BEYOND GLOBAL ERROR: PATH–FREQUENCY MEMBERSHIP INFERENCE IN DENOISING GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Denoising generative models expose training membership through their predictions on noisy inputs. Existing attacks often summarize prediction residuals into error statistics, compressing the available evidence before membership discrimination. Focusing on rectified flow, we uncover a richer structure: member–nonmember differences vary across spatial-frequency bands, and these frequency responses change along the interpolation path. Global error conceals how residual energy is distributed across bands, while isolated path positions miss complementary responses. To exploit this structure, we introduce PaFMIA (Path–Frequency Membership Inference Attack), which represents queried residuals jointly across path positions and spatial frequencies. PaFMIA either retains the complete band measurements or learns position-dependent weights that aggregate them, then fits a lightweight membership classifier using labeled source-model examples. The resulting attack learns from both the distribution of residual energy and its variation along the path. All frequency measurements reuse existing predictions, adding no generator queries. Our mathematical analysis characterizes information lost through aggregation, identifies conditions for complementary positions to improve detection at a fixed query budget, and derives frequency-dependent membership responses in a linear rectified-flow model. Experiments on four image datasets demonstrate the value of this representation: at 32 queries, PaFMIA reaches 67.41% TPR at 1% FPR on TinyImageNet, compared with 19.78% for DIME adapted to rectified flow. Controlled ablations identify the contributions of frequency resolution and joint positions, while variance-preserving flow matching and DDPM experiments extend the approach across flow and diffusion objectives.

1 INTRODUCTION

Membership inference asks whether a candidate example was used to train a model. For image generators, the answer exposes information about the training set even without reconstructing a memorized image. Effective membership inference is therefore important for privacy auditing: a weak attack can leave membership leakage undetected. The quality of the representation supplied to the attack is central to this problem.

Many attacks form their inputs using manually specified or empirically selected statistics, such as prediction error, reconstruction consistency, or errors measured at selected noise levels. These choices determine which parts of the model's response reach the attack. Learning a classifier over the resulting inputs improves how they are used, but cannot recover information already discarded during feature construction. Generative predictors return high-dimensional responses whose structure varies along a noise–data path. Reducing each response to a global error mixes image components with different relationships to membership. This raises a broader question: **how should an attacker represent queried predictions to capture the membership information that conventional summaries overlook?**

For diffusion and flow models, existing attacks exploit step-wise consistency, prediction-error geometry, frequency filtering, and learned profiles across positions (Duan et al., 2023; Lian et al., 2026; Do & Alabi, 2026; Rao & Moyer, 2026; Sesmat et al., 2026). We study the coupling between two

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.

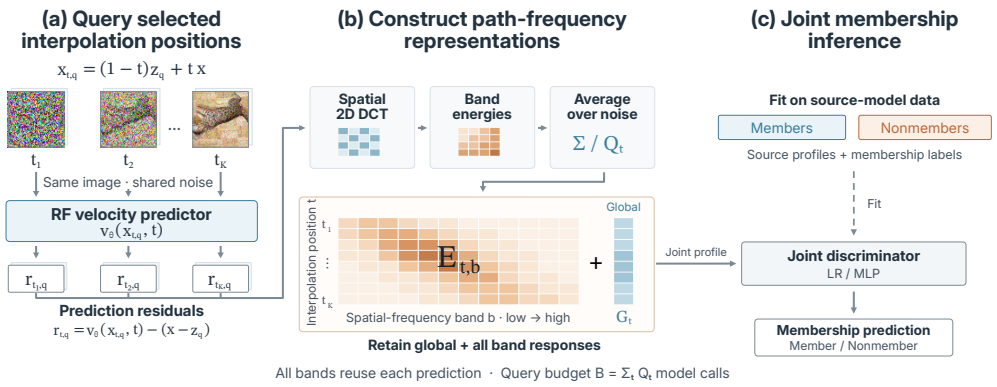


Figure 1: PaFMIA with rectified flow. (a) Shared-noise queries produce residuals at selected positions. (b) Global and bandwise energies form a joint profile. (c) A source-trained LR/MLP learns membership discrimination and is frozen for target evaluation. All bands reuse each prediction; the profile is schematic.

dimensions of these responses: *spatial-frequency bands* and *generative-path positions*. Focusing on rectified flow (RF), we find that member–nonmember residual differences vary across bands and that their frequency patterns change along the interpolation path. Two images with similar total error can distribute that error differently across bands; two positions can expose complementary membership responses for the same image. The useful signal lies in both the frequency allocation of residual energy and how that allocation changes along the path.

We introduce PaFMIA (Path–Frequency Membership Inference Attack), which represents prediction residuals jointly across spatial DCT bands and selected path positions (Figure 1). Its *full-profile* variant retains individual band energies for a lightweight classifier; its *learned-aggregation* variant learns position-dependent frequency weights before classification. Attack fitting uses source-model examples with known membership, and the fitted parameters are fixed for separately trained targets. Both variants reuse the queried predictions, so frequency decomposition requires no additional generator calls.

Our analysis connects the representation to three mechanisms. Global error is a fixed weighted sum of band energies, and testing theory characterizes when this compression loses membership information. Conditional-discrimination and shared-difficulty models explain how multiple positions supply new evidence and suppress common image variation. A linear RF analysis derives frequency-dependent membership-gap profiles along the path. We further characterize when compact aggregation preserves the relevant information and when retaining individual bands is beneficial.

Experiments on four image datasets validate PaFMIA. At 32 queries, the full profile reaches 67.41% TPR at 1% FPR on TinyImageNet, compared with 19.78% for DIME-RF, and improves over a stronger independently selected global-profile MLP by 17.69 points. Learned aggregation captures much of the benefit on several distributions. Controlled VP-FM and DDPM experiments extend the representation beyond straight-path RF.

Our contributions are threefold:

1. **Discovery.** We identify membership information in the joint path–frequency structure of prediction residuals: frequency-resolved responses vary along the path, revealing distinctions that global error summaries conceal.
2. **Method.** We develop PaFMIA, which exploits this structure through complete band profiles or learned position-dependent aggregation, using the same generator queries as the underlying residual measurements.
3. **Explanation and validation.** We connect information compression, complementary positions, and frequency-dependent RF responses through mathematical analysis, and test these mechanisms with controlled ablations, four-dataset attack evaluations, and flow/diffusion extensions.

2 BACKGROUND AND PROBLEM SETUP

2.1 PREDICTION TASKS IN DIFFUSION AND FLOW MODELS

We study unconditional image generators that learn a prediction task along a path connecting noise and data. Rectified flow (RF) is our primary setting; variance-preserving flow matching (VP-FM) and denoising diffusion probabilistic models (DDPM) provide two extensions. This viewpoint connects models through their noisy inputs and prediction residuals while retaining each model's native training target (Liu et al., 2023; Lipman et al., 2023; Ho et al., 2020).

For RF, let $x \in \mathbb{R}^{C\times H\times W}$ be an image and $z \sim \mathcal{N}(0, I)$ independent noise. With the noise-to-data convention, the straight interpolation path and target velocity are

$$x_t = (1-t)z + tx, \qquad u(x, z) = x - z, \qquad t \in [0, 1]. \tag{1}$$

A velocity predictor v_θ is trained by minimizing the mean squared difference between $v_\theta(x_t, t)$ and $u(x, z)$. The interpolation coordinate t specifies the noise–data mixture at which the predictor is queried. It is independent of the numerical step count used by an ODE solver to generate an image.

RF also admits a denoising interpretation. Its estimate of the clean image is $\widehat{x} = x_t + (1-t)v_\theta(x_t, t)$, and hence

$$\widehat{x} - x = (1-t)\big[v_\theta(x_t, t) - (x - z)\big]. \tag{2}$$

Thus a velocity residual describes clean-image reconstruction error up to a position-dependent scale. More generally, a Gaussian path has the form $x_t = \alpha_t x + \sigma_t z$: flow matching uses the reference velocity $\alpha'_t x + \sigma'_t z$, while our DDPM implementation predicts z. The candidate and sampled noise determine these targets. We compute the native velocity or noise residual before applying the same spatial-frequency analysis. Throughout, *path position* refers to a continuous interpolation coordinate for RF/VP-FM or a discrete diffusion timestep for DDPM.

2.2 MEMBERSHIP INFERENCE AND OBSERVABLE RESIDUALS

Given a trained generator and a candidate image x, membership inference predicts $m(x) = \mathbf{1}\{x \in D_{\text{train}}\}$. We consider prediction-interface access: the attacker can supply a noisy image and a path position and observe the predicted velocity or noise. The attacker does not need model parameters, gradients, or hidden activations. One generator query is one such prediction for one candidate, one position, and one noise realization.

For RF, the qth residual at position t is

$$x_{t,q} = (1-t)z_q + tx, \qquad r_{t,q}(x) = v_\theta(x_{t,q}, t) - (x - z_q). \tag{3}$$

The candidate image and sampled noise determine the reference target, making the complete residual image available from the prediction. Its global squared error is a conventional membership statistic. Our question is how much more useful evidence can be extracted from the structure of these residuals across spatial frequencies and path positions.

We train the attack on a source generator with known member and nonmember examples, then fix its parameters and query schedule for separately trained target generators. The source and targets use a common candidate-image population with different training memberships and seeds. Images used to fit or validate the attack are excluded from target testing. Section 4.1 specifies the experimental instances and Appendix A.1 gives the complete role assignments.

3 REVEALING AND LEARNING MEMBERSHIP SIGNALS

PaFMIA builds on three observations: global error compresses distinct spatial-frequency responses; different path positions supply complementary evidence; and RF training produces frequency-dependent responses along the path. Together, they motivate retaining the allocation of residual energy across bands and its variation across positions (Figure 1).

3.1 FREQUENCY STRUCTURE BEYOND GLOBAL ERROR

An image's total prediction error does not specify which spatial components produce it. Residuals concentrated in smooth structure or fine detail can have the same squared norm while conveying different membership evidence. Figure 2 shows that member–nonmember gaps vary across both frequency bands and path positions on TinyImageNet. Retaining this allocation gives the attack access to distinctions that global error discards. Our decomposition resolves spatial frequencies within each residual image; the other profile axis indexes the queried interpolation positions.
Let U be a channelwise orthonormal two-dimensional DCT, and P_b project onto one of $L = 12$ disjoint radial bands.

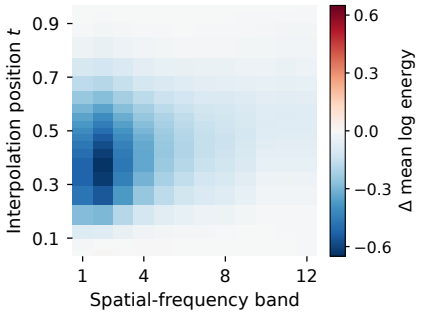


Figure 2: Member–nonmember log-energy gaps on TinyImageNet (two-target mean).

Write $d_b = C|\mathcal{B}_b|$ for the coefficient count in band b and $d = CHW$. With Q_t noise samples at position t, define

$$E_{t,b}(x) = \frac{1}{Q_t d_b} \sum_{q=1}^{Q_t} \|P_b U r_{t,q}(x)\|_2^2, \qquad G_t(x) = \frac{1}{Q_t d} \sum_{q=1}^{Q_t} \|r_{t,q}(x)\|_2^2. \tag{4}$$

Normalizing by d_b makes each band feature a mean squared residual per coefficient, so wider bands do not appear larger simply because they contain more coefficients. This separates average error intensity from band size. Orthogonality gives the exact compression identity

$$G_t = \sum_{b=1}^{L} \frac{d_b}{d} E_{t,b}. \tag{5}$$

Frequency information under compression. Let E collect the band energies, $C(E)$ be a deterministic summary such as the global profile, and $\beta_E(\alpha)$ the optimal detection power at FPR at most α. For the same measurements, $\beta_E(\alpha) \geq \beta_C(\alpha)$. Assume the member law is absolutely continuous with respect to the nonmember law. Equality at every FPR then holds exactly when the likelihood ratio is recoverable from C; failure of this recoverability condition gives strict improvement at some interior FPR (Appendix C, Proposition A1).

Table 2 tests this frequency increment with the same LR fitting procedure and query budget. A complementary experiment fixes eight positions, 32 queries, and LR. After standardization, a source-fitted linear regression predicts log-band features from the log-global profile. We freeze this regression and add its residual features to the global profile: TinyImageNet TPR rises from 38.03% to 57.07% at 1% FPR. Thus useful band structure remains after removing the linearly predictable component (Appendix A.6).

3.2 COMPLEMENTARY EVIDENCE ACROSS PATH POSITIONS

Different noise–image mixtures probe different predictor responses. An image may have large residuals at two positions because it is difficult to predict, while their change reveals membership. Joint measurements expose both absolute levels and relative responses. Table 2 compares one and two positions at 32 total queries: on TinyImageNet, the band-profile TPR rises from 15.42% to 36.78% despite halving repetitions per position.

An additional position can supply conditional evidence. Suppose nonredundant observations (X_1, X_2) are jointly Gaussian within each membership class, with shared positive-definite covariance Σ and member–nonmember mean differences δ_j. Their squared Mahalanobis separation satisfies

$$D_{12}^2 = D_1^2 + \delta_{2|1}^\top S^{-1} \delta_{2|1}, \quad \delta_{2|1} = \delta_2 - \Sigma_{21}\Sigma_{11}^{-1}\delta_1, \tag{6}$$

where $S = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. The increment is positive exactly when X_2 retains class contrast after accounting for X_1. Since optimal power is $\Phi(\Phi^{-1}(\alpha) + D)$, this improves every interior-FPR optimum while preserving the first observation's precision. Appendix C.3 proves the result and the symmetric condition for exceeding both single-position attacks.

With existing measurements retained, X_1 can stack all preceding positions and X_2 an additional position. The same decomposition then extends recursively to any finite position set: each new position contributes its remaining membership contrast (Appendix C.3, Corollary G1.1). The two-position ablation isolates this mechanism; the final attack selects multiple positions, whose budget-specific results appear in Appendix B.1.

Fixed-budget gains. Writing $Y_t = a_t m + U + \eta_t$ separates shared image difficulty U from query noise η_t. Across-position contrasts cancel U while retaining the membership response $(a_{t_2} - a_{t_1})m$. At a fixed budget, the contrast improves detection when this cancellation outweighs the increased query noise from fewer repeats per position. The Gaussian model makes this tradeoff explicit; Appendix A.8 gives the condition and shared-noise experiment, with the proof in Appendix C.3.

