($\Delta_{\mathrm{vis}} = +0.014$, $\Delta_{\mathrm{aud}} = -0.035$), consistent with a model that is near chance on AVLips and not strongly routing through either modality. TriRoute, in contrast, shows a strongly asymmetric profile — full visual masking lowers AVLips AUC from 0.789 to 0.475 ($\Delta = -0.314$), while full audio masking is approximately neutral ($\Delta = -0.003$). Three-way factorization without push-pull again sits in between ($\Delta_{\mathrm{vis}} = -0.048$, $\Delta_{\mathrm{aud}} = +0.048$).

Table 16: AVLips masking intervention, mean over three seeds; baseline AUC equals Table 15. SDs of Δ are 0.007–0.024. Δ_{vis} is the change after applying full visual masking; Δ_{aud} is the change after applying full audio masking (both $p = 1.0$). The asymmetric profile replicates the FAVC FV-RA pattern: only TriRoute shows a large $|\Delta_{\mathrm{vis}}|$.

Model	Baseline AUC	Δ_{vis} (full)	Δ_{aud} (full)
Two-way baseline	0.533	$+0.014$	-0.035
Three-way, no push-pull	0.588	-0.048	$+0.048$
TriRoute (ours)	**0.789**	$\mathbf{-0.314}$	-0.003

Q Leading-Silence Robustness Sweep

Because TriRoute is speech-anchored, we test how it degrades as speech evidence is withdrawn, by progressively zeroing the first 1–10 audio frames (mean over the three seeds). FV-RA is comparatively

stable across the sweep, dropping by at most 0.033 from its unperturbed value, while RV-FA falls by roughly 0.09, from 0.999 to approximately 0.91. The asymmetry is the expected one: RV-FA is the category whose label is carried by the audio stream, so removing audio evidence should hurt it, whereas FV-RA decisions run through the visual route and are largely unaffected. These results indicate graceful degradation under moderate loss of speech information. They do not, however, evaluate genuinely speech-free content, minimal-lip-motion videos, or predominantly visual manipulations, which we identify as important directions for future evaluation. This sweep is distinct from the leading-silence *trimming* control in Appendix V, which removes a dataset artifact rather than testing robustness to audio loss. The full per-trim table is in Appendix R.

R Full Leading-Silence Trim Table

Table 20 (Appendix V) reports trimming-induced changes (Δ) only. Table 17 below reports the full pre-trim and post-trim AUCs that produce those Δ values, including all FAVC categories. The asymmetry highlighted in the main text is visible directly here: the audio-fake categories (RV-FA, FV-FA) lose substantial AUC after trimming, while FV-RA stays within 0.008 AUC of its untrimmed value across all three models.

Table 17: Full leading-silence trim AUCs on FakeAVCeleb categories, mean over three independent seeds. "Untrimmed" is the reference condition reported in the main results; "Trimmed" removes raw-waveform leading silence prior to AV-HuBERT feature extraction. $\Delta = \mathrm{Trimmed} - \mathrm{Untrimmed}$. FV-RA is highlighted; it is the visual-routing diagnostic and stays flat across all three models, while RV-FA and FV-FA (audio-fake categories) drop substantially.

Model	Condition	Overall	RV-FA	**FV-RA**	FV-FA
Two-way baseline	Untrimmed AUC	0.760	1.000	0.483	0.998
Two-way baseline	Trimmed AUC	0.685	0.962	0.479	0.859
Two-way baseline	Δ	-0.074	-0.038	$\mathbf{-0.004}$	-0.139
Three-way, no push-pull	Untrimmed AUC	0.814	1.000	0.600	0.999
Three-way, no push-pull	Trimmed AUC	0.763	0.950	0.592	0.909
Three-way, no push-pull	Δ	-0.051	-0.050	$\mathbf{-0.008}$	-0.090
TriRoute (ours)	Untrimmed AUC	0.922	0.999	0.838	0.994
TriRoute (ours)	Trimmed AUC	0.866	0.810	0.833	0.898
TriRoute (ours)	Δ	-0.056	-0.189	$\mathbf{-0.005}$	-0.096

S Additional Qualitative Routing Examples

Figure 6 reports additional FV-RA test clips drawn from the same "two-way fails / TriRoute corrects" pool described in Section 6.4. The pattern is consistent across cases: the audio waveform appears as ordinary speech, the lip-region frames carry the manipulation evidence, the two-way baseline assigns very low fake probabilities, and TriRoute assigns high fake probabilities. Figure 7 complements these successes with FV-RA cases where TriRoute is still confidently wrong. Together, the success and failure examples illustrate both the audio-shortcut failure mode that motivates the three-way factorization and the remaining limits of the current routing mechanism, complementing the aggregate evidence in Tables 5, 11, and 12.

T Unsupervised Scaling Stress Test

We tested whether simply increasing the amount of unsupervised training data recovers the FV-RA routing behavior seen in supervised TriRoute.

