BLACK-BOX ATTRIBUTION UNDER DOMAIN SHIFT: WHEN ARE GENERATIVE MECHANISMS IDENTIFIABLE?

Anonymous authors
Paper under double-blind review

ABSTRACT

We study when a black-box behavioral response that separates generators on a source domain remains valid attribution evidence after domain shift. We distinguish source recognition, population identifiability over a declared nuisance family, and finite-development certification. Response-family overlap gives a global testing obstruction. In a structured common-location model, finitely observed domains cannot certify unseen-domain attribution without a coverage assumption; under declared linear coverage, only the nuisance-orthogonal mechanism contrast survives, with an exact binary Gaussian benchmark and a finite-development recovery bound. In a controlled training-parameterization stress test, $\epsilon/x_0/v$ models are 9/9 separable within MNIST and Fashion-MNIST but 3/9 under frozen transfer. A nuisance projection reaches 17/18 on the development domains yet falls to 4/9 on held-out KMNIST. Deployed diffusion, Flow Matching, and consistency pipelines remain direction-dependent even with a frozen S^2-CLIP representation (8/9 versus 5/9). Synthetic checks recover the predicted Gaussian risk and coverage-rank behavior. Our claim is methodological: attribution validity is relative to a declared nuisance family, and finite source domains do not certify unseen coverage for free.

1 INTRODUCTION

Generative models leave behavioral traces that are used for model attribution, fingerprinting, provenance, and forensic analysis (Yu et al., 2019; 2021; Nie et al., 2023; Song et al., 2024; Song & Itti, 2025). Recent fingerprinting work increasingly evaluates transfer beyond the source setting, which makes the scope of an attribution claim part of the problem. A source-domain classifier asks whether observed generators are separable. Attribution asks whether the label remains meaningful when factors unrelated to the label change.

Our first experiment is intentionally narrow: it is a controlled stress test of *training-parameterization traces*, not a claim that ϵ-, x_0-, and v-prediction are functionally distinct deployed mechanisms. Active noise- and step-response curves separate these three training histories at the generator level within MNIST and within Fashion-MNIST: 9/9 in each domain. With all preprocessing frozen on the source domain, the same representation gives 3/9 in both transfer directions. The errors are systematic: one direction collapses all generators to ϵ, the reverse to x_0. Algebraically rewriting the output of a single trained ϵ model into x_0 and v forms under the same sampler leaves the measured response unchanged, ruling out the exposed output interface as the explanation. Checkpoint controls show that the source-domain trace also varies with optimization state.

Removing observed dataset information through a fitted nuisance projection does not suffice. A rank-2 projection estimated from MNIST and Fashion-MNIST raises training-trace accuracy to 17/18 while reducing dataset predictability to roughly chance. Once frozen, however, it gives 4/9 on KMNIST. The three KMNIST centroids move in strongly aligned directions that were not represented by the development-domain nuisance span. A separate study uses functionally distinct deployed pipelines—standard diffusion, Flow Matching (Lipman et al., 2023), and consistency/distillation (Song et al., 2023). Transfer remains direction-dependent. In the original low-order response space, same-pipeline domain drift exceeds the nearest source-pipeline separation in every direction–probe–pipeline condition. A frozen S^2-CLIP control on the same 18 generators improves several transfers,

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

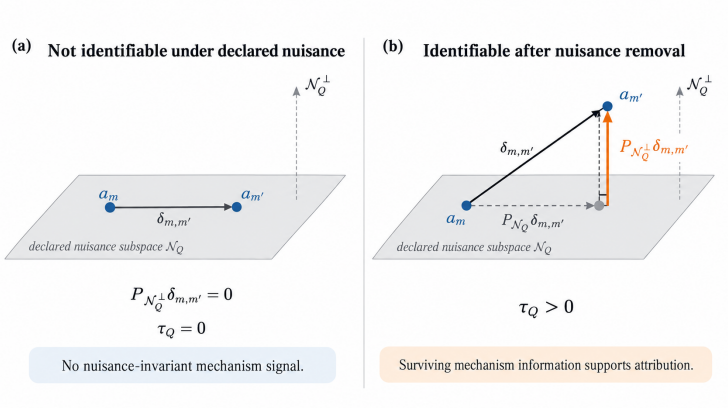

Figure 1: **Only nuisance-orthogonal mechanism variation supports attribution in the structured linear model.** The contrast decomposes as $\delta_{m,m'} = P_{\mathcal{N}_Q}\delta_{m,m'} + P_{\mathcal{N}_Q^\perp}\delta_{m,m'}$. If the orthogonal component vanishes (left), nuisance can reproduce the entire contrast and $\tau_Q = 0$; if it is nonzero (right), mechanism information survives nuisance removal. Here τ_Q is the minimum nuisance-orthogonal pairwise separation. The global response-orbit theory does not require nuisance variation to be linear.

but combined nearest-centroid accuracy is 8/9 for MNIST→Fashion-MNIST and 5/9 in reverse. This rules out the initial low-order representation as a sufficient explanation for the observed instability, without implying that every attribution representation must fail.

These failures point to a more basic question than whether a fingerprint classifier transfers:

> *When does a behavioral difference remain attributable to the generative mechanism after the domain changes, and when can an unseen nuisance make that distinction fundamentally ambiguous?*

The answer must be relative to what is allowed to change. If the nuisance family is unrestricted, a future domain can imitate mechanism variation. If it is too narrow, a source-domain benchmark can certify a claim that fails outside the benchmark. We therefore make the nuisance family part of the statistical problem.

Let m denote a generative mechanism, d a domain, and Q a controlled black-box probe protocol. The response distribution is $P^Q_{m,d}$. For a declared admissible domain family $\mathcal{D}$, mechanism m induces the response orbit

$$\mathcal{O}^Q_m(\mathcal{D}) = \{P^Q_{m,d} : d \in \mathcal{D}\}. \tag{1}$$

Recognition on one domain concerns isolated points on these orbits. Attribution over $\mathcal{D}$ requires the orbits to remain distinguishable. This leads to three levels of theory. First, response-family overlap yields a global information-theoretic obstruction. Second, finite development domains cannot certify unseen-domain attribution without a coverage assumption: the same observed data can admit incompatible nuisance completions. Third, under structured nuisance variation, the mechanism information that survives is exactly the component transverse to nuisance. In a Gaussian model that component determines the exact minimax attribution risk; with finitely many development domains it can be recovered only if the observed domain contrasts sufficiently excite the nuisance space.

For the schematic in Figure 1, a_m denotes the reference response mean of label m, $\delta_{m,m'} := a_{m'} - a_m$ is the contrast from m to m', and $\mathcal{N}_Q$ is the nuisance subspace associated with the fixed probe/representation Q. We reserve $P^Q_{m,d}$ for response laws; when the subscript is a subspace, $P_{\mathcal{S}}$ denotes its orthogonal projector. Throughout the structured theory, Q is fixed and we write $\mathcal{N}$ for $\mathcal{N}_Q$.

Our contributions are threefold. First, we separate source recognition, population identifiability, and finite-development certification, and formalize the last two relative to a declared nuisance family. Second, we prove a source-only coverage obstruction and derive conditional positive results under

structured coverage: exact linear transversality, a sharp binary Gaussian benchmark, and finite-development nuisance recovery. Third, we evaluate these distinctions with frozen stress tests that retain negative results, use the trained generator as the replication unit, and explicitly separate empirical non-transfer from information-theoretic impossibility.

2 Related Work and Positioning

Generative fingerprints and attribution. GAN fingerprints identify source models from generated images (Yu et al., 2019); artificial and latent fingerprints add or recover model-specific traces for attribution (Yu et al., 2021; Nie et al., 2023). ManiFPT formalizes artifacts and fingerprints and studies attribution across diverse generative processes (Song et al., 2024); Riemannian-geometric fingerprints report generalization across unseen datasets, model types, and modalities (Song & Itti, 2025). These works establish that useful fingerprint signal exists. Our question is different: what statistical conditions make such signal valid mechanism evidence over a declared nuisance family?

Domain shift and inverse identification. Domain adaptation bounds target error for a fixed prediction problem using source error and a discrepancy between source and target distributions (Ben-David et al., 2010); invariant and domain-generalization methods seek predictors that transfer across environments (Arjovsky et al., 2019; Gulrajani & Lopez-Paz, 2021). Our object is different: the item being classified is itself a trained generator, and changing the domain can change both the fitted object and its black-box response. We therefore study inverse identification of a latent training/deployment label, rather than target prediction by a fixed classifier. This distinction is also why a domain-discrepancy bound does not by itself certify mechanism identifiability.