3.3 Why RF produces frequency-dependent path responses

Consider a diagonal linear RF predictor in a spatial-frequency basis, with $X \sim \mathcal{N}(0, \lambda)$ and independent $Z \sim \mathcal{N}(0, 1)$ in each coordinate. At each position it is fitted to n training examples with fresh noise integrated exactly. Membership compares a uniformly chosen training example with an independent population example. The expected nonmember–member squared-error gap is positive at interior positions, with large-sample profile

$$n\,\Delta_n(t; \lambda, 1) \longrightarrow H_\lambda(t) = \frac{4\lambda^2 t^2 (1-t)^2}{[t^2\lambda + (1-t)^2]^3}. \tag{7}$$

Larger spectral variance shifts the maximum of H_λ toward smaller t. Two such limiting profiles with unequal variances form a rank-two gap matrix at any two distinct interior positions. Appendix C, Propositions C1–C2, proves the finite-sample gap and limiting structure; Figure 4 illustrates these mechanisms.

These curves explain why one frequency's membership response need not remain a fixed multiple of another's along the path. A neural-model experiment constrains LR coefficients on standardized log-band features at eight positions and 32 queries. Shared coefficients use one frequency direction across positions; separable coefficients allow its scale to vary by position; unrestricted coefficients allow different frequency directions. TinyImageNet TPR rises from 42.18% to 54.75% and 57.91%, respectively (Appendix A.6). Position-specific scaling captures most of the gain, while unrestricted coefficients retain further joint variation.

3.4 PaFMIA: learning from path–frequency residual profiles

PaFMIA retains the profile $\{E_{t,b}\}_{t\in\mathcal{T}, b=1,\ldots,L}$, identifying both where the predictor is queried and which frequencies contribute to its error. We use two representations of these measurements, followed by a source-trained membership readout.

Full profile. For K selected positions $\mathcal{T}$, the full-profile variant uses

$$\phi(x) = \mathrm{vec}_{t\in\mathcal{T}}[\log(G_t(x) + \epsilon), \log(E_{t,1}(x) + \epsilon), \ldots, \log(E_{t,L}(x) + \epsilon)], \tag{8}$$

with $\epsilon = 10^{-12}$. Differences of log-transformed energies are log ratios of the stabilized energies. A linear readout can therefore use relative changes across bands or positions together with absolute log-energy levels; source standardization preserves this representational capacity. After logarithms, global error is a nonlinear function of band features; retaining it explicitly lets LR use both overall error magnitude and frequency allocation. The $(L+1)K$-dimensional representation supports any finite selected position set, including the single- and two-position cases.

Features are standardized with source-fitting means and scales. We fit a regularized LR or small MLP score $s_\psi = f_\psi(\phi)$ using binary cross-entropy, with membership probability $\mathrm{sigmoid}(s_\psi)$. Source validation selects the readout and its parameters, which are fixed for target evaluation. The MLP adds nonlinear combinations. Section 4.3 compares them on matched schedules.

Learned frequency aggregation. The compact variant learns position-dependent weights over the same bands:

$$\phi_{\mathrm{agg}}(x) = \left\{ \log(G_t + \epsilon), \log\left(\sum_b a_{t,b} E_{t,b} + \epsilon \right) \right\}_{t \in \mathcal{T}}, \quad a_{t,b} \geq 0, \quad \sum_b a_{t,b} = 1. \tag{9}$$

The weights are fitted with a logistic head using source membership labels, then frozen before fitting the final LR/MLP readout. These position-dependent weights are shared by all candidate images and stay fixed during target evaluation. The resulting representation has $2K$ coordinates.

The full profile flattens the position–band measurements directly for classification; learned aggregation first compresses each position's bands into a weighted energy summary. Both retain global error and reuse the same predictions. The frequency basis defines the measurement coordinates, while source learning determines how to use their membership evidence.

Their relative value depends on signal structure. In an independent-Gamma energy model with class-independent shapes and class-dependent rates, one suitable positive filter plus global energy per position preserves the likelihood ratio. With heterogeneous membership effects across coordinates, fixed rank-deficient linear aggregation can lose information. Appendix C, Propositions B1–B2, derives these complementary regimes.

Positions, repetitions, and model families. The inference budget is $B = \sum_{t \in \mathcal{T}} Q_t$. Source validation balances distinct predictor responses against more precise noise averages, reusing common noise indices across positions. All bands reuse each prediction, so decomposition adds no generator calls. An independent source nonmember subset calibrates the operational threshold; ROC evaluation varies the threshold of the frozen score. Section 4.1 specifies the tested schedules and budgets. For VP-FM and DDPM, PaFMIA uses the native velocity or noise-prediction residual and the same position–band construction (Section 4.4; Appendix A.2).

4 Experiments

4.1 Setup

Data and models. We evaluate unconditional RF on CIFAR-10, SVHN, TinyImageNet-64, and CelebA-64 (Krizhevsky, 2009; Netzer et al., 2011; Stanford CS231n, 2015; Liu et al., 2015). Each dataset has one source and two separately trained target U-Nets with the same training recipe. Each generator trains for 500,000 steps on 20,000 members drawn from a shared 40,000-image candidate population; membership assignments and training seeds differ across models. Evaluation uses EMA weights. CIFAR-10 and SVHN use 32×32 images; the other datasets use 64×64 images.

Attack fitting and evaluation. Source fitting and validation use 2,000 images each, with 4,000 separate source nonmembers for threshold calibration. Target testing uses 14,000 members and 14,000 nonmembers per model, excluding source attack-fitting and validation images. All attack parameters, schedules, normalization, and readout choices are selected on the source and frozen for targets. We report TPR at 1% FPR, averaged over the two targets. Additional metrics, operational thresholds, and per-target paired intervals appear in Appendix B.

Queries and readouts. PaFMIA supports arbitrary finite position sets. We evaluate source-selected subsets of up to 21 RF positions at budgets $B \in \{2, 4, 8, 16, 32, 64, 128\}$, distributing queries as evenly as possible across positions and reusing noise indices. An offline greedy search adds positions by source-validation gain; candidate sets are then compared using their budget-specific repeat allocations. For example, 32-query TinyImageNet uses eight positions with four repeats each. Twelve DCT bands plus global error yield $13K$ full-profile features, versus $2K$ for learned aggregation, without frequency-specific model calls. Source validation selects LR or a 16-unit MLP by partial AUROC up to 5% FPR. Same-schedule ablations isolate representation effects. Full training and search details appear in Appendix A.

Comparisons. We evaluate RF error, JPEG-calibrated RF error, FreMIA-RF, DIME-RF, and SimA-MC-RF (Rao & Moyer, 2026; Lian et al., 2026; Do & Alabi, 2026; Rao et al., 2026). Their

source searches cover RF21 at 32 inference queries, counting all required model calls. A stronger global-profile MLP with 64–32 hidden units independently selects its input transform and schedule, providing a learned multi-position comparison inspired by Sesmat et al. (2026). This comparator tests overall attack strength; matched-schedule global profiles in ablations and model extensions isolate representation effects on the same queried predictions. RF adaptations and fitting details appear in Appendices A.4 and A.3.

4.2 Main results on four datasets

Table 1: RF membership inference at 32 queries: TPR at 1% FPR (%), two-target means. All configurations are source-selected; bold marks the best result per dataset.

Attack	CIFAR-10	SVHN	TinyImageNet	CelebA
RF error	59.90	67.80	5.51	38.63
RF error + JPEG	80.39	92.32	17.68	79.64
FreMIA-RF	65.12	89.76	15.30	38.86
DIME-RF	88.35	94.69	19.78	73.12
SimA-MC-RF	16.25	51.38	4.03	16.68
Global-profile MLP	96.43	**95.98**	49.72	88.05
PaFMIA (learned aggregation)	**97.13**	95.75	65.69	**89.60**
PaFMIA (full profile)	97.00	95.26	**67.41**	89.26

Joint residual evidence strengthens membership inference. Both PaFMIA variants improve over the five external attacks on all four datasets (Table 1). The clearest gain is on TinyImageNet: the full profile reaches 67.41% TPR, compared with 19.78% for the strongest external attack, and adds 17.69 points over the independently selected global-profile MLP. The frequency gain remains positive for both targets at 32 and 128 queries (Appendix B), showing that learned global-error summaries leave useful residual structure untapped.

Compact versus complete profiles. Learned aggregation captures most of the improvement and leads on CIFAR-10 and CelebA at 32 queries. Retaining all bands is most useful on TinyImageNet, whereas SVHN is already well captured by global profiles. This variation motivates the complementary compression regimes in Section 3.4.

Query efficiency. Figure 3 shows that informative representations also reduce query cost: on TinyImageNet, four-query PaFMIA reaches 55.65% TPR, exceeding the 128-query independently selected global-profile MLP (53.20%). Gains taper at larger budgets. Within the plotted comparison, all representations share each budget's selected schedule and select their own LR/MLP readout on the source.

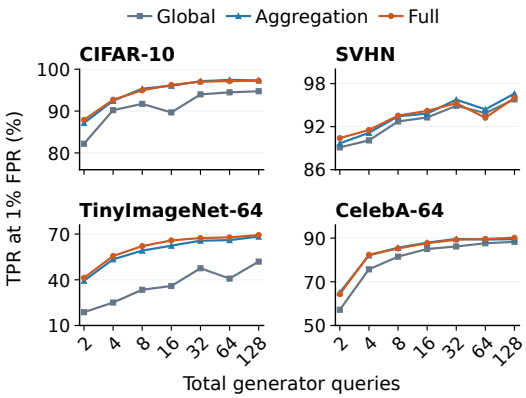


Figure 3: Query-budget comparison: TPR at 1% FPR (%), two-target means.

4.3 Ablations: positions, bands, and learned readouts

Table 2 fixes LR and the 32-query budget, varying position count and frequency resolution. One position receives 32 repeats; two positions receive 16 each.

Position and frequency contributions. Both dimensions contribute across all four datasets. On TinyImageNet, their combination raises TPR from 4.75% to 36.78%. Both targets' 95% paired bootstrap intervals show a positive two-position frequency increment. This comparison isolates the ingredients of source-selected PaFMIA.

Table 2: Position–frequency ablation: TPR at 1% FPR (%), two-target means; fixed LR and 32 queries. Search details: Appendix A.7.

	One position		Two positions	
Dataset	Global	Bands	Global	Bands
CIFAR-10	46.98	75.60	85.54	90.18
SVHN	44.14	86.90	93.51	94.69
TinyImageNet	4.75	15.42	7.68	36.78
CelebA	29.17	48.88	63.51	70.97

Representation and classifier effects. Table 3 compares LR and MLP at a common schedule. The frequency benefit appears with both readouts; nonlinear combinations add further gains on TinyImageNet. On the other datasets, their performance is often close. Together with the residualization result in Section 3.1, these comparisons tie the improvement to the information supplied to the classifier.

Table 3: Readout ablation at 32 queries: TPR at 1% FPR (%), two-target means. Agg. and Full denote the PaFMIA variants. Bold: best per dataset.

Dataset	Head	Global	Agg.	Full
CIFAR-10	LR	92.22	96.09	96.85
	MLP	93.98	**97.13**	97.00
SVHN	LR	94.74	**95.95**	94.95
	MLP	94.90	95.75	95.26
TinyImageNet	LR	38.03	56.47	57.73
	MLP	47.58	65.69	**67.41**
CelebA	LR	86.16	88.55	88.84
	MLP	86.69	**89.60**	89.26

Position selection. In the fixed eight-position mechanism panel, uniform positions and source-selected positions perform similarly (58.18% versus 57.91% on TinyImageNet). The joint representation is effective under both schedules (Appendix B.5).

4.4 Extension to VP-FM and DDPM

On CIFAR-10, we retain the backbone and source/target evaluation structure while replacing RF with VP-FM or DDPM. PaFMIA decomposes each model's native velocity or noise-prediction residual at positions matched by log signal-to-noise ratio. Both frequency representations improve on global profiles (Table 4), extending the benefit beyond straight-path RF to another flow path and a discrete diffusion objective. The complete 32-/128-query comparison appears in Appendix B.9.

Table 4: Model-family extension on CIFAR-10 at 32 queries: TPR at 1% FPR (%), two-target means, with matched schedules and source-selected heads. Bold: best per model.

Model	Global profile	PaFMIA (aggregation)	PaFMIA (full)
RF	93.98	**97.13**	97.00
VP-FM	91.46	96.93	**97.23**
DDPM	97.88	99.55	**99.57**

5 Related work

Membership inference and attack representations. Early membership attacks trained classifiers over model outputs (Shokri et al., 2017); later work strengthened calibration and low-false-positive evaluation (Carlini et al., 2022). These approaches demonstrate that both the available response and the attack statistic matter. PaFMIA studies the representation supplied to the attack classifier. A prediction residual contains spatial structure that a scalar error compresses; collecting residuals along a generative path further reveals how that structure changes. We characterize the membership information retained by these joint measurements and learn to exploit it with full-profile and compact representations.

Diffusion and flow attacks. Generative memorization has motivated extraction and membership inference (Carlini et al., 2023). SecMI uses step-wise prediction consistency (Duan et al., 2023),

PIA uses proximal initialization for efficient attacks (Kong et al., 2024), and quantile-regression attacks adapt thresholds to individual examples (Tang et al., 2024). MiO combines evidence across times for ownership verification (Jiang et al., 2026); SimA and DIME develop score-based and query-efficient membership statistics (Rao et al., 2026; Do & Alabi, 2026). For RF, Rao & Moyer (2026) study prediction-error membership signals, while Sesmat et al. (2026) analyze leakage along the flow path and learn from multi-position global-error profiles. PaFMIA resolves the spatial-frequency composition of those errors at each position. Comparisons with learned global profiles and established attack statistics quantify the additional evidence exposed by this representation.

Spatial frequency and generative paths. FreMIA filters spatial-frequency components of diffusion residuals to improve membership inference (Lian et al., 2026). PaFMIA examines how frequency-dependent membership evidence changes along the generative path. It retains the joint position–band structure and studies when learned aggregation captures its evidence and when complete profiles provide further gains. The position dimension follows the straight interpolation of RF, the broader probability paths of flow matching, and the discrete noise levels of diffusion (Liu et al., 2023; Lipman et al., 2023; Ho et al., 2020). Their native prediction tasks provide a common entry point for the residual representation. Our information-compression analysis, RF mechanism, and controlled experiments explain the role of both dimensions.