Relationship to supervised TriRoute. The unsupervised variant shares the frozen AV-HuBERT front end and the shared-factor alignment objective, but it is *not* the three-way supervised routing model. It is trained on real clips with a temporal audio-visual correspondence objective and scores clips by correspondence anomaly; it has no detection BCE ($\mathcal{L}_{\mathrm{fake}}$), no residual domain-prediction term ($\mathcal{L}_{\mathrm{dom}}$), and no task adversarial term ($\mathcal{L}_{\mathrm{adv}}$), because none of these is available without manipulation labels or a supervised nuisance partition. The experiment therefore isolates one question—does more unsupervised data alone recover cross-domain FV-RA?—and is not a variant of the routing objective.

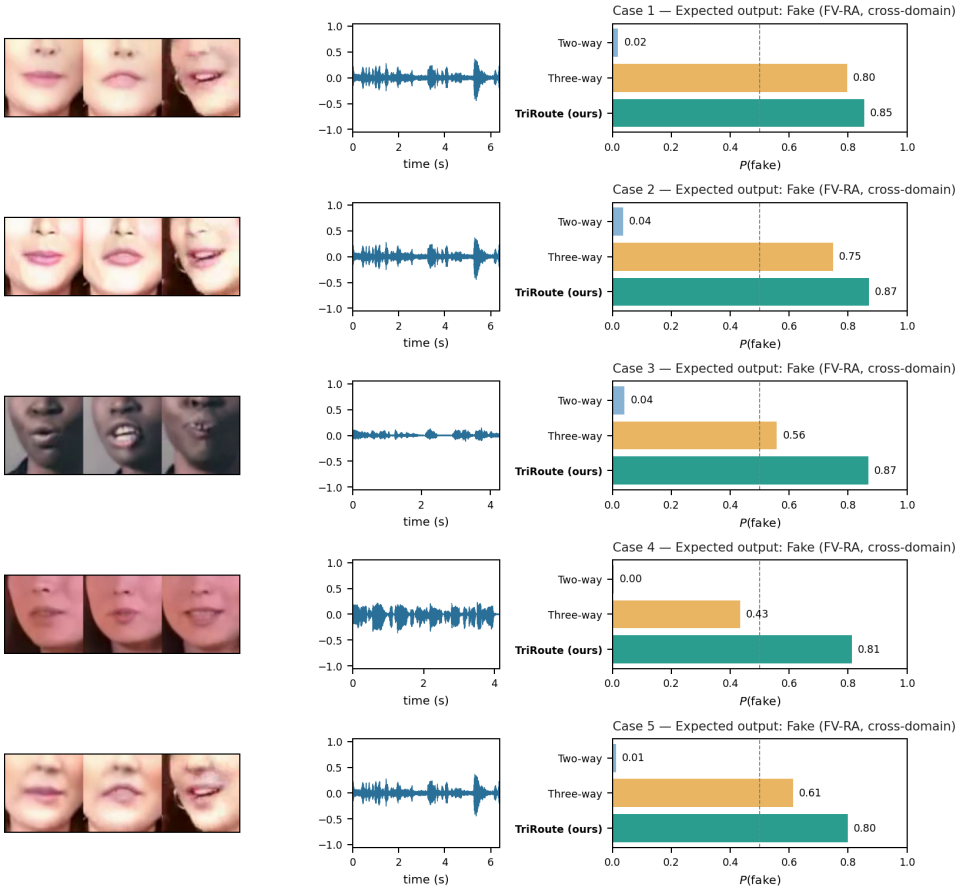

Figure 6: Additional FV-RA success cases where TRIROUTE corrects the two-way baseline's audio-shortcut failure. For each row, *left:* three lip-region frames; *middle:* audio waveform; *right:* model fake-probabilities. The two-way baseline assigns low fake probability, while TRIROUTE assigns high fake probability.

Result. Table 18 reports FAVC AUC at the source-validation-selected epoch across two unsupervised training scales (120K and 250K clips). Both checkpoints are selected exclusively by the in-domain source-validation criterion, and FakeAVCeleb is evaluated once after selection; no FakeAVCeleb score participates in choosing an epoch. FV-RA reaches 0.771 ± 0.029 at 120K and 0.742 ± 0.034 at 250K (three seeds each), against 0.838 for supervised TRIROUTE. The supported conclusion is therefore narrow: increasing unsupervised data alone does not improve cross-domain FV-RA monotonically. This non-monotonicity is consistent with the paper's position that the transfer benefit studied here stems from routing structure rather than data volume, but it does not establish that unsupervised scaling could not reach comparable performance with more data or a different objective.

Table 18: Unsupervised scaling stress test on FakeAVCeleb, mean $\pm$ SD over three seeds. Models are trained unsupervised at two data scales; the epoch is chosen by the AV1M source-validation criterion alone, and FakeAVCeleb is then evaluated once. Scaling from 120K to 250K clips lowers rather than raises FV-RA. The supervised TRIROUTE row repeats the three-seed means of Table 1.