Attribution limits and nuisance. Cebrian et al. (Cebrian et al., 2026) study LLM-output attribution through identification in the limit and probabilistic overlap, showing that source models can be non-identifiable even with extensive output evidence. Their ambiguity is driven by overlap within a model hypothesis class. Our ambiguity is domain-indexed: the candidate attribution labels are fixed, but each label induces a family of black-box responses as the nuisance domain varies. This yields a separate certification question—whether finitely observed development domains cover the nuisance directions needed for an unseen-domain claim.

Classical identification and invariant statistics study observable equivalence in the presence of nuisance (Rothenberg, 1971; Eaton, 1989). The mathematical tools below—testing bounds, invariant projection, Gaussian testing, and subspace perturbation—are classical; our contribution is their use to separate source recognition, population identifiability over a declared nuisance family, and finite-development certification for generative attribution. Davis–Kahan and random-matrix tools enter only after this attribution-validity problem is specified (Davis & Kahan, 1970; Vershynin, 2018).

3 Problem Formulation

Let $\mathcal{M} = \{1, \dots, M\}$ denote candidate attribution labels. Their ontology is experiment-specific. In Study I, the label records the *historical training parameterization* (ϵ, x_0, or v) under a common diffusion framework and sampler. Because these parameterizations are analytically related, Study I is a training-trace stress test, not a claim that they are functionally distinct deployed mechanisms. In Study III, the label denotes a deployed generation pipeline (standard diffusion, Flow Matching, or consistency/distillation); native samplers are then part of the operational mechanism. We treat the two studies as separate attribution objects. Appendix B summarizes the attribution objects and claims.

A domain $d \in \mathcal{D}$ specifies an admissible data environment. The family $\mathcal{D}$ is part of the claim: it may be a finite declared collection, a bounded distribution family, or a structured nuisance model. Attribution claims must therefore specify the nuisance family over which they are intended to hold.

A controllable black-box probe protocol Q specifies allowed interventions (here noise rescaling or sampling-budget perturbations) and a measurable feature map ϕ_Q applied to returned samples. The theoretical observation unit is one independent *probe-batch statistic*: for a probe batch $X_{1:B}$ of size B returned under Q, we write

$$Z = \phi_Q(X_{1:B}) \sim P^Q_{m,d}. \tag{2}$$

Repeated independent probe batches give $Z_1, \ldots, Z_n$. This distinction matters because the empirical signatures below are generator-level summaries of batches, not collections of independent images. The interface exposes neither model weights nor training logs. Our experiments evaluate cross-domain attribution with aggregated generator signatures; they do not estimate the finite probe-batch rate in Theorem 4.

The response orbit of mechanism m is Eq. (1). For an estimator $\widehat{m}$, define worst-case attribution risk

$$R_Q(\widehat{m}; \mathcal{D}) = \sup_{m \in \mathcal{M},\, d \in \mathcal{D}} P^Q_{m,d}(\widehat{m}(Z) \neq m), \qquad R^*_Q(\mathcal{D}) = \inf_{\widehat{m}} R_Q(\widehat{m}; \mathcal{D}). \tag{3}$$

A source-domain confusion matrix only establishes separation at sampled orbit points. Uniform attribution is a statement about the response families over $\mathcal{D}$.

Empirical and theoretical claims are kept separate. The experiments establish frozen non-transfer and measure the geometry of observed response features. They do not prove that the full black-box response distributions overlap. Information-theoretic impossibility is invoked only under the conditions of the theorems below.

4 Nuisance-Relative Identifiability Theory

4.1 A global response-family boundary

Define the closest pairwise orbit distance in total variation,

$$\Delta^Q_{\mathrm{TV}}(\mathcal{D}) = \min_{m \neq m'} \inf_{d, d' \in \mathcal{D}} \mathrm{TV}(P^Q_{m,d}, P^Q_{m',d'}). \tag{4}$$

Proposition 1 (Global response-orbit testing obstruction)**.** *For any probe protocol Q,*

$$R^*_Q(\mathcal{D}) \geq \frac{1 - \Delta^Q_{\mathrm{TV}}(\mathcal{D})}{2}. \tag{5}$$

If two mechanism response orbits intersect, uniform attribution over $\mathcal{D}$ has minimax error at least $1/2$ on that pair.

Proposition 1 is deliberately general. It says when information is absent from the declared black-box experiment, but it does not say whether source data are sufficient to certify that the relevant response families are separated. That is a distinct problem.

4.2 Finite development domains cannot certify unseen coverage for free

We now consider a structured mean model. Let $\mathcal{D}_{\mathrm{dev}} = \{d_0, \ldots, d_K\}$ be the development domains and write

$$\mu_{m,d} = a_m + b_{m,d}, \qquad b_{m,d} \in \mathcal{N}, \tag{6}$$

where $\mathcal{N} \subseteq \mathbb{R}^p$ is an unknown nuisance subspace. The nuisance deviation may depend on both mechanism and domain. Set $a_m := \mu_{m,d_0}$, so $b_{m,d_0} = 0$, and let the observed development span be

$$\mathcal{N}_{\mathrm{dev}} = \mathrm{span}\{\mu_{m,d_k} - \mu_{m,d_0} : m \in \mathcal{M},\ k = 1, \ldots, K\}. \tag{7}$$

Intuition. Development domains reveal only nuisance directions that actually occur in those domains. If a mechanism contrast has a component outside that observed span, the same development data are compatible with two worlds: one in which the component is mechanism-specific, and another in which an unseen nuisance direction reproduces it. The next theorem formalizes this ambiguity.

Theorem 2 (No source-only distributional certification under a common-location response family)**.** *In addition to Eq.* (6)*, suppose $P^Q_{m,d} = \mathsf{P}_Q(\cdot - \mu_{m,d})$ for a common location-family base law P_Q. Fix $m \neq m'$ and write $\delta_{m,m'} = a_{m'} - a_m$. If $P_{\mathcal{N}^\perp_{\mathrm{dev}}} \delta_{m,m'} \neq 0$, the finite development observations are compatible both with $\mathcal{N}_0 = \mathcal{N}_{\mathrm{dev}}$, where the pair has positive transverse separation, and with*

$$\mathcal{N}_1 = \mathcal{N}_{\mathrm{dev}} + \mathrm{span}\{P_{\mathcal{N}^\perp_{\mathrm{dev}}} \delta_{m,m'}\},$$

where an admissible unseen domain can satisfy $P^Q_{m,d_\star} = P^Q_{m',d_\star}$. Hence finite development responses alone cannot certify uniform unseen-domain attribution over all linear nuisance completions containing $\mathcal{N}_{\mathrm{dev}}$.

The nuisance deviation is deliberately allowed to depend on mechanism and domain: Theorem 2 asks what can be certified *without* extra structure on future shifts. The common-location assumption lifts a constructed mean collision to equality of full response laws; without it, the same construction gives no certification of mean-response identifiability. The global orbit result does not assume linear nuisance variation. The additive model below is a structured characterization used for transversality, sharp rates, and finite-domain recovery; the empirical subspace analyses are diagnostics of this structure rather than claims that MNIST–Fashion or KMNIST shifts are exactly linear.

4.3 Exact transversality: what survives a known nuisance family?

For a known nuisance subspace $\mathcal{N}$, define pairwise mechanism contrasts $\delta_{m,m'} = a_{m'} - a_m$ and the *transverse mechanism margin*

$$\tau_Q = \min_{m \neq m'} \|P_{\mathcal{N}^\perp} \delta_{m,m'}\|_2. \tag{8}$$

Theorem 3 (Exact linear transversality characterization)**.** *Under Eq.* (6)*, the following are equivalent:*

(i) there exists a linear map T with $T\mathcal{N} = \{0\}$ such that $Ta_1, \dots, Ta_M$ are pairwise distinct;

(ii) $\tau_Q > 0$.

For every nuisance-invariant linear map with $\|T\|_{\mathrm{op}} \leq 1$,

$$\|T\delta_{m,m'}\|_2 \leq \|P_{\mathcal{N}^\perp} \delta_{m,m'}\|_2, \tag{9}$$

so orthogonal projection preserves the largest Euclidean pairwise separation available to any norm-bounded linear invariant representation.