6 Discussion

Our results establish structured residual measurements as an effective basis for membership inference. Frequency decomposition exposes how error is allocated within an image; observations at multiple positions reveal how that allocation changes along the generative path. PaFMIA preserves both dimensions before classification. The fixed-readout and residualized-band experiments show that the resulting gains reflect useful information beyond the global-error profile. The structured-weight comparisons further identify how position-dependent frequency use contributes to the joint representation.

The largest frequency gains occur on TinyImageNet, where global error is least effective at distinguishing members. All four datasets use 20,000 training images; the contrast concerns their response structure at a common training-set size. TinyImageNet spans more diverse object categories and visual content, suggesting more heterogeneous residual allocations and path responses. The residualization experiment supports this interpretation: frequency features left after linear prediction from global errors substantially improve its attack, while changing little on SVHN. CIFAR-10 and SVHN also leave less improvement headroom because their strong global attacks are already near 96% TPR. Learned aggregation captures most of TinyImageNet's frequency gain, with complete profiles providing an additional increment. These patterns connect the compression regimes in Section 3.4 to the useful membership structure left unresolved by global error.

The evaluation uses unconditional models, native prediction access, and labeled source membership data. Source and target models share a candidate population; earlier exploratory analyses included these targets, while the reported comparisons select and freeze attack choices on source data before target evaluation. Disjoint-population shadow models and conditional generators offer further settings for studying how the learned residual representation transfers. Services exposing only generated samples require a different measurement interface.

7 Conclusion

PaFMIA exposes membership evidence in the spatial-frequency composition of prediction residuals and its variation along the generative path. Complete profiles and learned aggregation make this structure available to the attack without additional frequency-specific queries. Four-dataset experiments, controlled ablations, and flow/diffusion extensions demonstrate the value of this representation; the analysis explains its gains through information preservation and complementary predictor responses.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of parts of the experimental code and with mathematical derivations and proofs. We did not use generative AI tools to generate synthetic datasets, and other required-disclosure tasks that are not involved in this work are not applicable. Additionally, we used generative AI tools to improve the readability and presentation of the manuscript, to search for and identify relevant literature, and to assist with LaTeX formatting and code editing. All AI-assisted outputs were reviewed and verified by the authors. AI-assisted code was manually inspected, executed, and tested for correctness before being used in our experiments. Mathematical derivations and proofs assisted by AI were independently checked by the authors. References and related work identified with AI assistance were verified against the original publications. All AI-assisted textual and LaTeX edits were reviewed and revised by the authors before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The experiments study models trained for this evaluation on public image benchmarks. Membership inference is used to characterize privacy exposure, not to identify individuals or infer membership in deployed services. Stronger attacks can improve auditing and also carry misuse risks; the intended use of the released methodology is controlled evaluation with appropriate authorization. CelebA is used as an image-distribution benchmark, and no conclusions are drawn about any person's participation in external training datasets.

REPRODUCIBILITY STATEMENT

Appendix A details the model-training, data-splitting, query, feature-extraction, and attack-selection protocols. We provide an anonymous supplementary package containing the PaFMIA implementation, source-selected configurations, dataset split indices, and cached per-example attack scores. CPU scripts recompute the main experimental tables and query-budget curves from these scores, with additional code for paired-bootstrap evaluation and mechanism visualization. The accompanying README maps paper results to executable commands and documents the dependencies and procedures for data preparation, model training, feature extraction, and attack fitting.

REFERENCES

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649. URL https://doi.org/10.1109/SP46214.2022.9833649.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Tue Do and Daniel Alabi. DIME: Query-efficient framework for membership inference on diffusion models, 2026. URL https://arxiv.org/abs/2608.22824.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8717–8730. PMLR, 2023. URL https://proceedings.mlr.press/v202/duan23b.html.

Mads Dyrholm, Christoforos Christoforou, and Lucas C. Parra. Bilinear discriminant component analysis. *Journal of Machine Learning Research*, 8:1097–1111, 2007. URL https://jmlr.org/papers/v8/dyrholm07a.html.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Feng Jiang, Zuobin Xiong, An Huang, Zhipeng Cai, and Yingshu Li. Membership is ownership: A robust ownership verification framework for diffusion models, 2026. URL https://arxiv.org/abs/2608.28929. Accepted to the IEEE International Conference on Data Mining (ICDM), 2026.

Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rpH9FcCEV6.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Puwei Lian, Yujun Cai, Songze Li, and Bingkun Bao. Enhancing membership inference attacks on diffusion models from a frequency-domain perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QZlNaWuxO6.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL https://research.google/pubs/reading-digits-in-natural-images-with-unsupervised-feature-learning/.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://doi.org/10.1098/rsta.1933.0009.

Mingxing Rao and Daniel Moyer. Generalization and memorization in rectified flow. In *Computer Vision – ECCV 2026*, volume 17016 of *Lecture Notes in Computer Science*, pp. 537–554, Cham, 2026. Springer. doi: 10.1007/978-3-032-37383-0_31. URL https://link.springer.com/chapter/10.1007/978-3-032-37383-0_31.

Mingxing Rao, Bowen Qu, and Daniel Moyer. Score-based membership inference on diffusion models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Ckvsu5xRmf.

Thomas Sesmat, Gabriel Meseguer-Brocal, and Geoffroy Peeters. Where rectified flows leak: Characterising membership signals along the interpolation path. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Ty5X41WbJw.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41. URL https://doi.org/10.1109/SP.2017.41.

Stanford CS231n. Tiny imagenet challenge, 2015. URL https://cs231n.stanford.edu/2015/project.html. Official course challenge webpage.

Shuai Tang, Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47819–47829. PMLR, 2024. URL https://proceedings.mlr.press/v235/tang24g.html.

A Experimental details

A.1 Data, models, and attack roles

CIFAR-10 and SVHN use 32×32 RGB images; TinyImageNet and CelebA use 64×64 RGB images. Each setting uses a fixed candidate population of 40,000 images. Each generator is trained on 20,000 members, with the other 20,000 serving as nonmembers. Source and target membership assignments and training seeds differ, but the candidate population is shared: target model training data are not disjoint from the source model's training data.

Image identities define global attack roles. Source fitting and validation use 2,000 images each, while source nonmember calibration uses 4,000 images. Target testing contains 14,000 members and 14,000 nonmembers. The target test set has zero overlap with source attack fitting and validation images. A separate target-nonmember calibration subset is used only for the explicitly labeled recalibration setting. Model training membership and attack-fitting membership are different concepts: the attack learns from residual measurements of designated images, while the generator was trained on its full member subset.

We use the U-Net implementation vendored from Rao and Moyer's official rf_gen_mem code, revision d86af325, derived from Phil Wang's rectified-flow-pytorch. All generators use this U-Net with base dimension 128, dimension multipliers $[1, 2, 2, 2]$, three image channels, two residual streams, four attention heads of dimension 32, full attention in the final two stages, and dropout 0.1. Training uses batch size 64, Adam with learning rate 10^{-4}, betas $(0.9, 0.999)$, epsilon 10^{-8}, no weight decay, and no learning-rate scheduler. Gradients are clipped to norm 0.5; training uses FP32 and no image augmentation. All models run for 500,000 steps (1,600 nominal image exposures per training member). The RF loss is mean-squared velocity error with uniform sampling of interpolation coordinates and standard Gaussian noise. Evaluation uses EMA weights with decay 0.9999, updates beginning after step 100 and every 10 steps, with the implementation's recorded warmup.

CIFAR-10, SVHN, and TinyImageNet use their original training image pools without resizing or cropping in the frozen adapters. Images are decoded to float32 in $[0, 1]$, then normalized once to $[-1, 1]$. CelebA uses official partition 0, converts to RGB, crops pixel box $[0, 20, 178, 198]$, and bilinearly resizes to 64×64. Its evaluation class is set to a single class; face attributes are not used for class-stratified attack training. Cohorts and preprocessing manifests remain fixed across models.

The main RF analysis has four sources and eight target models. Model-family extensions use one source and two targets each for VP-FM and DDPM on CIFAR-10. The targets participated in earlier exploratory development; configurations for the reported comparisons are selected on source data and frozen for target evaluation.

A.2 Residual energies and query schedules

The normalized radius of spatial DCT coefficient (u, v) is

$$\rho(u,v) = \frac{1}{\sqrt{2}}\sqrt{\left(\frac{u}{H-1}\right)^2 + \left(\frac{v}{W-1}\right)^2}.$$

Its zero-based band index is $\min(\lfloor 12\rho \rfloor, 11)$. These are equal-width radial intervals, not equal-count bins; channelwise coefficients use the same partition. Equations (4)–(5) count all channels in d_b.

The RF interpolation-coordinate grid is

$$\{.047, .076, .119, .182, .269, .321, .378, .418, .450, .475, .500,$$
$$.525, .550, .582, .622, .679, .731, .818, .881, .924, .953\}.$$

The initial source scout uses four noise samples at every grid point, or 84 generator queries per source image. Starting from a single query position, a greedy search appends the coordinate yielding the best source-validation partial AUROC. Candidate prefixes are then evaluated with their actual budget-specific noise allocations. Additional source-prefix measurements used for schedule selection are offline development cost, separate from B at inference.

For VP-FM and DDPM, define $s = 1 - t$, $I(s) = 0.1s + (20 - 0.1)s^2/2$, $\alpha = e^{-I/2}$, and $\sigma = \sqrt{1 - e^{-I}}$. The path is $x_t = \alpha x + \sigma z$; VP-FM predicts $\alpha' x + \sigma' z$, whereas DDPM predicts z. The saved attack residuals are native velocity and noise-prediction residuals, respectively. Here t indexes the model's prescribed noise-to-data path; it does not count numerical solver steps. RF-grid log-SNR $2\log(t/(1-t))$ is mapped to the nearest of 1,000 VP/DDPM indices. This matches noise levels rather than numerically identical native coordinates. VP-FM training samples continuous $t \in [0, .999)$; DDPM uses 1,000 discrete diffusion timesteps.

Band energies average squared residuals **before** noise averaging. They are not the squared norm of an averaged residual. Every frequency view reuses the same prediction and the same noise realization. Shared query indices correspond to the same stateless noise across path positions; longer prefixes extend the same sequence.

Schedules use feasible $K \in \{1, 2, 3, 4, 8, 16, 21\}$ and $B \in \{2, 4, 8, 16, 32, 64, 128\}$. Each selected path position receives either $\lfloor B/K \rfloor$ or $\lceil B/K \rceil$ queries. Source fitting and validation choose the schedule through a band-feature LR scout; each feature family then selects its final LR or MLP readout on the source. Thus the same-schedule controls isolate representation effects at the schedule selected for PaFMIA (full profile), while the independently selected global-profile MLP checks a separate global-error schedule search.

For the original three RF datasets, the candidate pool includes global-error, band, and anchored-band greedy paths. CelebA uses global-error and band greedy paths; its previously frozen 32- and 128-query configurations are retained. These protocols match final query budgets and source-selection principles, but their candidate pools are not identical.

A.3 Readout fitting and filtering

For the main readouts, feature means and standard deviations are fitted only on source fitting images; standard deviations are floored at 10^{-6}. Regularization is selected from $\{10^{-4}, 10^{-3}\}$. Logistic regression uses L-BFGS-B with at most 2,000 iterations. The MLP uses 16 hidden tanh units and Adam with learning rate 0.003, at most 2,000 epochs, validation every 20 epochs, and 200-epoch patience. The objective is mean binary cross-entropy plus $\lambda\|\mathbf{W}\|_2^2/2$, excluding biases. Source validation chooses normalized partial AUROC up to 5% FPR; the final LR/MLP tie rule favors LR.

PaFMIA (learned aggregation), also called the learned filter, produces one nonnegative weighted energy sum per path position, together with global error. Its weights are $a_{t,b} = \mathrm{softmax}_b(\log d_b + \theta_{t,b})$. The source optimization jointly fits θ and an LR head by binary cross-entropy with LR weight penalty, a second-difference smoothness penalty on θ along frequency, and a small quadratic θ penalty. The search uses LR penalties $\{.1, .01, .001, .0001\}$, smoothness weights $\{.1, .001\}$, and two initializations, with at most 2,400 iterations per candidate and up to 8,000 refinement iterations for the source-validation selection. The extra quadratic coefficient is 10^{-4}, with the implementation's half-mean-square convention. Source-fitted filter weights may vary with path position and are

then frozen while selecting the final LR/MLP head. Thus the MLP row is not an end-to-end filter–MLP optimum. Additional fixed-MLP capacity controls set global-error and filter hidden widths to approximately match the parameter count of the band MLP; these are separate diagnostic comparisons and are not extra candidates in the final LR/MLP selector.

The global-profile MLP uses 64–32 hidden tanh layers and source-fit normalization. For budgets 32 and 128, source validation jointly selects its input transform (raw global-error means, their $\log(G_t + 10^{-12})$ values, or the concatenation), regularization, stopping epoch, and query schedule. It searches the union of the existing source-defined global-error, band, and anchored-band greedy candidate paths where available; CelebA uses global-error and band paths. These are independently scored candidate schedules, not newly generated greedy paths for each transform. Query allocations use measured noise prefixes. All source choices are frozen before target evaluation; the transform with the best target TPR is not selected retrospectively. The source-winning configuration within each transform is also reported in Appendix B.3.2. This is a budget-matched global-profile control inspired by Sesmat et al.; the original 11-position, 100-noise-per-position configuration would require 1,100 queries and is not used here.

A.4 External attack implementations

The external panel contains Monte Carlo RF prediction error, its JPEG-based image-complexity calibration, FreMIA's frequency-filtering construction, DIME-RF, and SimA-MC-RF. Existing methods retain their attribution. Complexity calibration is used as a baseline, not a contribution of the proposed representation.

The total budget counts all generator evaluations. In particular, DIME's base query is included: budgets 2 and 32 leave 1 and 31 additional reference queries. Source search considers the implemented DIME parameter grid; FreMIA uses its square-mask construction rather than the radial bands of the proposed representation. Each attack is evaluated with a configuration fixed from the source; target labels do not select its query position or attack parameters.

At 32 queries, all four datasets use the complete 21-coordinate RF grid for source selection of the external attacks. The CIFAR-10, SVHN, and TinyImageNet results use the completed RF21 rerun; the CelebA evaluation already used RF21. Each method retains its own published statistic and the stated RF adaptation. The historical two-query comparison in Appendix B.2 uses four candidate coordinates on the first three datasets and RF21 on CelebA. SecMI, PIA, MiO, and quantile-regression attacks are discussed as related work rather than evaluated rows.