Training scale	FAVC overall AUC	FAVC FV-RA AUC
120K clips	0.889 ± 0.014	0.771 ± 0.029
250K clips	0.876 ± 0.015	0.742 ± 0.034
TRIROUTE (supervised, ours)	0.922 ± 0.012	0.838 ± 0.024

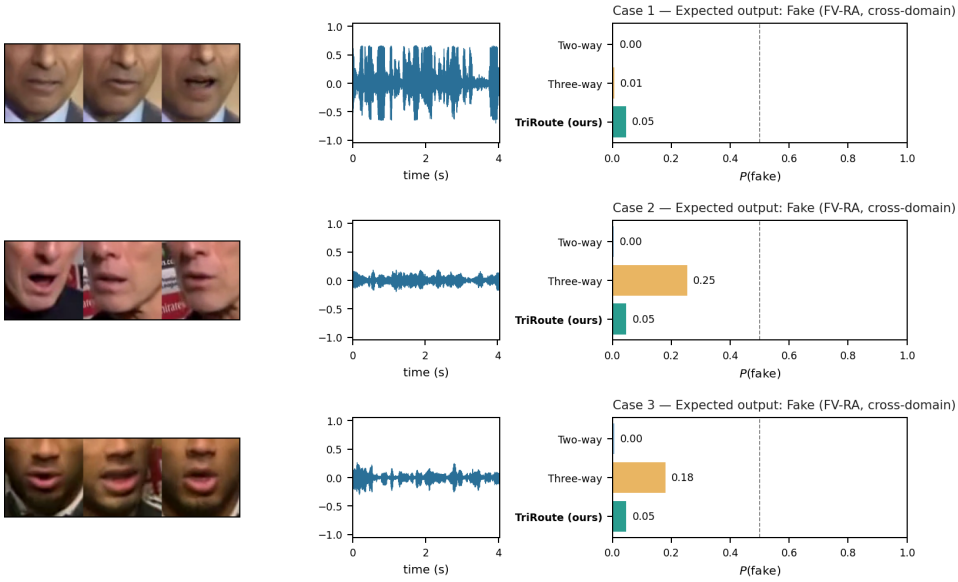

Figure 7: FV-RA failure cases where TRIROUTE remains confidently wrong. For each row, *left:* three lip-region frames; *middle:* audio waveform; *right:* model fake-probabilities. These cases show that explicit routing substantially reduces the audio-shortcut failure mode but does not eliminate all difficult FV-RA errors.

U POST-HOC FUSION RESULTS

Table 19 reports a post-hoc fusion result that combines the supervised TRIROUTE detector with an unsupervised kNN anomaly score. It is reported for completeness and is not part of the headline comparison.

Table 19: Post-hoc fusion: supervised TRIROUTE + unsupervised kNN anomaly score, on FakeAVCeleb (FAVC) 70/30 test split and AV-Deepfake1M (AV1M) validation set. Values are percentages. Reported for completeness only; not a headline result.

				Metric: AUC				Metric: AP			
				FakeAVCeleb		AV-Deepfake1M		FakeAVCeleb		AV-Deepfake1M	
Method	Mod.	Train type	Train data	Trim:×	Trim:✓	Trim:×	Trim:✓	Trim:×	Trim:✓	Trim:×	Trim:✓
TRIROUTE + source-gated kNN	AV	post-hoc	AV1M	95.7	92.8	99.6	85.4	99.9	99.8	99.7	93.7

V CORRECTED ARTIFACT CONTROL

A potential confound in AV1M-trained models is leading silence: fabricated clips sometimes begin with audio-free frames that are absent in authentic samples (Smeu et al., 2025). We therefore evaluate a leading-silence-trimmed version that removes raw-waveform leading silence prior to AV-HuBERT feature extraction. Table 20 reports the trimming-induced change $\Delta = \mathrm{AUC}_{\mathrm{trimmed}} - \mathrm{AUC}_{\mathrm{untrimmed}}$ per category. The pattern is a stark asymmetry: RV-FA and FV-FA, the two audio-fake categories, drop by tens of AUC points after trimming, while FV-RA stays flat across every model ($|\Delta_{\text{FV-RA}}| \leq 0.008$). If the FV-RA gain were a leading-silence artifact, trimming would erase it. That it does not rules out the audio-shortcut explanation for the visual-routing gain. The trimming-induced drops are instead concentrated where they should be, in the audio-fake categories, indicating that the artifact mainly inflates audio-driven scoring. The full pre-/post-trim AUCs that produce these Δ values are reported in Table 17.

Table 20: Trimming-induced change in AUC ($\Delta = \text{trimmed} - \text{untrimmed}$) on the FAVC categories, mean over three independent seeds. Audio-fake categories (RV-FA, FV-FA) drop substantially under trimming; FV-RA stays flat ($|\Delta| \leq 0.008$) across all three models. Seed SDs of $\Delta_{\text{FV-RA}}$ are at most 0.009. Pre-/post-trim raw AUCs are in Table 17.

Model	Δ_{Overall}	$\Delta_{\text{RV-FA}}$	$\Delta_{\text{FV-RA}}$ (key)	$\Delta_{\text{FV-FA}}$
Two-way baseline	-0.074	-0.038	$\mathbf{-0.004}$	-0.139
Three-way, no push-pull	-0.051	-0.050	$\mathbf{-0.008}$	-0.090
TRIROUTE (ours)	-0.056	-0.189	$\mathbf{-0.005}$	-0.096