Theorem 3 formalizes the central principle: a detectable mechanism difference is not attribution-valid evidence when declared nuisance variation can reproduce it. The margin τ_Q is the smallest pairwise separation that remains after removing those nuisance directions.

4.4 A sharp binary Gaussian benchmark for probe-batch complexity

The previous result is geometric. We next isolate its statistical consequence in a deliberately idealized reference model. The Gaussian assumption and known common covariance below are used to obtain a sharp benchmark; we do *not* assume that the empirical generator responses are exactly Gaussian or that their covariance is known. Consider two mechanisms with

$$Z_i = a_m + b + \xi_i, \qquad b \in \mathcal{N}, \qquad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \tag{10}$$

where $\Sigma \succ 0$ is known and the nuisance value may differ under the two hypotheses. Here Φ denotes the standard normal cumulative distribution function. Let

$$\widetilde{\mathcal{N}} = \Sigma^{-1/2}\mathcal{N}, \qquad \tau_\Sigma = \|P_{\widetilde{\mathcal{N}}^\perp} \Sigma^{-1/2}(a_1 - a_2)\|_2. \tag{11}$$

Theorem 4 (Sharp binary Gaussian minimax benchmark)**.** *Under Eq.* (10) *for $M = 2$ with equal prior weight on the two mechanisms, define*

$$R^*_{n,\mathrm{G}} = \inf_{\psi} \sup_{b_1, b_2 \in \mathcal{N}} \frac{1}{2}\{P^{\otimes n}_{1,b_1}(\psi = 2) + P^{\otimes n}_{2,b_2}(\psi = 1)\}.$$

Then

$$R^*_{n,\mathrm{G}} = \Phi\left(-\frac{\sqrt{n}\,\tau_\Sigma}{2}\right). \tag{12}$$

The minimax rule whitens the sample mean, projects onto $\widetilde{\mathcal{N}}^\perp$, and applies the Gaussian likelihood-ratio test. In particular, $\tau_\Sigma = 0$ gives exact non-identifiability. For $\tau_\Sigma > 0$ and target error $\alpha \in (0, 1/2)$, the condition $R^*_{n,\mathrm{G}} \leq \alpha$ is equivalent to

$$n \geq \frac{4}{\tau_\Sigma^2}\left[\Phi^{-1}(1-\alpha)\right]^2. \tag{13}$$

Consequently, as $\alpha \downarrow 0$, the query complexity obeys $n = \Theta(\tau_\Sigma^{-2} \log(1/\alpha))$.

The exact identity in Theorem 4 is binary. For $M > 2$, the exact minimax risk depends on the full projected configuration of mechanism means; Appendix A gives pairwise lower and nearest-centroid upper bounds with the same $\Theta(\tau_\Sigma^{-2} \log(1/\alpha))$ query-complexity scaling. Thus adding response coordinates is useful only if they increase nuisance-orthogonal Mahalanobis separation; coordinates concentrated in nuisance directions can increase raw signal without improving the attribution rate.

4.5 Recovering nuisance geometry from finitely many development domains

The remaining question is whether $\mathcal{N}$ can be estimated from development data. Let $c_{m,k} = \mu_{m,d_k} - \mu_{m,d_0}$ and stack the $L = MK$ development contrasts as $C \in \mathbb{R}^{p \times L}$. Define

$$G = \frac{1}{L} CC^\top, \qquad r = \operatorname{rank}(C), \qquad g = \lambda_r(G) > 0, \tag{14}$$

and let $\Lambda = \sqrt{\lambda_1(G)}$. The condition $g > 0$ means that the observed domain contrasts excite all nuisance directions represented by the model. Suppose $\widehat{C} = C + E$ is obtained from finite generator/probe-batch averages, and $\widehat{\mathcal{N}}$ is the top-r left singular subspace of $\widehat{C}$. Put

$$e_C = \frac{\|E\|_{\mathrm{op}}}{\sqrt{L}}, \qquad \rho = 2\Lambda e_C + e_C^2, \qquad \eta = \|P_{\mathcal{N}} - P_{\widehat{\mathcal{N}}}\|_{\mathrm{op}}. \tag{15}$$

Theorem 5 (Finite-development nuisance recovery under declared coverage)**.** *Assume $\mathcal{N} = \operatorname{col}(C)$ and $\rho < g/2$. Then*

$$\eta \le \frac{2\rho}{g}. \tag{16}$$

Suppose $\max_{m \neq m'} \|a_m - a_{m'}\|_2 \le A$ and $\|b_{m,d}\|_2 \le R$. Let $\widehat{a}_m$ be source-domain estimates of the reference mechanism means and define

$$\zeta = \max_m \|P_{\widehat{\mathcal{N}}^\perp}(\widehat{a}_m - a_m)\|_2.$$

For nearest-centroid attribution after projection onto $\widehat{\mathcal{N}}^\perp$, define the effective margin

$$\gamma_Q = \tau_Q - (A + 2R)\eta - 2\zeta. \tag{17}$$

Let $\mathcal{F}_{\mathrm{dev}}$ denote the sigma-field generated by the development data. Condition on $\mathcal{F}_{\mathrm{dev}}$, so that $\widehat{\mathcal{N}}$, $\widehat{a}_m$, η, ζ, and γ_Q are fixed. Assume $\gamma_Q > 0$ and that the projected mean of n fresh independent probe-batch statistics satisfies the conditional norm-concentration condition in Appendix A.5. Then nearest-centroid attribution satisfies

$$\sup_{m,d} \mathbb{P}(\widehat{m} \neq m \mid \mathcal{F}_{\mathrm{dev}}) \le 2 \exp\left(-\frac{n\gamma_Q^2}{8\sigma_{Q,\perp}^2} \right). \tag{18}$$

Equation (18) is a *conditional source-based bound under declared coverage*: finite domains and finite replication determine how accurately the nuisance subspace and reference mechanism means can be estimated; those errors reduce the transverse margin; the remaining margin controls probe-batch attribution. Theorem 5 is deterministic in the contrast-estimation error E and therefore permits correlated errors. This matches our development contrasts, which share a reference domain and are consequently not entrywise independent. Appendix A gives an independent-entry sub-Gaussian specialization only as an illustrative sufficient condition. If $\operatorname{rank}(C)$ is smaller than the true nuisance dimension, Theorem 2 applies: unobserved nuisance directions cannot be certified away from source data alone.

What the positive theorem does not certify. Theorem 5 is conditional on a declared nuisance family, complete development-domain coverage of that family, and population quantities such as τ_Q and $\sigma_{Q,\perp}$. These conditions are not estimated or certified in the real-generator experiments. Appendix A records a conditional positive-margin implication if a valid finite-sample confidence radius is available, but we do not instantiate a deployable certificate for trained generators. Theorem 2 explains why source data alone cannot provide a universal certificate that all future nuisance directions have been observed.

Assumption–experiment correspondence. The theoretical layers have different empirical roles. Failed transfer does not establish orbit overlap; the additive model is a structured explanation rather than a claim of exact linear shift; and the Gaussian theorem is a sharp benchmark rather than a fitted likelihood. The recovery theorem is conditional on $\mathcal{N} = \mathrm{col}(C)$: KMNIST illustrates why coverage matters but does not verify it. The deterministic bound permits the shared-reference correlations in our development contrasts. Appendix C adds a controlled synthetic check in which the nuisance family and transverse margin are known: Monte Carlo error follows the exact binary Gaussian formula, and nuisance-subspace recovery exhibits the predicted rank/replication behavior.

5 Experimental Protocol

The experiments are controlled stress tests of attribution validity. They establish frozen cross-domain non-transfer and measure response geometry; they do not infer information-theoretic impossibility from a failed classifier. The independently replicated unit is a trained generator seed, not an image.