A.5 Metrics, thresholds, and uncertainty

A larger score indicates stronger evidence of membership. ROC metrics are computed using the frozen attack score. Operational thresholds use the 99th percentile of source calibration nonmember scores, with the quantile's *higher* convention and the strict decision $s > \tau$. The target-OUT setting recomputes only the threshold from designated target nonmembers; it does not refit the score.

Each paired interval uses 2,000 shared bootstrap resamples stratified by class and membership; CelebA uses a single evaluation class. Intervals are percentile intervals for a fixed target model and attack. Reported means average two target models, while per-target intervals remain separate. The intervals are pointwise and are not adjusted for simultaneous inference across every experiment. Appendix B.3 distinguishes ranking metrics from realized threshold TPR and FPR.

A.6 Mechanism panel

Mechanism comparisons use eight path positions, four noise repetitions per position, and LR-style readouts. The source-selected and uniform-index position sets are evaluated at the same total budget. Frequency residualization fits a pooled linear projection of log-band features on the log-global profile using source-fitting data only; the frozen transformation is applied to targets. This removes linearly predictable components, not all possible nonlinear dependence.

Structured readouts include the global profile and constrain the band-weight matrix across path positions: shared frequency weights; a rank-one factorization; or unrestricted weights. Coordinate scaling is kept consistent with the constraints. These fixed-position comparisons isolate the effect of

weight-sharing constraints; the final attack separately selects its schedule and LR/MLP readout for each budget.

The mechanism panel covers the three original RF datasets and CIFAR-10 VP-FM/DDPM. CelebA contributes the completed main, budget, and external evaluations; a full CelebA mechanism replication is not assumed.

A.7 Factorial position and frequency comparison

The one-versus-two-position comparison uses source-frozen LR at a fixed total budget of 32 queries. Its four representations are global energy at one position, global-plus-band energies at one position, global energies at two positions, and global-plus-band energies at two positions. Source fitting and validation choose the configurations; both separately trained targets are then evaluated with those choices fixed. The single-position design uses all 32 noise repetitions at one coordinate, whereas the two-position design uses 16 repetitions per coordinate. Thus the comparison trades repetition precision for a second position without changing the inference budget. This factorial panel is separate from the final attack's budget-specific selection over multiple position counts and LR/MLP readouts.

For CIFAR-10, SVHN, and TinyImageNet, the historical source search enumerates single positions and pairs over nine candidate coordinates. CelebA reuses its previously frozen band-LR greedy source search: four noise samples at each of the 21 grid positions selected index 8 for one position and indices 8 and 19 for two positions, corresponding to $t = 0.450$ and $(0.450, 0.924)$. The subsequent four-cell fit uses the same source-fit normalization, log-energy feature definitions, LR implementation, regularization grid $\{0.1, 0.01, 0.001, 0.0001\}$, source role sizes, and 32-query allocation as the earlier panel. Thus inference cost and readout protocol are aligned across datasets, while their source position-search histories differ. CelebA uses existing residual measurements and adds no generator queries. Within each target, 2,000 class-by-membership-stratified paired bootstrap replicates compare the two-position band profile with the other three cells.

A.8 Supplementary mechanism illustrations and shared-noise evidence

Figure 2 uses the RF21 grid with four shared-noise draws per position, totaling 84 characterization queries per image. The plotted value is mean log-band energy for members minus that for nonmembers, averaged over two target models. This characterization measures the joint structure over the full grid; attack schedules are selected on source validation.

Fixed-budget shared-difficulty model. Write

$$Y_t = a_t m + U + \eta_t, \qquad \mathrm{Var}(U) = \tau^2, \quad \mathrm{Var}(\eta_t) = \sigma^2/Q_t. \tag{10}$$

Here U is centered Gaussian image difficulty independent of membership and query noise. The query noises are centered jointly Gaussian with membership-independent laws, $\sigma > 0$, and matched-noise correlation $|\rho| < 1$. Repeating one position reduces query noise but retains U; subtracting positions cancels U while retaining mean contrast $(a_{t_2} - a_{t_1})m$. For an even budget B, allocating $B/2$ queries to each of two positions beats the best B-query single-position observation in the finite candidate set whenever

$$\frac{B(a_{t_2} - a_{t_1})^2}{4\sigma^2(1-\rho)} > \frac{\max_t a_t^2}{\tau^2 + \sigma^2/B}. \tag{11}$$

Large shared variation, changing membership responses, and correlated query noise can favor the contrast. The complete two-position representation also retains the common response direction. Appendix C.3, Propositions G2–G3, provides the proof, the exact joint optimum, and the vector extension to frequency bands.

Shared-noise experiment. In the earlier two-query difference experiment, shared versus independent noise gives 87.30% versus 84.67% TPR at 1% FPR on CIFAR-10, 88.46% versus 75.90% on SVHN, and 24.03% versus 15.56% on TinyImageNet. The smaller shared-noise advantage at 32 queries is consistent with repetitions reducing perturbation variability.

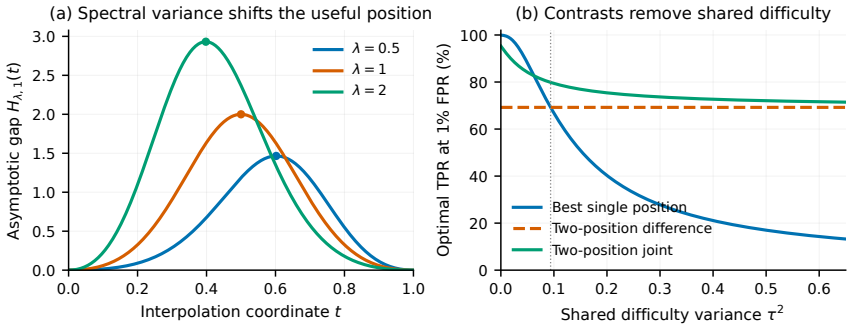


Figure 4: Analytic mechanisms. Left: spectral variance shifts the linear-RF membership-gap profile. Right: contrasts cancel shared difficulty and can improve fixed-budget detection in the Gaussian model.

B Additional results

Throughout this appendix, *full profile* and *learned aggregation* denote the two PaFMIA implementations. Global profile and global-profile MLP are comparison attacks.

B.1 Full query-budget results

TPR at 1% FPR (%), two-target means. The position count belongs to the common schedule used by the global-profile comparator and both PaFMIA variants.

Table 5: Full query-budget results: TPR at 1% FPR (%), averaged over two target models.

Queries	Position count	Global profile	Learned aggregation	PaFMIA (full profile)
CIFAR-10				
2	2	82.21	87.15	87.90
4	4	90.22	92.46	92.71
8	4	91.72	95.34	94.95
16	4	89.67	96.05	96.23
32	4	93.98	97.13	97.00
64	4	94.47	97.44	97.16
128	4	94.73	97.31	97.25
SVHN				
2	2	89.11	89.66	90.40
4	4	90.08	91.14	91.53
8	2	92.72	93.44	93.53
16	2	93.27	93.81	94.20
32	3	94.90	95.75	95.26
64	2	93.89	94.40	93.24
128	3	95.79	96.60	96.02

Continued on next page

Table 5 (continued)

Queries	Position count	Global profile	Learned aggregation	PaFMIA (full profile)
TinyImageNet-64				
2	2	18.73	39.40	41.27
4	4	25.05	53.44	55.65
8	4	33.46	59.16	62.11
16	4	35.94	62.39	65.86
32	8	47.58	65.69	67.41
64	4	40.81	66.04	67.85
128	8	51.96	68.36	69.39
CelebA-64				
2	2	57.20	65.08	64.32
4	4	75.69	82.09	82.39
8	3	81.52	85.61	85.25
16	4	85.00	87.90	87.67
32	4	86.16	89.60	89.26
64	4	87.62	89.33	89.62
128	4	88.29	89.49	90.18

B.2 Two-query external comparison

Table 6: Historical two-query external comparison: TPR at 1% FPR (%), averaged over two target models. The first three datasets use a four-coordinate source search; CelebA uses RF21.

Dataset	RF error	RF error + JPEG	FreMIA RF	DIME RF	SimA-MC RF	PaFMIA (full profile)
CIFAR-10	49.06	72.13	63.00	77.70	7.64	87.90
SVHN	25.27	46.79	82.86	80.05	12.35	90.40
TinyImageNet-64	3.46	14.44	13.77	18.49	1.94	41.27
CelebA-64	30.36	20.86	36.99	36.61	4.04	64.32

B.3 Additional metrics and operational thresholds

Two independent targets, arithmetic mean; values in percent. AUROC and ROC TPR differ from fixed-threshold operating points. The original result labels used below are `full` for the matched-schedule global-profile comparator, `profile` for the independently selected global-profile MLP, and `filter` for PaFMIA (learned aggregation). The row labeled *Full profile* denotes PaFMIA (full profile).

B.3.1 Ranking metrics

Table 7: Ranking metrics (%): arithmetic means over two independent target models.

Budget	Readout	AUROC	TPR at 1% FPR	TPR at 0.1% FPR
CelebA-64				
32	full	99.220	86.164	24.232
32	profile	99.258	88.046	26.646
32	filter	99.363	89.604	28.693
32	Full profile	99.385	89.264	29.679
128	full	99.323	88.293	26.821
128	profile	99.291	88.393	26.011
128	filter	99.406	89.489	31.593
128	Full profile	99.437	90.179	32.675
CIFAR-10				
32	full	99.762	93.982	74.261
32	profile	99.844	96.429	80.771
32	filter	99.836	97.132	78.993
32	Full profile	99.848	97.004	83.950
128	full	99.785	94.732	75.986
128	profile	99.865	97.050	84.157
128	filter	99.852	97.314	78.629
128	Full profile	99.855	97.246	83.450
SVHN				
32	full	99.775	94.900	79.182
32	profile	99.789	95.982	78.621
32	filter	99.817	95.754	81.018
32	Full profile	99.808	95.257	79.721
128	full	99.800	95.793	79.961
128	profile	99.829	96.511	83.268
128	filter	99.848	96.604	81.939
128	Full profile	99.839	96.018	81.757
TinyImageNet-64				
32	full	91.562	47.575	23.611
32	profile	92.225	49.718	30.407
32	filter	96.254	65.686	46.007
32	Full profile	96.311	67.407	49.118
128	full	92.516	51.961	30.336
128	profile	93.460	53.204	31.646
128	filter	96.588	68.364	53.604
128	Full profile	96.571	69.393	54.886

B.3.2 GLOBAL-PROFILE INPUT TRANSFORMATIONS

Each column below evaluates the source-winning schedule and readout within the named transform. The final transform is selected by source partial AUROC, not by these target outcomes. Thus the selected column need not maximize target TPR. These results replace the earlier raw-only, full-greedy global-profile control in the main comparison.

B.3.3 OPERATIONAL THRESHOLDS

Source-only: threshold fixed from source nonmembers. Recalibration: independent target nonmembers, with no target members, used to reset threshold. Both columns are actual target test TPR/FPR; nominal FPR is 1%.

Table 8: Global-profile MLP transform comparisons: two-target mean TPR at 1% FPR (%). Source selection ranges over the common candidate-schedule union.

Dataset	Budget	Raw	Log	Raw + log	Source pick
CIFAR-10	32	96.59	96.43	96.59	log
CIFAR-10	128	96.24	97.05	97.08	log
SVHN	32	95.98	95.91	95.83	raw
SVHN	128	96.78	96.51	96.64	log
TinyImageNet-64	32	50.14	49.23	49.72	raw + log
TinyImageNet-64	128	52.93	53.20	53.21	log
CelebA-64	32	87.35	88.05	87.44	log
CelebA-64	128	88.64	89.11	88.39	raw + log

Table 9: Operational threshold results (%): actual target test TPR and FPR at a nominal 1% FPR.

Budget	Readout	Source-only TPR	Source-only FPR	Recalibrated TPR	Recalibrated FPR
CelebA-64					
32	full	87.321	1.111	89.621	1.339
32	filter	90.975	1.146	91.625	1.275
32	Full profile	90.389	1.125	91.900	1.346
128	full	90.286	1.229	91.446	1.375
128	filter	91.354	1.171	92.907	1.414
128	Full profile	91.864	1.207	92.875	1.364
CIFAR-10					
32	full	93.004	0.850	94.454	1.071
32	filter	95.904	0.725	96.871	0.911
32	Full profile	95.889	0.696	96.714	0.939
128	full	93.793	0.843	94.582	0.971
128	filter	96.864	0.836	97.057	0.914
128	Full profile	96.832	0.825	97.264	0.996
SVHN					
32	full	95.479	1.161	94.318	0.943
32	filter	96.425	1.200	95.482	0.929
32	Full profile	93.993	0.725	95.211	0.979
128	full	96.636	1.318	95.621	0.993
128	filter	97.546	1.368	96.786	1.036
128	Full profile	94.086	0.568	96.418	1.132
TinyImageNet-64					
32	full	48.729	1.154	47.200	0.964
32	filter	65.714	1.007	64.400	0.854
32	Full profile	67.379	1.007	65.814	0.807
128	full	52.993	1.096	52.089	0.996
128	filter	69.043	1.121	67.025	0.804
128	Full profile	69.914	1.093	68.757	0.911

B.4 PER-TARGET PAIRED DIFFERENCES

All differences below are PaFMIA (full profile) minus comparator, in percentage points. The column labels `full`, `filter`, and `profile` refer respectively to the matched-schedule global profile, PaFMIA (learned aggregation), and independently selected global-profile MLP. Global-profile intervals use the source-selected raw/log/concatenated configuration, with identities aligned before 2,000 paired resamples. The global-error and filter comparisons retain their original matched-schedule evaluations.

Table 10: Per-target paired differences: PaFMIA (full profile) minus comparator, in percentage points, with 95% confidence intervals.

Budget	Target	vs. full Δ [95% CI]	vs. filter Δ [95% CI]	vs. profile Δ [95% CI]
CelebA-64				
32	102	2.81 [0.47, 4.32]	−0.18 [−1.98, 0.83]	1.04 [−1.21, 2.62]
32	103	3.39 [1.71, 5.42]	−0.50 [−1.87, 0.47]	1.40 [−0.26, 2.82]
128	102	1.68 [0.22, 3.10]	0.34 [−0.65, 1.20]	1.69 [0.13, 2.84]
128	103	2.09 [0.87, 3.99]	1.04 [−0.28, 2.56]	1.89 [0.61, 3.33]
CIFAR-10				
32	102	2.76 [1.84, 3.52]	−0.07 [−0.70, 0.31]	0.56 [−0.22, 1.01]
32	103	3.28 [2.27, 4.04]	−0.19 [−0.54, 0.18]	0.59 [0.16, 1.01]
128	102	2.38 [1.74, 3.26]	−0.01 [−0.41, 0.48]	0.19 [−0.41, 0.61]
128	103	2.65 [1.75, 3.31]	−0.13 [−0.43, 0.16]	0.20 [−0.16, 0.69]
SVHN				
32	102	0.31 [−0.31, 1.43]	−0.69 [−1.19, −0.22]	−0.94 [−1.58, −0.13]
32	103	0.41 [−0.05, 0.85]	−0.30 [−0.69, 0.19]	−0.51 [−1.04, 0.02]
128	102	0.06 [−0.54, 0.80]	−0.81 [−1.42, −0.27]	−0.71 [−1.38, −0.21]
128	103	0.39 [−0.12, 0.84]	−0.36 [−0.80, 0.04]	−0.28 [−0.81, 0.26]
TinyImageNet-64				
32	102	19.62 [17.47, 22.07]	1.55 [0.46, 2.98]	17.62 [15.84, 19.71]
32	103	20.04 [18.16, 21.64]	1.89 [0.61, 2.62]	17.76 [15.81, 19.51]
128	102	17.69 [15.76, 19.44]	1.31 [0.02, 2.34]	16.14 [13.86, 18.29]
128	103	17.17 [15.19, 19.06]	0.75 [−0.44, 1.81]	16.24 [13.88, 18.13]

B.5 Additional mechanism results

Fixed eight-position, 32-query LR mechanism panel; percentages are two-target means. The full-plus-residual representation preserves the unregularized linear span of full-plus-band features, so its numerical differences from free weights also reflect regularization.

Table 11: Fixed eight-position, 32-query LR mechanism panel (%): two-target means.

Setting	Global	Global + residuals	Shared	Rank one	Free	Uniform-index free
RF / CIFAR-10	91.07	93.79	92.86	96.21	96.51	94.83

Continued on next page

Table 11 (continued)

Setting	Global	Global + residuals	Shared	Rank one	Free	Uniform-index free
RF / SVHN	94.85	94.73	94.76	94.46	95.15	94.69
RF / TinyImageNet-64	38.03	57.07	42.18	54.75	57.91	58.18
VP-FM / CIFAR-10	94.75	94.90	95.55	96.91	96.55	96.09
DDPM / CIFAR-10	97.88	98.57	98.80	99.59	99.65	99.05

Residual-only classification is weak in the TinyImageNet panel (6.18% TPR), whereas combining the residuals with the global profile reaches 57.07%. The useful evidence is therefore the joint representation. Source-fitted residualization removes linear prediction from global features, not all nonlinear dependence.

The earlier shared-noise comparison fixes two source-selected interpolation coordinates and compares shared query indices with independent adjacent noise indices. At two queries, shared noise gives a positive paired difference in all six RF target tests. At 32 queries, the mean improvements shrink to 0.52, 0.23, and 1.44 percentage points on CIFAR-10, SVHN, and TinyImageNet, respectively, and most paired intervals cross zero.

A historical exploratory TinyImageNet control compared radial DCT groups with one fixed random grouping at 96 queries: 36.50% versus 10.72% TPR. This concerns one earlier model and one grouping, so it is supporting exploratory evidence rather than a replicated target-level result.

B.6 Previously evaluated compact learned summaries

An earlier multiple-position fusion panel considered one, two, or four softmax frequency readouts shared across path positions on standardized log-band features, with additional two-/four-readout sample-adaptive gates across path positions. These summaries differ from the main energy filter: they pool after logarithms and impose shared frequency directions. The source-selected pooling/gating candidate was evaluated on the two targets at budgets 32 and 128. On TinyImageNet, it achieved mean TPRs of 54.67% and 55.81%, compared with 65.49% and 67.80% for the same panel's unconstrained band MLP16. This panel uses its own fixed fitting protocol, distinct from the final source-selected readouts. Extended optimization was subsequently evaluated on the source only. The comparison separates the effects of summary form, weight sharing across positions, and optimization.

B.7 Readout ablations at matched schedules

Each row below uses the same measured residuals and query schedule for the global-profile comparator and both PaFMIA representations. The source independently selects regularization for each fixed LR or MLP readout. These fixed-head results complement the main table, whose head is itself chosen on source validation.

B.8 One versus two positions with and without frequency resolution

The fixed-budget LR comparison in Table 13 directly separates position count and frequency resolution. All four datasets use source-frozen LR and the same query allocation. The first three use the historical nine-coordinate source scan; CelebA reuses its RF21/Q4 greedy schedule, as detailed in Appendix A.7. The target mean excludes the source model.

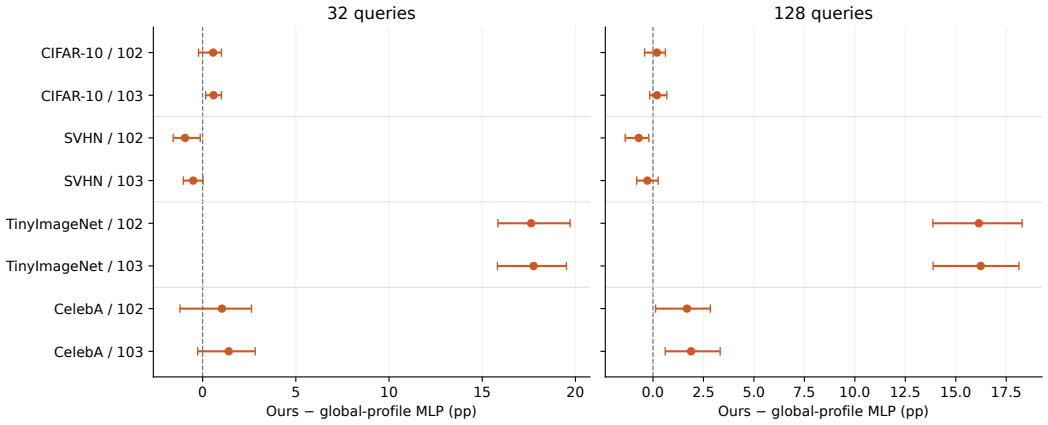


Figure 5: PaFMIA (full profile) minus independently selected global-profile MLP at 32 and 128 queries. Bars are within-target 95% paired percentile bootstrap intervals, not intervals on the two-model mean.

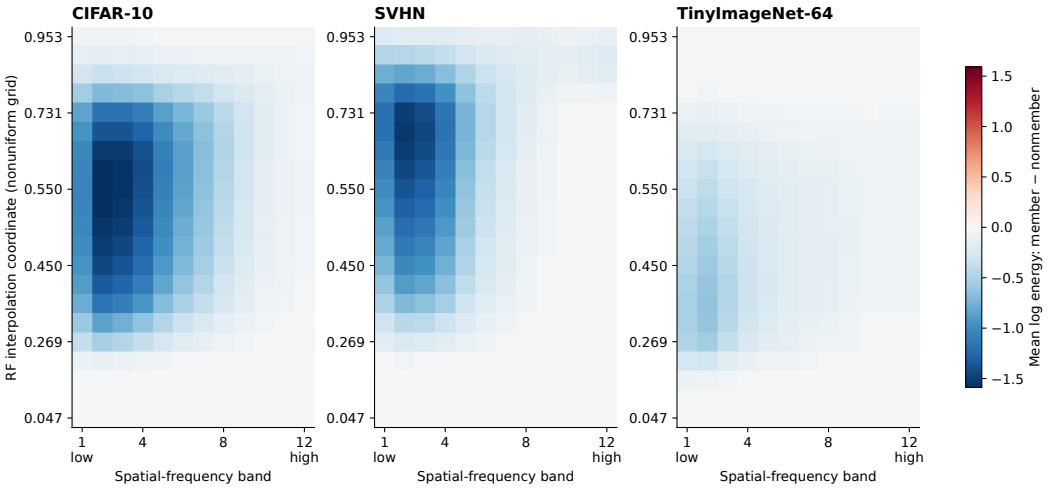


Figure 6: Measured mean log-band error for members minus nonmembers, averaged across two targets. Negative values indicate lower member error. Rows use the nonuniform RF21 grid and columns are twelve radial spatial-frequency bands. This descriptive analysis over the full coordinate grid uses four noises at each of 21 positions (84 nominal characterization queries); it is separate from the 32-query attack results. A common color scale is used across panels. These are observed mean contrasts, not classifier coefficients or estimates of conditional information.

Table 12: TPR at 1% FPR (%), two-target means. LR and MLP use matched schedules; MLP has 16 hidden tanh units. Learned aggregation and full profile are the two PaFMIA variants.

Dataset	B	Global profile		Learned aggregation		Full profile	
		LR	MLP	LR	MLP	LR	MLP
CIFAR-10	32	92.22	93.98	96.09	97.13	96.85	97.00
CIFAR-10	128	93.05	94.73	96.70	97.31	97.25	97.19
SVHN	32	94.74	94.90	95.95	95.75	94.95	95.26
SVHN	128	95.08	95.79	96.32	96.60	96.17	96.02
TinyImageNet-64	32	38.03	47.58	56.47	65.69	57.73	67.41
TinyImageNet-64	128	42.40	51.96	60.44	68.36	60.66	69.39
CelebA-64	32	86.16	86.69	88.55	89.60	88.84	89.26
CelebA-64	128	87.16	88.29	89.61	89.49	89.59	90.18

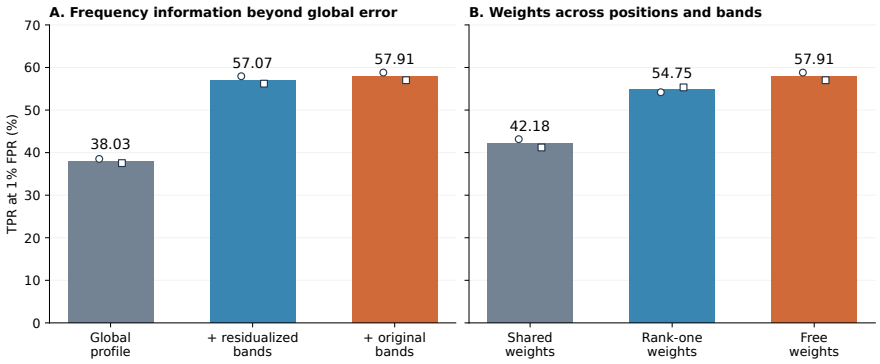


Figure 7: TinyImageNet mechanism ablations at eight positions, 32 total queries, and LR-style readouts. Left: band components remaining after source-fitted linear adjustment for global error. Right: increasingly flexible weights across positions and frequency bands. Means and individual target values are shown.

Table 13: Factorial comparison at $B = 32$: TPR at 1% FPR (%), two-target means. All four entries use LR; two positions receive 16 queries each.

	One position		Two positions	
Dataset	Global	With bands	Global	With bands
CIFAR-10	46.98	75.60	85.54	90.18
SVHN	44.14	86.90	93.51	94.69
TinyImageNet-64	4.75	15.42	7.68	36.78
CelebA-64	29.17	48.88	63.51	70.97

B.9 Model-family extension at both query budgets

Table 15 preserves the complete model-family comparison. All rows use CIFAR-10, two separately trained targets, matched schedules across representations, and source-selected LR/MLP heads. The RF global-profile entries are the matched-schedule comparator, distinct from the independently selected global-profile MLP in Table 1.

Table 14: Paired TPR gains of the two-position band profile over the other factorial cells, in percentage points with 95% within-target percentile intervals. Each interval uses 2,000 paired resamples.

Dataset	Target	vs. one/global	vs. one/bands	vs. two/global
CIFAR-10	102	44.91 [40.76, 49.38]	15.01 [12.00, 17.12]	4.53 [2.69, 6.01]
CIFAR-10	103	41.49 [36.69, 45.86]	14.15 [12.22, 16.24]	4.75 [3.58, 6.76]
SVHN	102	50.61 [48.29, 54.09]	8.68 [7.01, 9.86]	1.29 [0.64, 1.90]
SVHN	103	50.51 [46.18, 54.11]	6.90 [5.60, 8.29]	1.07 [0.42, 1.61]
TinyImageNet-64	102	31.26 [29.39, 33.67]	21.04 [18.85, 23.58]	28.61 [26.86, 30.64]
TinyImageNet-64	103	32.81 [29.29, 34.94]	21.69 [18.11, 24.10]	29.60 [26.31, 31.73]
CelebA-64	102	44.07 [40.69, 47.50]	24.62 [20.84, 28.15]	7.47 [5.33, 11.07]
CelebA-64	103	39.52 [36.43, 44.49]	19.55 [16.85, 24.28]	7.45 [4.77, 11.04]

Table 15: Model-family extension at 32 and 128 queries: TPR at 1% FPR (%), two-target means. Bold marks the best representation for each model and budget.

Model	Queries	Global profile	PaFMIA (learned aggregation)	PaFMIA (full profile)
RF	32	93.98	**97.13**	97.00
RF	128	94.73	**97.31**	97.25
VP-FM	32	91.46	96.93	**97.23**
VP-FM	128	96.01	97.48	**97.77**
DDPM	32	97.88	99.55	**99.57**
DDPM	128	98.86	99.50	**99.52**

C MATHEMATICAL ANALYSIS

The analysis follows the paper's discovery sequence: spatial-frequency information and its compression (§C.1–C.2), complementary observations across interpolation positions and query allocation (§C.3–C.4), a fresh-noise RF mechanism (§C.5), and conditions under which compact frequency summaries suffice (§C.6). Section C.7 provides a supplementary rank-constrained interpretation. Each result states its statistical experiment and assumptions. The letters A–G identify analysis blocks and remain attached to their equations and propositions.

C.1 THE MEASURED OBSERVATION AND OPTIMAL TESTING

C.1.1 EXACT INFORMATION GEOMETRY

Fix a trained target, an interpolation-coordinate schedule, query counts, and a noise protocol. Rows index RF path positions; the DCT resolves spatial frequencies within each residual image. Condition on the source-fitting data so that filter weights and transformations are fixed. All comparisons in this section use the same realized measurement record. Write $E = (E_{tb}) \in [0,\infty)^{K\times L}$ for the mean squared DCT-band residuals, and $\omega_b = d_b/d > 0$, where $\sum_b \omega_b = 1$. Then

$$\begin{aligned} G_t &= \sum_b \omega_b E_{tb}, & F_{tb} &= \log(E_{tb}+\epsilon), \\ g_t &= \log(G_t+\epsilon) = \log \sum_b \omega_b e^{F_{tb}}. \end{aligned} \tag{A1}$$

For $\epsilon > 0$ this includes zero energies; for $\epsilon = 0$ assume positive energies. The transform from E to F is invertible on its image. A frozen filter with $a_{tb} \geq 0$ and $\sum_b a_{tb} = 1$ retains

$$h_t = \log\left(\sum_b a_{tb}E_{tb} + \epsilon\right) = \log \sum_b a_{tb} e^{F_{tb}}. \tag{A2}$$

Thus g and (g,h) are deterministic compressions of F. Explicitly appending g can help a restricted fitted classifier, but does not add statistical information to F.