Two attribution regimes. Study I compares ϵ-, x_0-, and v-prediction diffusion models on MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017), with three seeds per label/domain and a shared architecture, schedule, sampler, training configuration, and paired noise bank. This is a fine-grained training-trace setting: the sampling interface is controlled, while the labels record which analytically related prediction parameterization was used during training. Study II estimates a nuisance projection from MNIST and Fashion-MNIST, freezes the full representation, and evaluates three seeds per mechanism on KMNIST (Clanuwat et al., 2018). Study III compares standard diffusion, Flow Matching, and consistency/distillation with matched architecture and training budget but native DDIM, Euler, and consistency samplers. It therefore concerns deployed generative mechanisms, not causal recovery of training loss; an intervention such as NFE $= 4$ is pipeline-relative and need not have identical operational meaning across samplers. A frozen multiscale S^2-CLIP control reuses the same completed Study III generators and samples, without retraining or target-domain adaptation (Radford et al., 2021; Shi et al., 2024).

Probes and frozen transfer. Noise-response uses scales $\{0.5, 0.75, 1.0, 1.25, 1.5\}$ and step-degradation uses NFE $\{1, 2, 4, 8, 16, 32\}$. The original response space uses sample standard deviation, mean pixel variance, and mean pairwise distance, yielding 15-, 18-, and 33-dimensional noise, step, and combined signatures. A strong control reuses the same completed outputs and replaces these statistics with frozen S^2-CLIP features followed by a source-only 10-dimensional Isomap and Gaussian geometric-JS distance to each generator's own baseline (Radford et al., 2021; Shi et al., 2024). Nearest centroid is primary and logistic regression secondary. All representation fitting, feature filtering, scaling, and classifier fitting are source-only; Appendix D gives the full S^2-CLIP protocol, counts, and audits.

6 Results

6.1 Strong source-domain signal does not establish attribution

Figure 2 shows the first boundary. Passive generator-level LOSO accuracy is 7/9 on MNIST and 5/9 on Fashion-MNIST; active probes raise both to 9/9. Under strict source-only transfer, Passive, Active, and Active without NFE 1/2 are all 3/9 in both directions. The errors collapse to a single class rather than spreading randomly.

An interface control shows that analytically converting one trained ϵ model into ϵ, x_0, and v outputs leaves the measured response unchanged under the common sampler. Checkpoint controls show a separate nuisance: Fashion-MNIST stays 9/9 across 50k, 25k, and quality-matched checkpoints, whereas MNIST changes from 9/9 to 7/9 and 6/9. Thus the source trace is not invariant to optimization state.

6.2 Seen-domain invariance does not certify unseen nuisance coverage

Generator-local normalization improves strict MNIST↔Fashion transfer from 3/9 to 6/9 but maps every v seed to x_0. A nuisance subspace estimated from same-mechanism contrasts across the two development datasets performs much better on those seen domains: rank-2 projection gives 17/18

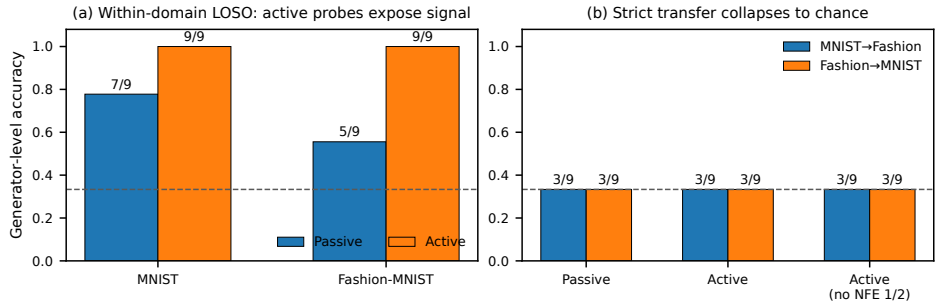

Figure 2: **Signal is not attribution.** Active response curves separate $\epsilon/x_0/v$ generators within MNIST and Fashion-MNIST, but the same frozen representation collapses to chance under strict bidirectional transfer. Dashed line: three-class chance.

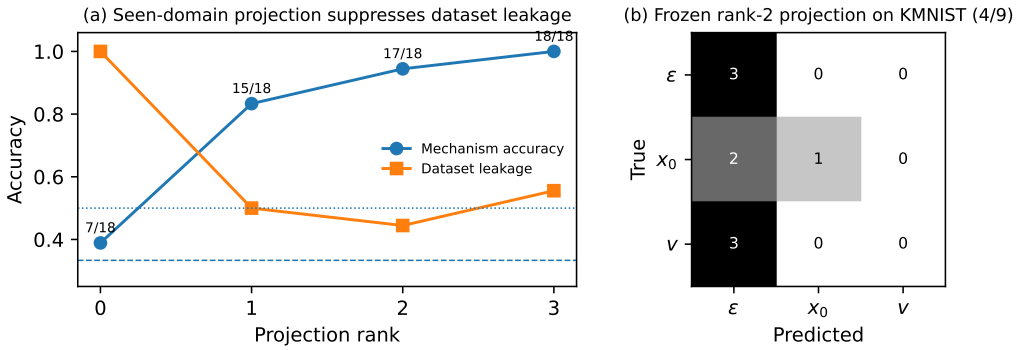

Figure 3: **Seen-domain invariance is not unseen-domain coverage.** A nuisance projection largely removes MNIST/Fashion dataset information while preserving mechanism labels, but the frozen representation drops to 4/9 on KMNIST.

nearest-centroid mechanism accuracy (18/18 logistic regression) while dataset predictability falls from 18/18 to 8/18 for both classifiers.

The frozen KMNIST result is 4/9 for both classifiers. Figure 3 summarizes the failure. The KMNIST mechanism centroids are shifted by 0.723, 1.023, and 1.061 relative to the frozen MNIST+Fashion centroids, whereas their within-KMNIST seed radii are 0.285, 0.129, and 0.068. The three shift directions are strongly aligned (pairwise cosine 0.834, 0.871, and 0.970). This pattern is consistent with incomplete nuisance coverage. It illustrates why the positive coverage condition in Theorem 5 must be assumed or certified; it does not verify $\mathcal{N} = \mathrm{col}(C)$, nor is it used as evidence that the real shift is exactly linear.

6.3 Stronger representations improve transfer but do not remove direction dependence

Study III removes the objection that $\epsilon/x_0/v$ are algebraically related parameterizations. Under the original low-order response summary, combined-response logistic regression is 9/9 for MNIST→Fashion-MNIST but 4/9 in reverse; combined nearest centroid is 6/9 and 3/9.

A frozen S^2-CLIP control tests whether the observed non-transfer is specific to weak low-order statistics. On the same 18 generators, combined nearest-centroid transfer improves from 6/9 to 8/9 for MNIST→Fashion-MNIST but only from 3/9 to 5/9 in reverse; the stronger representation improves this transfer without removing directional dependence. This is a representation control, not evidence that every purpose-built attribution method must fail. Appendix D reports the full classifier and probe breakdown.

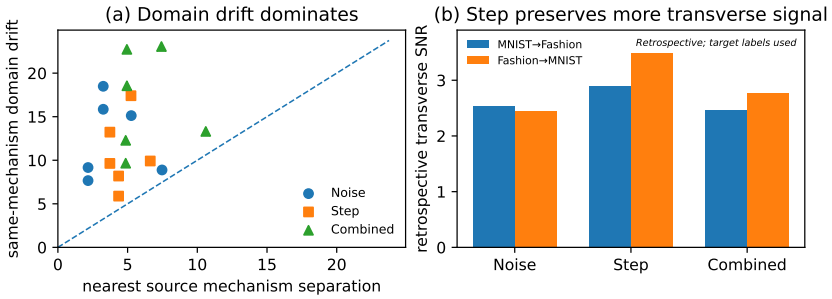

Figure 4: **Observed response geometry.** (a) All 18 conditions have domain drift above nearest source mechanism separation. (b) A target-label-dependent retrospective transverse-SNR diagnostic ranks step-degradation highest in both directions; it is explanatory and not deployable.

Observed geometry and transversality. In the pre-specified low-order response space, all preprocessing is source-fitted. Across all 18 direction×probe×mechanism conditions, same-mechanism domain drift exceeds the nearest source mechanism separation (median ratio 2.99), and leave-one-seed-block analyses preserve the 18/18 inequality. Seed RMS is smaller than domain drift. Because S^2-CLIP changes the geometry, this is a diagnostic of the frozen low-order representation rather than a representation-independent law; full ranges are in Appendix D.