Proposition A1. Preservation of the optimal ROC envelope

Let P_0 and P_1 be the nonmember and member laws of an observation X on a standard Borel space, with $P_1 \ll P_0$. Let $\Lambda = dP_1/dP_0$, and let $C = C(X)$ be a measurable compression. For randomized tests taking values in $[0,1]$, define

$$\begin{aligned} \beta_X(\alpha) &= \sup_{\mathbb{E}_0\varphi\leq\alpha} \mathbb{E}_1\varphi, \\ \beta_C(\alpha) &= \sup_{\substack{\mathbb{E}_0\varphi\leq\alpha \\ \varphi \text{ measurable with respect to } C}} \mathbb{E}_1\varphi. \end{aligned} \tag{A3}$$

For every $\alpha \in [0,1]$, $\beta_X(\alpha) \geq \beta_C(\alpha)$. Equality holds at every α if and only if Λ is measurable with respect to C, modulo P_0-null sets. Otherwise the inequality is strict at some $\alpha \in (0,1)$.

Proof. Every compressed test is a full-observation test, proving weak dominance. Put $M = \mathbb{E}_0[\Lambda \mid C]$. For every bounded measurable u,

$$\mathbb{E}_1 u(C) = \mathbb{E}_0[\Lambda u(C)] = \mathbb{E}_0[M u(C)],$$

so M is the likelihood ratio of the compressed experiment.

Likelihood-ratio thresholding, with randomization on a boundary if necessary, attains each optimal envelope. Indeed, for a size-α test $\varphi^* = \mathbf{1}\{\Lambda > c\} + \gamma\mathbf{1}\{\Lambda = c\}$ and any test ψ of size at most α,

$$\mathbb{E}_1(\varphi^* - \psi) = \mathbb{E}_0[(\Lambda - c)(\varphi^* - \psi)] + c\mathbb{E}_0(\varphi^* - \psi) \geq 0.$$

At sizes zero and one use constant tests. For interior sizes a nonnegative threshold and suitable boundary randomization exist. If $\Lambda = M$, the threshold tests are C-measurable, proving equality.

To prove necessity without a second-moment assumption, introduce

$$\begin{aligned} H_X(c) &= \sup_{0\le\varphi\le 1} (\mathbb{E}_1\varphi - c\mathbb{E}_0\varphi) \\ &= \mathbb{E}_0(\Lambda - c)_+ \\ &= \sup_{\alpha\in[0,1]} [\beta_X(\alpha) - c\alpha], \qquad c \ge 0. \end{aligned} \tag{A4}$$

Pointwise maximization gives the first identity, and optimal size-constrained tests give the last. The compressed version is $H_C(c) = \mathbb{E}_0(M-c)_+$. Complete envelope equality implies equality of these hinge expectations for every c.

Use the strictly convex function

$$\psi(u) = u - \log(1+u) = \int_0^\infty \frac{(u-c)_+}{(1+c)^2}\, dc, \quad u \ge 0.$$

It obeys $0 \le \psi(u) \le u$, so all expectations exist. Tonelli's theorem gives $\mathbb{E}_0\psi(\Lambda) = \mathbb{E}_0\psi(M)$. Strict convexity makes

$$D = \psi(\Lambda) - \psi(M) - \psi'(M)(\Lambda - M)$$

nonnegative, with equality only when $\Lambda = M$. Since $\psi'(M)$ is bounded and C-measurable, its linear term has zero expectation. Thus $\mathbb{E}_0 D = 0$ and $\Lambda = M$ almost surely.

If $\Lambda \ne M$ on a positive-probability set, the same argument gives strict inequality of the convex expectations. Conditional Jensen and the integral representation imply $H_X(c) > H_C(c)$ for some $c > 0$. Set $\alpha = P_0(\Lambda > c)$. The corresponding threshold test is optimal, and

$$\beta_C(\alpha) - c\alpha \le H_C(c) < H_X(c) = \beta_X(\alpha) - c\alpha.$$

Such an α is interior: either endpoint would rule out the strict support-function gap. □

This applies classical Neyman–Pearson testing and comparison of statistical experiments to frequency-profile compression; see Neyman & Pearson (1933) and Blackwell (1953).

A particular low FPR requires a particular condition. If $P_0(\Lambda = c_\alpha) = 0$ and $P_0(\Lambda > c_\alpha) = \alpha$, the size-$\alpha$ optimal test is unique almost surely. If $\mathbf{1}\{\Lambda > c_\alpha\}$ is not C-measurable, then $\beta_X(\alpha) > \beta_C(\alpha)$. To prove this, a compressed optimal test with equal power would also be optimal in the full experiment. Equality in the threshold argument forces it to equal the unique full test, a contradiction. This condition characterizes strict improvement at a specified operating point, including $\alpha = 0.01$.

Counterexample to strictness at every FPR. Let P_0 be uniform on four points and let $\Lambda = (1/4, 3/4, 5/4, 7/4)$. Retain the first and fourth points separately but merge the middle pair. Information is lost, but both envelopes equal $(7/4)\alpha$ for $0 \le \alpha \le 1/4$. At $\alpha = 1/2$ their powers differ: $3/4$ versus $11/16$.

Reading the optimal envelope. The proposition concerns optimal tests of the same measurement record under $P_1 \ll P_0$. Fitted LR/MLP performance additionally depends on estimation and the score class. Comparisons between different fixed-budget schedules are addressed by the separate allocation models below.

C.2 Conditional linear separation and a coherent ROC model

Proposition D1. Exact Schur decomposition of Fisher separation

Let $Z = (g, f)$ have finite second moments in both classes. Fix a pooled **within-class** covariance $\Sigma = \rho\,\mathrm{Cov}_1(Z) + (1-\rho)\,\mathrm{Cov}_0(Z) \succ 0$, with $0 < \rho < 1$, and mean difference $\delta = (\delta_g, \delta_f)$. Define

$$A = \Sigma_{fg}\Sigma_{gg}^{-1}, \quad S = \Sigma_{ff} - \Sigma_{fg}\Sigma_{gg}^{-1}\Sigma_{gf}, \quad \delta_r = \delta_f - A\delta_g.$$

Then the best squared linear Fisher separation satisfies

$$\begin{aligned} \sup_{w\ne 0} \frac{(w^\top\delta)^2}{w^\top\Sigma w} &= \delta^\top\Sigma^{-1}\delta \\ &= \delta_g^\top\Sigma_{gg}^{-1}\delta_g + \delta_r^\top S^{-1}\delta_r. \end{aligned} \tag{D1}$$

The additional term is nonnegative and is positive exactly when $\delta_r \neq 0$.

Proof. Cauchy–Schwarz for $\Sigma^{1/2}w$ and $\Sigma^{-1/2}\delta$ gives the first equality; if $\delta \neq 0$ equality is attained at $w = \Sigma^{-1}\delta$, and if $\delta = 0$ the ratio is zero. The invertible transformation $(g, f) \mapsto (g, f - Ag)$ sends the covariance to $\mathrm{diag}(\Sigma_{gg}, S)$ and the mean difference to (δ_g, δ_r). Applying the same quadratic form in these coordinates proves the decomposition. Positive definiteness implies $S \succ 0$. □

This classical linear-discrimination identity (Fisher, 1936) quantifies the extra linear mean-separation available beyond the global profile. Its moment and covariance assumptions permit application to energy features; the Gaussian construction below supplies an exact ROC interpretation.

The empirical residualizer regresses on all source-fitting samples without within-class centering and includes regularization. It need not equal A in equation D1. For any fixed matrix $\widehat{A}$, however, the linear score classes on (g, f) and $(g, f - \widehat{A}g)$ are identical before regularization, since the coordinate change is invertible. Their fitted penalized solutions can differ.

C.2.1 A VALID AUXILIARY GAUSSIAN MODEL

Equation (A1) prevents (g, F) from having a nondegenerate joint Gaussian distribution: its support lies on a deterministic nonlinear graph. This does not force its covariance matrix to be singular; nonlinear deterministic relations can still yield a positive-definite covariance.

For a consistent positive-energy model, instead use unsmoothed coordinates

$$g_t = \log \sum_b \omega_b E_{tb}, \qquad u_{tb} = \log(E_{tb}/E_{tL}), \quad b < L.$$

They are in one-to-one correspondence with positive energies:

$$E_{tL} = \frac{e^{g_t}}{\omega_L + \sum_{b<L} \omega_b e^{u_{tb}}}, \qquad E_{tb} = E_{tL} e^{u_{tb}}. \tag{D2}$$

Thus arbitrary real (g, u) induce positive energies obeying the global-energy identity.

Assume, only in this auxiliary model, $(g, u) \mid M = m \sim \mathcal{N}(\mu_m, \Sigma)$ with the same $\Sigma \succ 0$. Set $D^2 = (\mu_1 - \mu_0)^\top \Sigma^{-1} (\mu_1 - \mu_0)$. The log likelihood ratio is an affine linear score. For $D > 0$ its score mean difference is D^2 and variance in each class is D^2. Consequently its optimal ROC and AUROC are

$$\beta(\alpha) = \Phi\big(\Phi^{-1}(\alpha) + D\big), \qquad \mathrm{AUROC} = \Phi(D/\sqrt{2}). \tag{D3}$$

These follow by setting the nonmember Gaussian quantile as threshold and, for AUROC, subtracting independent member/nonmember scores. For $D = 0$ both classes coincide and randomized testing gives $\beta(\alpha) = \alpha$.

Applying equation D1 to (g, u) shows that a nonzero residual mean contrast strictly improves every interior ROC point in this positive-energy Gaussian-coordinate model.

C.3 WHEN TWO INTERPOLATION COORDINATES IMPROVE MEMBERSHIP INFERENCE

Here two-position observations are queries at two RF interpolation coordinates. We first characterize complementary information when an observation is added, then analyze reallocating a fixed query budget. We use classical Neyman–Pearson testing and Gaussian discrimination (Neyman & Pearson, 1933; Fisher, 1936). The Gaussian models describe nonredundant real-valued coordinates, such as the log-positive-energy coordinates in Section C.2.1; the deterministic identity between global and band energies remains as specified there.

Proposition G1 (Complementary evidence across path positions). Let $(X_1, X_2) \mid M = m$ be jointly Gaussian with class means (μ_{m1}, μ_{m2}) and a common positive-definite block covariance Σ. Define $\delta_j = \mu_{1j} - \mu_{0j}$,

$$A = \Sigma_{21}\Sigma_{11}^{-1}, \quad S = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, \quad r = \delta_2 - A\delta_1.$$

Then the optimal squared Gaussian separation decomposes as

$$D_{12}^2 = \delta^\top \Sigma^{-1} \delta = \delta_1^\top \Sigma_{11}^{-1} \delta_1 + r^\top S^{-1} r. \tag{12}$$

The optimal power at false-positive rate $\alpha \in (0,1)$ is

$$\beta_D(\alpha) = \Phi\big(\Phi^{-1}(\alpha) + D\big)\,. \tag{13}$$

It strictly exceeds the first-position optimum at every interior false-positive rate if and only if $r \neq 0$. It strictly exceeds both single-position optima if and only if this condition and its counterpart with the blocks exchanged both hold.

Proof. The invertible transformation $(X_1, X_2) \mapsto (X_1, X_2 - AX_1)$ diagonalizes the within-class covariance into $\mathrm{diag}(\Sigma_{11}, S)$ and maps the mean contrast to (δ_1, r). Invariance of the Mahalanobis quadratic form gives equation 12. Since $S \succ 0$, the increment is positive exactly when $r \neq 0$. The Gaussian log-likelihood ratio is affine. Its class-mean difference and each class variance equal D^2, which yields equation 13 by thresholding at the nonmember quantile. When $D = 0$, the classes coincide and randomized testing gives power α. Exchanging the blocks proves the final statement. □

Corollary G1.1 (Successive evidence from multiple positions). Fix any integer $K \geq 2$. Let $X_{1:K} = (X_1^\top, \ldots, X_K^\top)^\top$ collect nonredundant finite-dimensional measurement blocks, with $X_{1:K} \mid M = m \sim \mathcal{N}(\mu_m, \Sigma)$ and the same $\Sigma \succ 0$ in both classes. Compare prefixes of this fixed observation, retaining earlier measurements at their original precision; positions need not be independent. Write $\delta_j = \mu_{1j} - \mu_{0j}$ and $X_{<k} = X_{1:k-1}$. For $k = 2, \ldots, K$, define

$$\begin{aligned} r_k &= \delta_k - \Sigma_{k,<k}\Sigma_{<k,<k}^{-1}\delta_{<k}, \\ S_k &= \Sigma_{kk} - \Sigma_{k,<k}\Sigma_{<k,<k}^{-1}\Sigma_{<k,k}. \end{aligned} \tag{14}$$

The optimal squared Mahalanobis separations satisfy

$$D_{1:k}^2 = D_{1:k-1}^2 + r_k^\top S_k^{-1} r_k. \tag{15}$$

Thus adding position k strictly improves optimal power at every interior FPR exactly when $r_k \neq 0$; otherwise the optimal ROC is unchanged.

Proof. Apply Proposition G1 successively to $(X_{<k}, X_k)$: positive definiteness gives $S_k \succ 0$, and strict improvement follows from equation 13. □

The Schur identity also holds for optimal linear Fisher separation under finite second moments and positive-definite pooled within-class covariance, as in Proposition D1. The Gaussian common-covariance model gives the exact ROC expression. More generally, adding a retained observation weakly improves the optimal ROC by test-class inclusion. We next account explicitly for the cost of splitting a fixed query budget.

Proposition G2 (A fixed-budget nuisance-cancellation mechanism). Fix a finite nonempty candidate set of interpolation coordinates $\mathcal{T}$ and a positive even integer budget B. Consider the abstract observation model

$$Y_t(q) = a_t M + U + \varepsilon_t(q), \quad U \sim \mathcal{N}(0, \tau^2), \quad \varepsilon_t(q) \sim \mathcal{N}(0, \nu^2/q), \qquad \tau^2 \geq 0, \quad \nu^2 > 0.$$

The image-specific nuisance U is shared across path positions and is independent of membership and all query noise; noise laws are class-independent. The two-time design allocates $q = B/2$ queries to each of t_1, t_2 and has jointly zero-mean Gaussian query noise with covariance $(\nu^2/q)\left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, where $-1 < \rho < 1$. A single-time design uses all B queries at one candidate. Let $A_{\max}^2 = \max_{t \in \mathcal{T}} a_t^2$ and $\Delta a = a_{t_2} - a_{t_1}$. Then

$$D_{\mathrm{single,max}}^2 = \frac{A_{\max}^2}{\tau^2 + \nu^2/B}, \qquad D_{\mathrm{diff}}^2 = \frac{B(\Delta a)^2}{4\nu^2(1-\rho)}. \tag{16}$$

Consequently the difference alone strictly outperforms every candidate single-time optimal attack at every interior false-positive rate whenever

$$(\Delta a)^2(B\tau^2 + \nu^2) > 4\nu^2(1-\rho)A_{\max}^2. \tag{17}$$

Proof. A single-time observation has class-mean contrast a_t and common variance $\tau^2 + \nu^2/B$. In $Y_{t_2} - Y_{t_1}$ the nuisance cancels exactly, leaving contrast Δa and variance $2\nu^2/q - 2\rho\nu^2/q = 4\nu^2(1-\rho)/B$. Equation equation 16 follows. Comparing these separations and using the strict monotonicity in equation 13 proves the claim. □

The full two-time optimal separation in this model is

$$D^2_{\text{joint}} = \frac{(a_{t_1} + a_{t_2})^2}{4\tau^2 + 4\nu^2(1+\rho)/B} + \frac{B(a_{t_2} - a_{t_1})^2}{4\nu^2(1-\rho)}. \tag{18}$$

Indeed, the orthonormal sum and difference coordinates have zero covariance, hence are independent conditionally on membership. Their separations add. Condition equation 17 is sufficient, not necessary, for the full joint attack to win.

Counterexamples within the fixed-budget model. If $a_{t_1} = a_{t_2}$ and $\rho = 0$, the fixed-budget joint observation matches each selected position's single-position optimum while the difference has no signal. If $\tau^2 = \rho = 0$, $a_{t_1} \neq 0$, and $a_{t_2} = 0$, equally splitting the budget halves the first position's single-position squared separation. Positive shared-noise correlation favors the difference but can harm the sum when the membership signal is common across path positions. The $1/q$ variance and Gaussian laws here define an abstract observation model; they are not derived from taking a logarithm after averaging actual squared residuals.

Proposition G3 (Spatial-frequency-resolved contrasts). Extend the fixed-budget model to vectors:

$$Y_t(q) = a_t M + U + \varepsilon_t(q), \quad U \sim \mathcal{N}(0, T), \quad \mathrm{Cov}(\varepsilon_t(q)) = V/q, \quad T \succeq 0, \quad V \succ 0.$$

Retain the independence assumptions. Each single-time query-noise vector is Gaussian, and the selected pair has jointly zero-mean Gaussian query noise with block covariance $q^{-1}\begin{pmatrix} V & \rho V \\ \rho V & V \end{pmatrix}$, where $|\rho| < 1$. Write $d = a_{t_2} - a_{t_1}$ and $R = 4(1-\rho)V/B$. Then

$$D^2_{\text{band,diff}} = d^\top R^{-1} d, \qquad D^2_{\text{single},t} = a_t^\top (T + V/B)^{-1} a_t.$$

The band difference strictly outperforms all candidate single-time complete band-vector observations if the first quantity exceeds the maximum of the second. For any specified nonzero linear aggregate ℓ,

$$D^2_{\text{aggregate}} = \frac{(\ell^\top d)^2}{\ell^\top R \ell} \leq d^\top R^{-1} d, \tag{19}$$

with equality if and only if d is proportional to $R\ell$, including $d = 0$.

Proof. Subtracting observations at the two path positions cancels U and gives covariance R. The Gaussian separation formula yields the first statements. Applying Cauchy–Schwarz to $R^{1/2}\ell$ and $R^{-1/2}d$ gives equation 19 and its equality condition. □

Spatial-frequency-resolved contrasts expose class differences canceled by a specified aggregate. In this Gaussian model the optimal signed projection $\ell \propto R^{-1}d$ preserves all discriminative information when $d \neq 0$, showing why a learned summary can also be effective. This projection acts on the model's observation coordinates; the global log-energy statistic instead takes the log-sum-exp form in equation A1.

Connection to the RF model and empirical interpretation. These propositions identify a mechanism for large gains from combining path positions: a second position can preserve a changing membership effect while canceling shared image-level variation. The separate fresh-noise linear RF experiment in Section C.5 derives spectral class-mean effects that vary along the interpolation path. Its expected-gap analysis and the Gaussian observation models here address complementary parts of this explanation, with their respective training and covariance assumptions stated explicitly.

C.4 Repeated queries, common noise, and allocation across path positions

Proposition E1. Exact covariance under shared noise prefixes

Conditional on a fixed image and model, let $Y_{tq} = Y_t(Z_q)$ have finite second moments, where Z_q are independent, identically distributed copies of Z, the query counts Q_t are positive in-

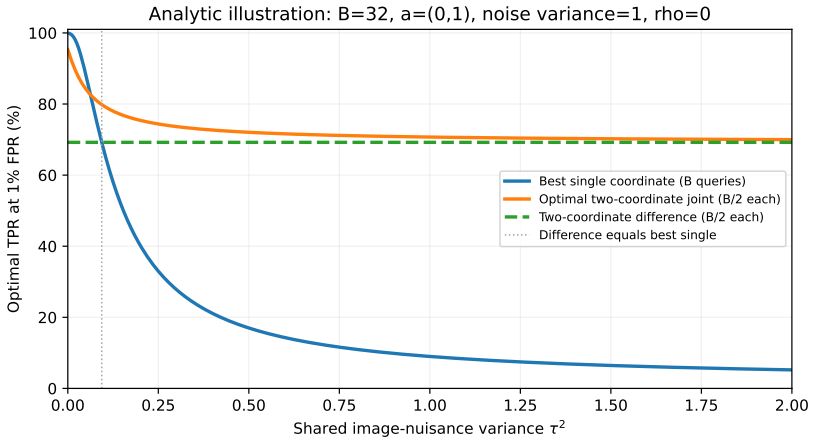


Figure 8: Analytic illustration of the fixed-budget observation model. In the two-candidate model of Proposition G2, $B = 32$, $a_{t_1} = 0$, $a_{t_2} = 1$, $\nu^2 = 1$, and $\rho = 0$. Varying the shared image-nuisance variance shows when a fixed-budget two-time difference overcomes its reduced number of repeats per path position. The full two-time optimum retains the sum direction as well.

tegers, and the same prefix is shared across path positions. Write $\widehat{\mu}_t = Q_t^{-1} \sum_{q=1}^{Q_t} Y_{tq}$ and $C_{st} = \mathrm{Cov}(Y_s(Z), Y_t(Z) \mid x)$. Then

$$\mathrm{Cov}(\widehat{\mu}_s, \widehat{\mu}_t \mid x) = \frac{\min(Q_s, Q_t)}{Q_s Q_t} C_{st} = \frac{C_{st}}{\max(Q_s, Q_t)}. \tag{E1}$$

For equal counts Q,

$$\mathrm{Var}(\widehat{\mu}_s - \widehat{\mu}_t \mid x) = \frac{C_{ss} + C_{tt} - 2C_{st}}{Q}. \tag{E2}$$

Independent noises across path positions instead give zero Monte Carlo cross covariance.

Proof. Expand the covariance of the two sums. Terms with different query indices vanish by independence; exactly $\min(Q_s, Q_t)$ terms use the same index and contribute C_{st}. The variance identity follows by bilinearity. □

Positive covariance improves estimation of a difference, negative covariance worsens it, and a different signed linear combination has a different criterion. Moreover, conditional image-level Monte Carlo covariance must be distinguished from variation across images. The exact $1/Q$ formula concerns means of Y; the implemented $\log(\widehat{\mu} + \epsilon)$ is not an average of logged per-query measurements.

How noise affects a low-FPR operating point. For a fixed scalar Gaussian score with class means μ_0, μ_1 and standard deviations $\sigma_0, \sigma_1 > 0$, thresholding high scores at $0 < \alpha < 1$ yields

$$\mathrm{TPR}(\alpha) = \Phi\left(\frac{\mu_1 - \mu_0 - z_\alpha \sigma_0}{\sigma_1}\right), \quad z_\alpha = \Phi^{-1}(1 - \alpha). \tag{E3}$$

At $\alpha = 0.01$ and mean gap 0.1, changing variances from $(1, 1)$ to $(0.9, 0.1)$ decreases power, despite reducing both variances. For example, form a difference of two Gaussian components, each of variance $1/2$: independent components give variance one, while shared-noise covariances 0.05 in class zero and 0.45 in class one give variances 0.9 and 0.1. Both covariance matrices are positive definite, so this counterexample is realizable. Equal-variance Gaussian separation is the special case where reducing a common variance while retaining a positive mean gap improves this expression.

Proposition E2. A continuous allocation result for independent noise

For a fixed linear score, restrict attention to a nonempty set of path positions with positive per-position Monte Carlo variance contributions $v_t > 0$. Independent noises give variance $\sum_t v_t/Q_t$. For $B > 0$, continuous counts $Q_t > 0$, and $\sum_t Q_t = B$, we have

$$\min \sum_t \frac{v_t}{Q_t} = \frac{(\sum_t \sqrt{v_t})^2}{B}, \qquad Q_t^* = \frac{B\sqrt{v_t}}{\sum_s \sqrt{v_s}}. \tag{E4}$$

Proof. Cauchy–Schwarz gives $(\sum_t \sqrt{v_t})^2 \le (\sum_t v_t/Q_t)(\sum_t Q_t)$. Equality holds precisely at the displayed allocation. □

If every variance contribution is zero, the objective is identically zero. Otherwise zero-variance coordinates are omitted before applying equation E4. If they are required to receive positive counts, equation E4 can be an unattained infimum rather than the stated minimum. Required minimum counts and integer allocations need constrained optimization; they are not solved by exact equality in equation E4. Shared prefixes add the cross terms in equation E1, and a fitted nonlinear classifier introduces further issues. Thus equation E4 is not claimed to optimize the implemented equal-allocation source search.

An uninformative-position counterexample at fixed cost. Let $B \ge 2$ be even and $\delta > 0$. Suppose one path position returns independent $\mathcal{N}(m\delta, 1)$ measurements in class $m \in \{0, 1\}$, while another returns class-independent noise, with all observations mutually independent across both path positions and query indices. All B queries at the first position give squared Gaussian separation $B\delta^2$. Splitting the same budget equally across the two gives $B\delta^2/2$; the second position carries no likelihood-ratio contribution. More distinct path positions can help when they add useful directions, but a fixed-budget redesign is not a nested observation.

C.5 A FRESH-NOISE RF MODEL WITH SPECTRAL VARIATION

Prior work analyzes membership signals along the RF interpolation path (Sesmat et al., 2026). We study an explicit diagonal linear ERM fitted separately at each path position, integrating fresh noise during training and evaluation. This isolates data membership from memorization of fixed data–noise pairs and permits a direct derivation of spectral variation in the membership loss gap.

C.5.1 TRAINING AND MEMBERSHIP EXPERIMENT

Let $X_1, \dots, X_n$ be independent $\mathcal{N}(0, \lambda)$ data, with $\lambda > 0$ and integer $n \ge 1$. Let $Z \sim \mathcal{N}(0, b)$ be independent noise, $b > 0$. At each interpolation coordinate $t \in [0, 1]$ fit an intercept-free scalar predictor $v_w(y, t) = w_t y$ to the integrated-noise RF objective

$$\widehat{R}_t(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_Z \left[w\{tX_i + (1-t)Z\} - (X_i - Z) \right]^2 . \tag{C1}$$

Coefficients at distinct path positions are optimized separately. An IN observation chooses a training datum uniformly and uses fresh noise. An OUT observation uses a new independent datum and fresh noise.

Write $a = 1 - t$, $s = n^{-1}\sum_i X_i^2$, and $D = t^2 s + a^2 b$. Almost surely $s > 0$. Expanding the objective gives

$$\widehat{R}_t(w) = (wt - 1)^2 s + (wa + 1)^2 b.$$

The strictly convex quadratic has minimizer

$$\widehat{w}_t = \frac{ts - ab}{D}, \qquad \widehat{w}_t t - 1 = -\frac{ab}{D}, \qquad \widehat{w}_t a + 1 = \frac{ts}{D}. \tag{C2}$$

Proposition C1. Exact expected membership loss gap

For each fixed $0 < t < 1$, let $\Delta_n(t; \lambda, b) = \mathbb{E}_{\mathcal{D}}[R_{\mathrm{OUT}}(\mathcal{D}, t) - R_{\mathrm{IN}}(\mathcal{D}, t)]$. Then

$$\Delta_n(t; \lambda, b) = \frac{4\lambda b^2}{n} t^2 (1-t)^2 \mathbb{E}\left[\frac{s}{\{t^2 s + (1-t)^2 b\}^3} \right] > 0. \tag{C3}$$

The expectation is over training datasets. At the two endpoints the expected gap equals zero.

Proof. Conditional on the data, the two fresh-noise risks are

$$R_{\mathrm{IN}} = \frac{a^2 b^2 s + t^2 s^2 b}{D^2} = \frac{bs}{D}, \qquad R_{\mathrm{OUT}} = \frac{a^2 b^2 \lambda + t^2 s^2 b}{D^2}.$$

Therefore

$$R_{\mathrm{OUT}} - R_{\mathrm{IN}} = \frac{a^2 b^2 (\lambda - s)}{D^2}. \tag{C4}$$

In particular, this conditional gap can be negative for a realized dataset.

Now s has a Gamma distribution with shape $n/2$ and **scale** $2\lambda/n$. If its density is p, then $(sp(s))' = (n/2 - ns/(2\lambda))p(s)$. Integration by parts consequently gives

$$\mathbb{E}[(s-\lambda)f(s)] = \frac{2\lambda}{n}\mathbb{E}[sf'(s)]. \tag{C5}$$

For $f(s) = (t^2 s + a^2 b)^{-2}$, the derivative is $-2t^2(t^2 s + a^2 b)^{-3}$. At zero the boundary term $sp(s)f(s)$ is of order $s^{n/2}$; at infinity the Gamma exponential tail dominates. Both boundaries vanish for every $n \geq 1$ at a fixed interior path position, and the relevant expectations are finite because $a^2 b > 0$. Substitution of equation C5 into equation C4 proves equation C3.