Because target labels are used to estimate the nuisance span, transverse-SNR is strictly retrospective. Projecting source mechanism contrasts away from the observed same-mechanism domain shifts gives SNRs of 2.90/3.49 for step, 2.54/2.44 for noise, and 2.47/2.77 for combined response (MNIST→Fashion / reverse). Step is highest despite the combined signature having more coordinates, consistent with the structured geometry; the S^2-CLIP control further shows that probe ranking can depend on the response representation.

7 Discussion and Limitations

Claims. Source recognition separates sampled responses, whereas population identifiability requires separation over a declared nuisance family. Finite-development certification further asks whether the observed domains support such a claim on unseen domains. Theorem 2 characterizes why this certification is not automatic, while Theorem 5 gives the complementary positive guarantee under declared nuisance coverage. The real-generator studies examine this distinction under frozen transfer protocols rather than estimate population-level theoretical quantities.

Representation. S^2-CLIP improves transfer asymmetrically (8/9 versus 5/9), showing that representation choice changes the observed attribution geometry without removing its directional dependence. The Gaussian benchmark motivates probe designs that preserve nuisance-orthogonal information; practically calibrated certificates and optimized probe selection are natural extensions of this framework.

Scope. The empirical study uses a controlled three-domain setting with independently trained generator seeds as the replication unit; generated images are repeated measurements. Study I examines related training parameterizations, whereas Study III considers deployed pipelines with native samplers. S^2-CLIP serves as a stronger frozen-representation control rather than a SOTA attribution benchmark.

8 Conclusion

Black-box attribution requires separable response families over a declared nuisance family and evidence that development domains sufficiently cover that family; source recognition alone provides neither guarantee. Our analysis distinguishes detectable mechanism-related variation from variation that remains identifiable after nuisance effects are removed, and shows why finite development domains cannot certify unseen-domain attribution without additional coverage assumptions.

AI USE STATEMENT

Generative AI tools were used only for language polishing and figure/table preparation. They were not used for research-question refinement, theoretical formulation, proof drafting, experimental design, result interpretation, literature search, or manuscript organization. All scientific claims, proofs, experiments, numerical results, and citations were authored, executed, and verified by the authors. The authors take full responsibility for the final content.

ETHICS STATEMENT

Mechanism attribution can be used in provenance and forensic settings, where overconfident conclusions may have material consequences. This work studies limits on such conclusions. We recommend that attribution systems state their nuisance assumptions and abstain outside supported coverage rather than treating a source-domain fingerprint as conclusive evidence.

REPRODUCIBILITY STATEMENT

The main-text results correspond to recorded generator-level experiments with source-only preprocessing and frozen target evaluation. Appendix D records the model counts, probe grids, feature definitions, classifiers, controls, nuisance-projection protocol, and geometry calculations. The supplementary source package includes the fixed-seed synthetic script and CSV outputs together with recorded empirical summaries, predictions, preprocessing audits, and geometry tables used to check the reported counts. Appendix D states the training budget and sampler choices available from the recorded runs, and Appendix A gives complete proofs. Large checkpoints and generated-image tensors are not required to compile the paper and are not embedded in the Overleaf source.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Manuel Cebrian, Andres Abeliuk, and Jan Arne Telle. Can adversarial attacks by large language models be attributed? *PLOS Complex Systems*, 3(2):e0000085, 2026. doi: 10.1371/journal.pcsy.0000085.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1989.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Guangyu Nie, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributing image generative models using latent fingerprints. In *International Conference on Machine Learning*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Thomas J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, 2024.

Hae Jin Song and Laurent Itti. Riemannian-geometric fingerprints of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11425–11435, 2025.

Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. ManiFPT: Defining and analyzing fingerprints of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10791–10801, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7556–7566, 2019.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

A Proofs

A.1 Proof of Proposition 1

Fix $m \neq m'$ and $d, d' \in \mathcal{D}$. Any multiclass estimator induces a binary test between $P = P^Q_{m,d}$ and $P' = P^Q_{m',d'}$ by retaining the decisions m, m' and assigning all other outputs arbitrarily. Under equal priors, the minimum average binary error is $(1 - \mathrm{TV}(P, P'))/2$. Hence the larger conditional error is at least this value, and the multiclass worst-case risk is at least as large. Taking a sequence of admissible pairs approaching the infimum in Eq. (4) proves Eq. (5). If the response orbits intersect, the infimum total variation distance is zero.

A.2 Proof of Theorem 2

Let $P_{\mathrm{dev}} = P_{\mathcal{N}_{\mathrm{dev}}}$ and let $v = P_{\mathcal{N}_{\mathrm{dev}}^{\perp}} \delta_{m,m'}$, which is nonzero by assumption. Consider the two completions

$$\mathcal{N}_0 = \mathcal{N}_{\mathrm{dev}}, \qquad \mathcal{N}_1 = \mathcal{N}_{\mathrm{dev}} + \mathrm{span}\{v\}.$$

Every observed development contrast belongs to $\mathcal{N}_{\mathrm{dev}}$, so both completions are compatible with all development means. Under $\mathcal{N}_0$, the pairwise transverse margin is $\|v\|_2 > 0$. Under $\mathcal{N}_1$, we have $\delta_{m,m'} = P_{\mathrm{dev}} \delta_{m,m'} + v \in \mathcal{N}_1$. For a new domain d_*, set $b_{m,d_*} = \delta_{m,m'}$ and $b_{m',d_*} = 0$. Both are admissible in $\mathcal{N}_1$, and

$$\mu_{m,d_*} = a_m + \delta_{m,m'} = a_{m'} = \mu_{m',d_*}.$$

Thus the same observed development data admit one completion with positive transverse separation and another with an unseen-domain mean collision. Under the common-location assumption in Theorem 2, equality of the two means implies equality of the complete response laws,

$$P^Q_{m,d_*} = P^Q_{m',d_*}.$$

Any certification rule based only on the development responses must return the same answer in both worlds and therefore cannot be uniformly valid over both completions. Without the common-location assumption, the construction still establishes non-certifiability of mean-response identifiability, but not equality of arbitrary higher-order response laws.

A.3 Proof of Theorem 3

Suppose $\tau_Q > 0$ and take $T = P_{\mathcal{N}^\perp}$. This map vanishes on $\mathcal{N}$. For every $m \neq m'$,

$$\|Ta_{m'} - Ta_m\|_2 = \|P_{\mathcal{N}^\perp}\delta_{m,m'}\|_2 \geq \tau_Q > 0,$$

so all projected mechanism means are distinct. Conversely, if $\tau_Q = 0$, some $\delta_{m,m'}$ lies in $\mathcal{N}$. Every nuisance-invariant linear map vanishes on $\mathcal{N}$, hence $T\delta_{m,m'} = 0$ and the pair cannot be separated. For maximality, nuisance invariance implies $T = TP_{\mathcal{N}^\perp}$, so for every contrast δ,

$$\|T\delta\|_2 = \|TP_{\mathcal{N}^\perp}\delta\|_2 \leq \|T\|_{\mathrm{op}}\|P_{\mathcal{N}^\perp}\delta\|_2 \leq \|P_{\mathcal{N}^\perp}\delta\|_2.$$

A.4 Proof of Theorem 4

Whiten the observations: $Y_i = \Sigma^{-1/2}Z_i$. Let $\widetilde{a}_m = \Sigma^{-1/2}a_m$ and $\widetilde{\mathcal{N}} = \Sigma^{-1/2}\mathcal{N}$. The projected sample mean

$$\bar{Y}_\perp = P_{\widetilde{\mathcal{N}}^\perp}\frac{1}{n}\sum_{i=1}^{n} Y_i$$

has covariance $n^{-1}I$ on $\widetilde{\mathcal{N}}^\perp$ and mean $P_{\widetilde{\mathcal{N}}^\perp}\widetilde{a}_m$, independent of the nuisance value. The equal-prior likelihood-ratio rule on this projected statistic therefore has error

$$\Phi\left(-\frac{\sqrt{n}}{2}\|P_{\widetilde{\mathcal{N}}^\perp}(\widetilde{a}_1 - \widetilde{a}_2)\|_2\right) = \Phi\left(-\frac{\sqrt{n}\tau_\Sigma}{2}\right)$$

uniformly over all nuisance values, giving an upper bound on the minimax risk.

For the matching lower bound, decompose $\widetilde{a}_1 - \widetilde{a}_2 = \delta_\perp + \delta_\parallel$ with $\delta_\parallel \in \widetilde{\mathcal{N}}$. Choose nuisance values so that the whitened nuisance under mechanism 1 is zero and under mechanism 2 is $\delta_\parallel$. The resulting two simple Gaussian hypotheses have mean difference exactly $\delta_\perp$ and covariance I. Their optimal equal-prior error for n observations is $\Phi(-\sqrt{n}\|\delta_\perp\|_2/2)$. The composite minimax risk is at least the Bayes risk of this admissible pair, which matches the projection rule. Equation (13) follows algebraically from Eq. (12); the stated $\Theta(\tau_\Sigma^{-2}\log(1/\alpha))$ scaling follows from the standard Gaussian tail asymptotics.

Corollary 6 (Multiclass Gaussian consequence)**.** *Consider $M > 2$ mechanisms under the same whitened Gaussian nuisance model and let*

$$\tau_{\Sigma,\min} = \min_{m \neq m'} \left\| P_{\widetilde{\mathcal{N}}^\perp}\Sigma^{-1/2}(a_m - a_{m'}) \right\|_2 .$$

For worst-case multiclass error $R^*_{n,M}$,

$$\Phi\left(-\frac{\sqrt{n}\,\tau_{\Sigma,\min}}{2}\right) \leq R^*_{n,M} \leq (M-1)\Phi\left(-\frac{\sqrt{n}\,\tau_{\Sigma,\min}}{2}\right),$$

where the upper bound is achieved by projected nearest-centroid classification followed by a union bound. Thus the exact multiclass constant depends on the full projected geometry, but the worst-pair query complexity has the same $\Theta(\tau_{\Sigma,\min}^{-2}\log(1/\alpha))$ scaling for fixed M.

Proof. The lower bound follows by restricting any multiclass rule to the closest pair and applying Theorem 4. For the upper bound, under true mechanism m, an error implies that at least one projected competitor m' beats m. The pairwise Gaussian error against each competitor is at most $\Phi(-\sqrt{n}\,\tau_{\Sigma,\min}/2)$; summing over at most $M-1$ competitors gives the stated bound. □

A.5 Proof of Theorem 5

Norm-concentration convention. For the statement of Theorem 5, condition on $\mathcal{F}_{\mathrm{dev}}$ and define $\sigma_{Q,\perp}$ by the uniform tail condition

$$\mathbb{P}\big(\big\|P_{\widehat{\mathcal{N}}^\perp}\{\overline{Z}_n - \mu_{m,d}\}\big\|_2 \geq t \,\big|\, \mathcal{F}_{\mathrm{dev}}\big) \leq 2\exp\left(-\frac{nt^2}{2\sigma_{Q,\perp}^2}\right), \qquad t>0,\ (m,d)\in\mathcal{M}\times\mathcal{D}.$$

Thus $\sigma_{Q,\perp}$ is a conditional norm-concentration scale for the fresh probe-batch mean, not a per-coordinate variance parameter.

Let

$$\widehat{G} = \frac{1}{L}\widehat{C}\widehat{C}^\top = G + \frac{1}{L}(CE^\top + EC^\top + EE^\top).$$

Since $\|C\|_{\mathrm{op}}/\sqrt{L} = \sqrt{\lambda_1(G)} = \Lambda$ and $\|E\|_{\mathrm{op}}/\sqrt{L} = e_C$,

$$\|\widehat{G} - G\|_{\mathrm{op}} \leq 2\Lambda e_C + e_C^2 = \rho.$$

The eigenvalues of G satisfy $\lambda_r(G) = g$ and $\lambda_{r+1}(G) = 0$. Under $\rho < g/2$, Davis–Kahan gives

$$\|P_{\mathcal{N}} - P_{\widehat{\mathcal{N}}}\|_{\mathrm{op}} \leq \frac{2\rho}{g},$$

which is Eq. (16).

Let $P = P_{\mathcal{N}^\perp}$ and $\widehat{P} = P_{\widehat{\mathcal{N}}^\perp}$. Since $\|P - \widehat{P}\|_{\mathrm{op}} = \eta$, for every pair $m \neq m'$,

$$\|\widehat{P}(a_m - a_{m'})\|_2 \geq \tau_Q - \eta A.$$

For a target mean $\mu_{m,d} = a_m + b_{m,d}$, the distance to its own estimated projected centroid is at most

$$\|\widehat{P}\mu_{m,d} - \widehat{P}\widehat{a}_m\|_2 \leq \eta R + \zeta,$$

because $Pb_{m,d} = 0$. Its distance to a competing estimated centroid m' is at least

$$\|\widehat{P}\mu_{m,d} - \widehat{P}\widehat{a}_{m'}\|_2 \geq \tau_Q - \eta A - \eta R - \zeta.$$

Hence the deterministic gap between the nearest wrong centroid and the correct centroid is at least $\gamma_Q = \tau_Q - (A+2R)\eta - 2\zeta$. If the projected probe-batch sample mean deviates from its target mean by less than $\gamma_Q/2$, the same ordering is preserved and nearest-centroid attribution is correct. The assumed concentration bound gives Eq. (18).

The preceding argument is deterministic in E and therefore remains valid when contrast errors are correlated, as happens when several development contrasts share the same reference-domain estimate.

A.6 Conditional positive-margin implication

The next corollary is a *certificate template*, not an implemented deployment algorithm: this paper does not derive or calibrate a data-dependent radius $r_{K,S,n}$ for the real-generator experiments.

Corollary 7 (Conditional positive-margin implication under declared nuisance coverage)**.** *Assume the deployment nuisance family is contained in the declared subspace $\mathcal{N}$ and the coverage condition of Theorem 5 holds. Let $\widehat{\tau}_Q$ be an empirical transverse margin and, for some $\beta \in (0,1)$, suppose an externally justified or separately derived radius $r_{K,S,n}(\beta)$ satisfies*

$$\mathbb{P}(|\widehat{\tau}_Q - \tau_Q| \leq r_{K,S,n}(\beta)) \geq 1-\beta.$$

If $\widehat{\tau}_Q > r_{K,S,n}(\beta)$, then, conditional on the declared coverage assumption, $\tau_Q > 0$ and the applicable finite-query guarantee can be evaluated with lower confidence margin $\widehat{\tau}_Q - r_{K,S,n}(\beta)$. If the inequality is not met, the available evidence supplies no positive lower confidence bound on τ_Q. The result does not test whether the declared nuisance family covers all future domains.

Proof. On the stated confidence event,

$$\tau_Q \geq \widehat{\tau}_Q - r_{K,S,n}(\beta).$$

The strict inequality $\widehat{\tau}_Q > r_{K,S,n}(\beta)$ therefore yields a positive lower bound on the population transverse margin, conditional on the coverage assumption, and this lower bound can replace τ_Q in the corresponding finite-query guarantee. When the inequality fails, the event provides no positive lower bound. Theorem 2 rules out interpreting the corollary as a source-only test that all future nuisance directions have been observed. □

A.7 Illustrative independent-noise specialization

Corollary 8 (Independent-entry sub-Gaussian specialization)**.** *Under the setting of Theorem 5, suppose additionally that the entries of E are independent centered sub-Gaussian random variables with ψ_2 norm at most κ_E. Then for any $\beta \in (0,1)$, with probability at least $1-\beta$,*

$$e_C \le C\kappa_E\left(1+\sqrt{\frac{p}{L}}+\sqrt{\frac{\log(1/\beta)}{L}}\right) \tag{19}$$

for a universal constant C.

Proof. Standard sub-Gaussian random-matrix norm bounds give

$$\|E\|_{\mathrm{op}} \le C\kappa_E\big(\sqrt{L}+\sqrt{p}+\sqrt{\log(1/\beta)}\big)$$

with probability at least $1-\beta$. Dividing by $\sqrt{L}$ yields Eq. (19). □

Corollary 8 is only an illustrative sufficient condition. The shared-reference contrasts used in our experiments are block-correlated, so the empirical discussion relies on the deterministic perturbation statement in Theorem 5, not on entrywise independence. In a hierarchical implementation, independently trained generator seeds and repeated probe batches can still reduce the mean-estimation terms entering E and ζ at their usual square-root rates under the corresponding block-level assumptions.

A.8 A coverage-rank corollary

Corollary 9 (Rank is necessary for complete linear nuisance recovery)**.** *Under Eq.* (6)*, if the true nuisance dimension is $r_* > \mathrm{rank}(C)$, then the development contrasts do not uniquely determine the nuisance subspace. In particular, there exist distinct r_*-dimensional nuisance subspaces containing $\mathrm{col}(C)$ and agreeing on all development means.*

Proof. Any r_*-dimensional extension of $\mathrm{col}(C)$ is compatible with all observed development contrasts. If $\mathrm{rank}(C) < r_*$, at least one additional direction can be chosen in more than one way, so the nuisance subspace is not uniquely determined from the development data. □

B Attribution Ontology and Claim–Evidence Map

Study	Attribution label	Sampling control	Supported claim
I	historical diffusion prediction parameterization ($\epsilon/x_0/v$)	common diffusion sampler and interface	training-trace recognition/transfer stress test; not functional mechanism ontology
II	same Study-I labels after nuisance projection	representation frozen before KMNIST	seen-domain nuisance removal does not certify unseen coverage
III	deployed pipeline (diffusion / Flow Matching / consistency)	native DDIM / Euler / consistency samplers	deployed-mechanism attribution; not causal isolation of training loss
S^2-CLIP	same Study-III labels	same generators and outputs, stronger frozen feature map	stronger-representation control, not a SOTA attribution benchmark

Table 1: Ontology of the experimental labels. The word "mechanism" has different operational scopes across the controlled studies; the tasks are not pooled into one benchmark.

C Controlled Synthetic Validation of the Structured Theory

To check the quantitative statements in a setting where all assumptions are known, we run a small synthetic experiment with a fixed random seed. This experiment is separate from the trained-generator stress tests and is used only to verify the structured theory.

For Theorem 4, we generate balanced binary Gaussian responses with identity covariance and nuisance already projected out, using transverse separations $\tau_\Sigma \in \{0, 0.75, 1.5\}$ and $n \in \{1,2,4,8,16,32,64\}$ independent probe batches. Each point uses 50,000 Monte Carlo trials. Figure 5a shows that the empirical error lies on the exact curve $\Phi(-\sqrt{n}\tau_\Sigma/2)$; when $\tau_\Sigma = 0$, the error remains at chance.

Claim	Theoretical support	Empirical role
source recognition does not imply cross-domain attribution	response-orbit formulation	9/9 within-domain → 3/9 strict transfer
finite development domains do not certify future nuisance coverage	Theorem 2 under stated assumptions	17/18 seen-domain projection → 4/9 frozen KMNIST stress test
only nuisance-orthogonal contrast supports invariant linear attribution	Theorem 3	retrospective mechanism–nuisance geometry
stronger features need not yield uniform attribution	representation-dependent Q	S^2-CLIP improves selected transfers but remains directional

Table 2: Claim–evidence map. Empirical failures motivate and stress-test the theory; they are not used as proofs of information-theoretic impossibility.

For finite-development recovery, we generate a $p = 20$ response space with a true nuisance dimension $r = 3$. Development contrasts are sampled inside the nuisance subspace and observed with Gaussian estimation noise whose scale decreases as $S^{-1/2}$ for $S \in \{1, 4, 16\}$ replications. When the number of independent development contrasts $L < r$, the observed contrasts are rank deficient and the full nuisance projector is not recoverable; once $L \geq r$, recovery improves with both contrast coverage and replication (Figure 5b). This is a sanity check of the rank and perturbation logic, not a calibration of the real-generator experiments.

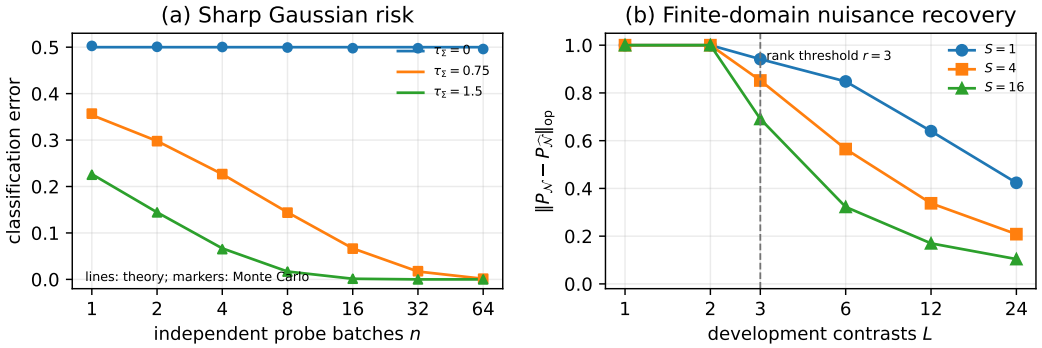

Figure 5: **Synthetic validation of the structured theory.** (a) Monte Carlo binary attribution error (markers) matches the exact Gaussian minimax risk (lines). (b) Nuisance-projector error is unrecoverable under rank-deficient development contrasts and decreases after the true rank is covered; additional replication reduces estimation noise.

The script and generated CSV tables are included in the supplementary source package for exact reproduction.

D Additional Experimental Details

D.1 Diffusion-parameterization study

The classes are ϵ-, x_0-, and v-prediction. Each dataset contains three independently trained generators per class. Generator-level LOSO uses one generator seed from each class as test data per fold. Active features summarize noise-rescaling and step-degradation response curves; passive features use only the standard generation condition. All normalization and feature filtering are fit inside the training fold.

For the interface-equivalence control, one fixed ϵ model is algebraically converted into x_0 and v outputs and all three representations are passed through the same sampler. The resulting response curves coincide up to numerical precision. For the checkpoint control, Active accuracy is 9/9, 7/9, and 6/9 on MNIST at 50k, 25k, and approximately quality-matched checkpoints, while Fashion-MNIST is 9/9 in all three settings.

D.2 NUISANCE PROJECTION AND KMNIST

The development-domain projection results are:

Rank	NC mech.	LR mech.	NC dataset	LR dataset
0	7/18	13/18	18/18	18/18
1	15/18	18/18	9/18	9/18
2	17/18	18/18	8/18	8/18
3	18/18	17/18	10/18	11/18

Rank 2 is selected before KMNIST because it gives the best balance between nearest-centroid mechanism retention and dataset suppression. This development-domain choice is a representation hyperparameter, not an estimate or certificate of the true future nuisance dimension. On KMNIST both classifiers are 4/9 with identical confusion matrices: ϵ is 3/3 correct, x_0 is 1/3 correct with two predictions as ϵ, and v is 0/3 with all three predictions as ϵ.

D.3 DISTINCT-PIPELINE STUDY

The three pipeline families use the same MNISTUNet backbone and training budget: 50k training steps, batch size 256, learning rate 2×10^{-4}, and zero weight decay. Standard diffusion, Flow Matching, and consistency/distillation use their native DDIM, Euler, and consistency multistep samplers, respectively. Noise-response scales are 0.5, 0.75, 1.0, 1.25, 1.5 and step-degradation uses NFE 1, 2, 4, 8, 16, 32. Every condition is summarized by sample standard deviation, mean pixel variance, and mean pairwise distance. The response dimensions are therefore 15, 18, and 33 for noise, step, and combined features. The operational pilot rule was fixed as combined-response nearest-centroid accuracy at least 6/9 in both transfer directions with all three predicted classes. The recorded result fails this rule.

For MNIST→Fashion, combined nearest centroid predicts standard diffusion 3/3, Flow Matching 0/3 (all as standard diffusion), and consistency 3/3. For Fashion→MNIST, combined nearest centroid predicts all nine generators as Flow Matching. Combined logistic regression is 9/9 in the first direction and 4/9 in the reverse. These asymmetric results are retained rather than selecting a favorable classifier or direction.

D.4 FROZEN S^2-CLIP STRONG-REPRESENTATION CONTROL

The control reuses the 18 Study III generators and their completed samples; no generator is retrained and no target-domain statistic is used to tune the representation. Each generator contributes 10 probe conditions with 256 images per condition, for 180 distributions and 46,080 images. Images are encoded by frozen CLIP ViT-B/32 with the parameter-free S^2 multiscale wrapper at scales 1 and 2 (Radford et al., 2021; Shi et al., 2024). The resulting features are reduced with a 10-dimensional Isomap fitted only on balanced source-domain landmarks. Each condition is represented by its Gaussian geometric-JS distance to the *same generator's* baseline (noise scale 1.0 or NFE 32). All Isomap fitting, covariance regularization, variance filtering, standardization, and classifier fitting are source-only.

The primary nearest-centroid results are:

Evaluation	Noise	Step	Combined
Fashion-MNIST LOSO	9/9	9/9	9/9
MNIST LOSO	9/9	6/9	7/9
Fashion-MNIST→MNIST	7/9	4/9	5/9
MNIST→Fashion-MNIST	6/9	6/9	8/9

Logistic regression gives the same qualitative conclusion; for combined response it is 9/9 for MNIST→Fashion-MNIST and 5/9 in reverse. Figure 6 shows the primary confusion matrices. The main residual confusion is standard diffusion versus Flow Matching. This control addresses one specific alternative explanation—that the original cross-domain failures arise only because standard

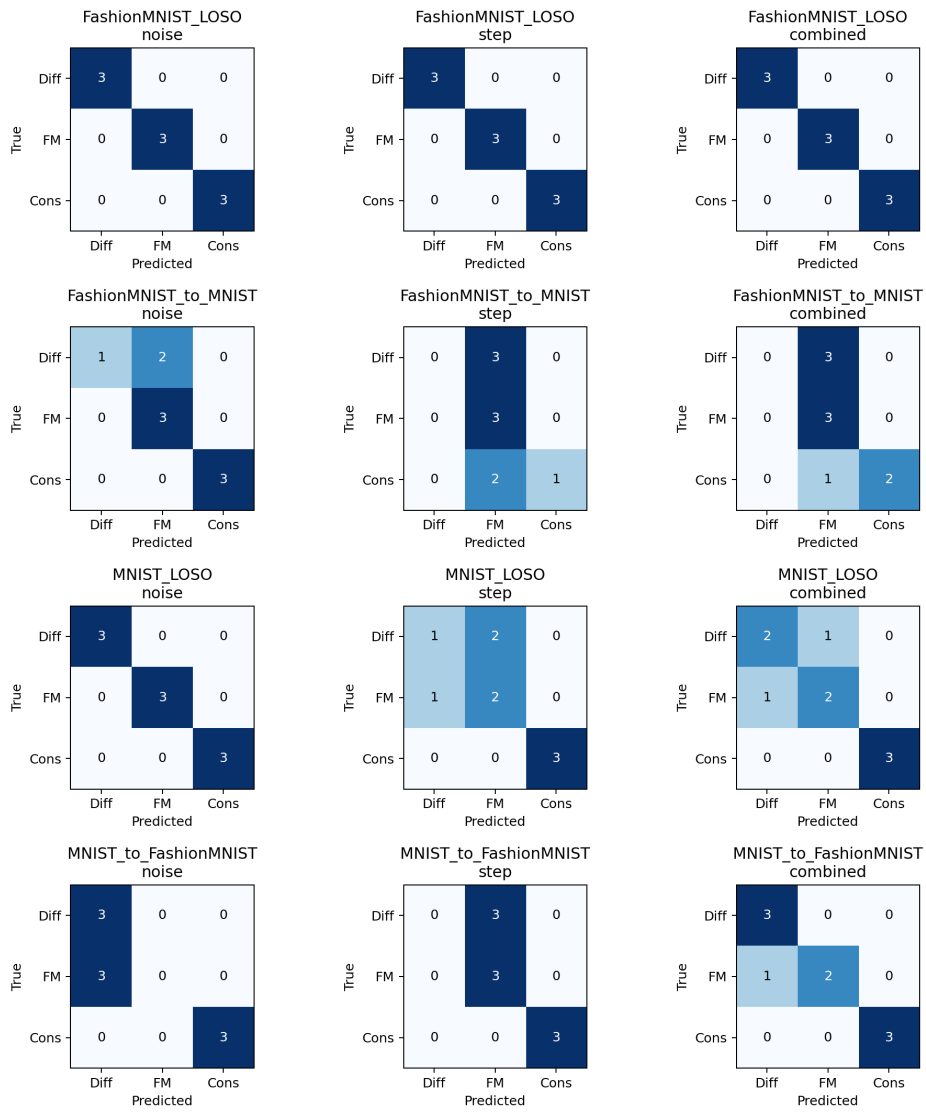

Figure 6: **S^2-CLIP control.** Generator-level nearest-centroid confusion matrices for within-domain LOSO and strict bidirectional transfer. Stronger features improve selected cells but retain direction-dependent confusion.

deviation, pixel variance, and pairwise distance are too weak—but it does not imply that every stronger representation must fail.

D.5 Geometry diagnostics and seed-block robustness

For each source-target direction, source-only standardization is applied to source and target response vectors. Source mechanism separation is the pairwise centroid distance. Domain drift is the distance between source and target centroids of the same mechanism. Seed RMS is the root mean squared distance of trained-generator seeds to their mechanism centroid.

Across all 18 direction–probe–mechanism conditions, domain drift exceeds the nearest source mechanism separation. Recomputing centroids after dropping the same seed index across every mechanism and both datasets preserves the 18/18 inequality for dropped seed blocks 0, 1, and 2. The corresponding median drift/separation ratios are 3.124, 3.246, and 2.752; the full-data value is 2.989.

The target-aware identifiability gap used during diagnosis is computed only after target labels are available and is therefore retrospective. Its Spearman correlations with mechanism-level target accuracy are 0.618 overall, 0.365 for MNIST→Fashion, 0.776 for Fashion→MNIST, 0.621 for noise-response, 0.833 for step-degradation, and 0.293 for combined response. The 18 observations are dependent because they share generators and transfer directions, so these correlations are not treated as confirmatory significance tests.

For the retrospective transverse-SNR diagnostic in Figure 4(b), the nuisance span is estimated after target labels are revealed from the three same-mechanism cross-domain centroid shifts. Source mechanism contrasts are projected onto its orthogonal complement. The score is the minimum projected pairwise mechanism separation divided by the mean projected seed RMS across source and target mechanisms. The scores are 2.54, 2.90, and 2.47 for noise, step, and combined features under MNIST→Fashion, and 2.44, 3.49, and 2.77 under Fashion→MNIST. Because this construction uses target mechanism labels, it is explanatory rather than deployable.

E Notation Summary

Symbol	Meaning
$P^Q_{m,d}$	response distribution for attribution label m, domain d, probe/representation Q
$P_{\mathcal{S}}$	orthogonal projector onto a subspace $\mathcal{S}$ (distinguished by its subspace subscript)
$\mathcal{O}^Q_m(\mathcal{D})$	response orbit over the declared nuisance family
$\mathcal{N}, \widehat{\mathcal{N}}$	true and estimated structured nuisance subspaces
τ_Q	minimum Euclidean nuisance-orthogonal mechanism margin
τ_Σ	binary Mahalanobis transverse margin in the Gaussian benchmark
C, E	development-contrast matrix and its estimation error
g, Λ	smallest positive and largest square-root eigenvalue scales of $CC^\top/L$
e_C, ρ, η	contrast-estimation scale, Gram perturbation, projector error
A, R, ζ	mechanism diameter, nuisance radius, reference-mean estimation error
γ_Q	effective finite-development attribution margin
$\sigma_{Q,\perp}$	norm-concentration scale of the projected probe-batch mean
κ_E	sub-Gaussian entry scale in Corollary 8

Table 3: Notation used in the structured finite-development theory.

F Claim Scope and Statistical Units

Study I probes weak, closely related training traces under a common sampler; Study III probes larger deployed-pipeline differences and therefore includes sampler behavior. We do not treat these labels as one homogeneous attribution task. The admissible nuisance family is also claim-specific. The global theory allows any declared $\mathcal{D}$; the linear model is a structured special case in which nuisance removal can be characterized exactly. The empirical MNIST/Fashion/KMNIST studies should be read as controlled stress tests of the validity distinction, not as universal quantitative claims about all generative models.

Generated images are repeated measurements from a trained generator, not independent training replicates. All reported generator-level accuracies therefore use trained seeds as the experimental unit. Probe-batch replication affects response estimation; it does not increase the number of independently trained models.

Computational environment. Experiments used Python 3.11.16 and PyTorch 2.8.0 with CUDA 12.8; the main dependencies were NumPy 2.4.6, SciPy 1.17.1, scikit-learn 1.9.0, POT 0.9.7.post1, and GeomLoss 0.3.1. GPU-based preprocessing used an NVIDIA GeForce RTX 4090. Full package versions are provided with the code.