At $t = 0$, $\widehat{w} = -1$, and the conditional risks are s and λ, whose expectations agree. At $t = 1$, $\widehat{w} = 1$, and both risks equal b. These endpoint statements follow directly from the risks, rather than an unjustified endpoint substitution in equation C5. □

Proposition C2. Spectral differences produce nonseparable asymptotic path profiles

For every fixed interior t,

$$n\Delta_n(t;\lambda,b) \longrightarrow H_{\lambda,b}(t) = \frac{4\lambda^2 b^2 t^2 (1-t)^2}{\{t^2\lambda + (1-t)^2 b\}^3}. \tag{C6}$$

Set $b = 1$. The limiting profile has a unique maximum at $t^* = r^*/(1+r^*)$, where $r^* > 0$ solves

$$1 + 2r - 2\lambda r^2 - \lambda r^3 = 0. \tag{C7}$$

Its maximizing path coordinate decreases with λ. Moreover, if $\lambda_1 \neq \lambda_2$, any two distinct interior path positions produce a rank-two matrix

$$\begin{pmatrix} H_{\lambda_1,1}(t_1) & H_{\lambda_2,1}(t_1) \\ H_{\lambda_1,1}(t_2) & H_{\lambda_2,1}(t_2) \end{pmatrix}. \tag{C8}$$

Proof. For fixed interior t, $u/(t^2 u + a^2 b)^3$ is bounded and continuous on $[0,\infty)$. Since $s_n \to \lambda$ almost surely, bounded convergence in equation C3 proves equation C6. This is pointwise in the interior.

For $b = 1$ and $r = t/(1-t) > 0$,

$$H_{\lambda,1}(r) = \frac{4\lambda^2 r^2 (1+r)^2}{(1+\lambda r^2)^3}.$$

Its logarithmic derivative is $2/r + 2/(1+r) - 6\lambda r/(1+\lambda r^2)$, whose sign is that of the left side of equation C7. The stationary equation is

$$\begin{aligned} \lambda = q(r) &= \frac{1+2r}{r^2(2+r)}, \\ q'(r) &= -\frac{4+7r+4r^2}{r^3(2+r)^2} < 0. \end{aligned} \tag{C9}$$

The function q decreases from infinity to zero. Hence there is exactly one stationary point, the derivative changes from positive to negative there, and it is the unique global maximum. The maximizing r, and hence t, decreases with λ. When $\lambda = 1$ the peak is $t = 1/2$.

For two spectral variances, put $u = r^2$. Then

$$\begin{aligned} \frac{H_{\lambda_1,1}(t)}{H_{\lambda_2,1}(t)} &= \left(\frac{\lambda_1}{\lambda_2}\right)^2 \left(\frac{1+\lambda_2 u}{1+\lambda_1 u}\right)^3, \\ \frac{d}{du}\log\frac{H_{\lambda_1,1}}{H_{\lambda_2,1}} &= \frac{3(\lambda_2-\lambda_1)}{(1+\lambda_2 u)(1+\lambda_1 u)} \neq 0. \end{aligned} \tag{C10}$$

The ratio is strictly monotone along the interpolation path, so the determinant in equation C8 cannot vanish. □

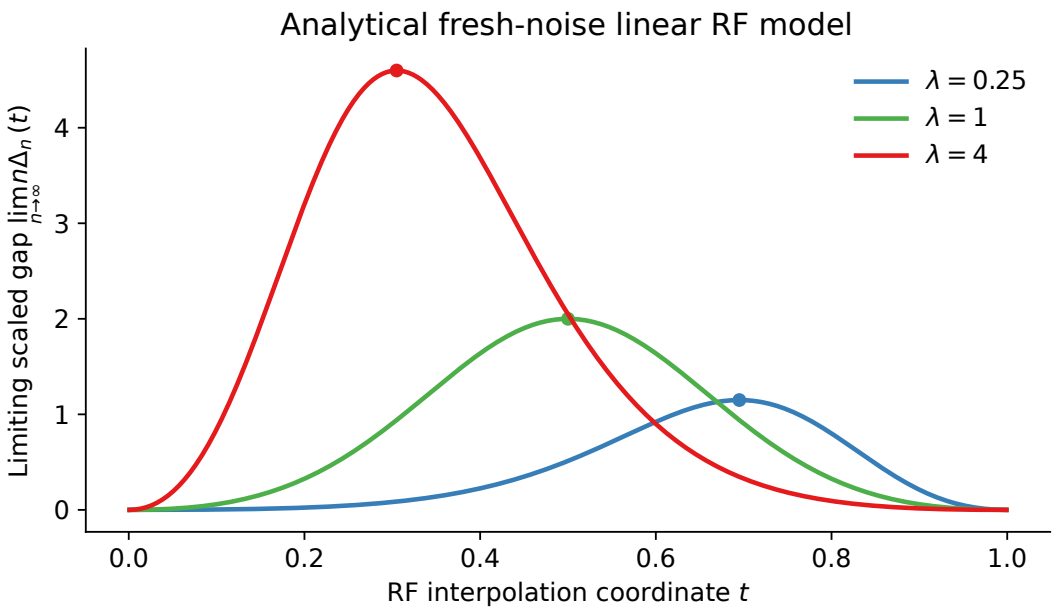


Figure 9: Analytical limiting profiles of the linear RF model in equation C6, with noise variance one. Dots mark the unique maxima: approximately 0.695, 0.500, and 0.305 for data variances 0.25, 1, and 4. Increasing spectral variance shifts the maximum toward smaller interpolation coordinates.

For a Gaussian image whose covariance is diagonal in the DCT basis, with constant variance λ_b inside each band, independently fitting each coordinate and averaging its squared loss gives the same expected per-coordinate band gap by linearity of expectation. Unequal band variances therefore generate joint path-position and spatial-frequency structure in this model. Equal variances give identical limiting columns and a rank-one nonzero matrix.

Mechanism and interpretation. Unequal spectral variances generate nonproportional membership-gap curves along the RF path. The result concerns gaps averaged over training datasets and fresh noise: its convergence is pointwise at interior coordinates, and the rank statement applies to the limiting expected-gap matrix. Low-FPR discrimination and finite-query allocation are analyzed separately in Sections C.2–C.4.

Independent fitted-model validation of the fresh-noise RF prediction. We tested the finite-sample expected-gap prediction of Proposition C1 in the same idealized linear RF setting, using independently sampled training datasets rather than plotting the derived formula alone. We fixed $n \in \{32, 128, 512\}$, data variance $\lambda \in \{0.25, 1, 4\}$, noise variance $b = 1$, and $t \in \{0, 0.1, \ldots, 1\}$ before the formal run. For each (n, λ) setting, we sampled 1,024 independent training datasets and independently fit a scalar predictor without an intercept at every interpolation coordinate. Training used weighted least squares on Gauss–Hermite quadrature rows representing the integrated fresh-noise RF objective. The analytic optimizer was used only as a post-fit implementation check. Evaluation directly integrated the fitted model's squared residuals: the member risk averaged over the training images, while the nonmember risk integrated an independent draw from the population Gaussian distribution. Both risks integrated fresh evaluation noise. Because these integrands are quadratic, two-node Gaussian quadrature evaluates the relevant expectations exactly up to numerical precision; a three-node calculation served as a check. Thus the experiment comprised 9,216 independent training datasets and 101,376 scalar fits. It uses population integration for the independent nonmember/evaluation-noise distributions, rather than a finite sampled test set.

Results. Figure 10 compares measured mean gaps with the finite-n prediction and the asymptotic profile. The empirical intervals are nominal pointwise 95% Student-t intervals over independent training datasets; the dataset-level gaps need not be Gaussian, so these are not exact finite-sample confidence intervals. Observations at different interpolation coordinates for the same dataset remain

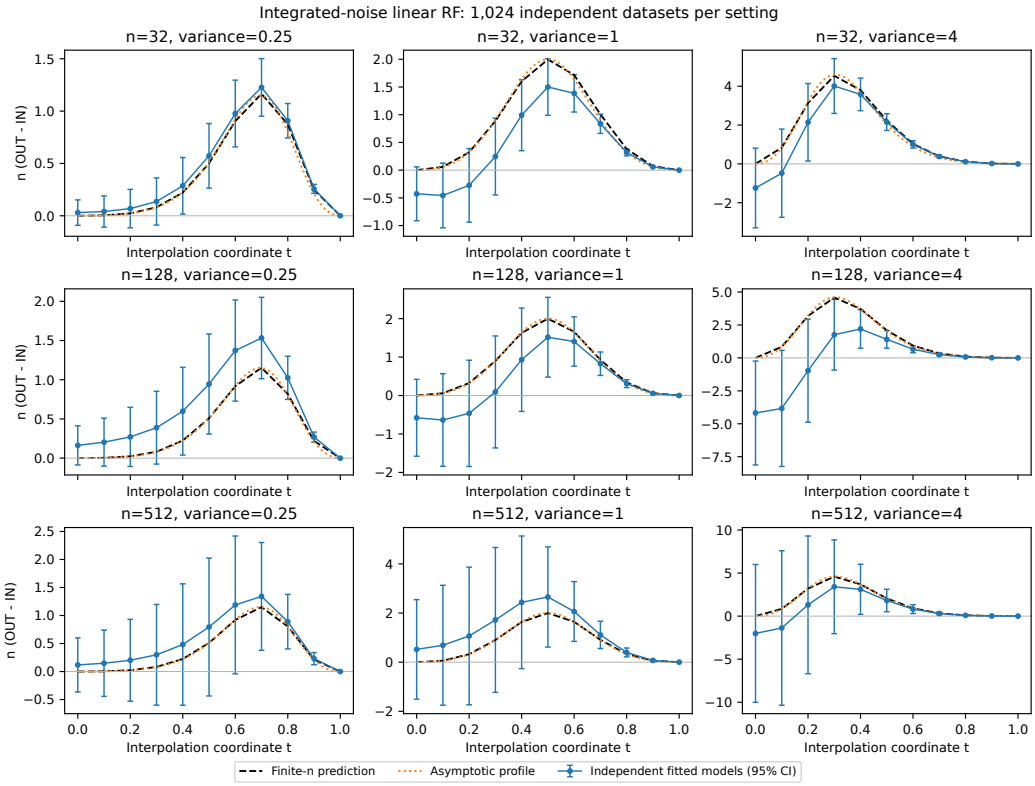


Figure 10: Independent fitted-model validation in the idealized fresh-noise linear RF model, with 1,024 training datasets per (n, λ) setting. Points and bars show measured scaled gaps $n(\mathrm{OUT} - \mathrm{IN})$ and nominal pointwise 95% intervals. The dashed black curve is the finite-n prediction; the dotted curve is the asymptotic profile. Interpolation coordinates share training datasets. This is a synthetic model experiment, distinct from the analytic two-time illustration in Figure 8 and from the neural-generator attack benchmarks.

paired. Among the 90 nondegenerate points with $t < 1$, the finite-n prediction lay within 82 intervals. The nine $t = 1$ endpoints were numerical-zero checks and are reported separately from interval coverage. The largest absolute mean discrepancy divided by its estimated standard error was 2.084. All eight noncovered points belonged to the same $(n, \lambda) = (128, 4)$ group at interpolation coordinates 0 through 0.7 and therefore do not constitute eight independent discrepancies. The empirical and predicted discrete-grid maximizers agreed in eight of nine settings: the exception was $(128, 4)$, whose empirical maximizer was 0.4 compared with the predicted 0.3. This grid-maximum comparison is descriptive and does not establish significant peak localization. At both $n = 32$ and $n = 512$, increasing λ from 0.25 to 1 to 4 moved the measured and predicted grid maximizers from 0.7 to 0.5 to 0.3. The maximum optimizer-check error was 2.00×10^{-15}, and the maximum difference between the two- and three-node risk integrations was 2.67×10^{-15}. Interval coverage is descriptive rather than a formal acceptance criterion: the run's completion status indicates successful execution and numerical checks, not proof of the statistical model. These results support the finite-sample prediction in the specified linear RF experiment; they do not validate Gaussian feature or shared-nuisance assumptions for trained neural generators, establish low-FPR attack improvements, or validate the separate two-time nuisance model.

C.6 When a small spatial-frequency summary succeeds or fails

Two positive-energy models isolate the structure that determines whether a compact frequency summary is sufficient. Gamma distributions in this section use **shape–rate** notation.

Proposition B1. A homogeneous Gamma rate shift is compressible

Suppose the coordinates are independent in each class and

$$E_{tb} \mid M = m \sim \text{Gamma}(k_{tb}, r_{m,tb}), \qquad k_{tb} > 0,\ r_{m,tb} > 0. \tag{B1}$$

The shapes are identical across classes. There exists one nonnegative normalized filter per path position such that the full global-plus-filter profile preserves the entire optimal ROC envelope.

Proof. Directly dividing Gamma densities gives

$$\log \Lambda(E) = c_0 + \sum_{t,b} c_{tb} E_{tb}, \qquad c_{tb} = r_{0,tb} - r_{1,tb}. \tag{B2}$$

For each t, choose $\kappa_t < \min_b c_{tb}/\omega_b$ and define

$$v_{tb} = c_{tb} - \kappa_t \omega_b > 0, \qquad A_t = \sum_b v_{tb}, \qquad a_{tb} = v_{tb}/A_t.$$

Then $a_{tb} > 0$, $\sum_b a_{tb} = 1$, and

$$\sum_b c_{tb} E_{tb} = \kappa_t G_t + A_t \sum_b a_{tb} E_{tb}. \tag{B3}$$

The two retained sums determine the likelihood ratio. Logging and adding a known ϵ are invertible on the retained sums. Proposition A1 therefore proves complete envelope preservation. □

The strict choice of κ_t places the weights inside the simplex, consistent with the softmax parameterization used by learned aggregation. This is a representation-level sufficiency result: the likelihood ratio is linear in raw energies, and hence generally nonlinear in the retained logged features. Recovering the weights and fitting that score are the subsequent estimation tasks.

The global profile alone is sufficient when each coefficient vector c_t is proportional to ω. Conversely, under the positive density in equation B1, a nonproportional coefficient varies on a positive-dimensional fiber of the global aggregation and cannot be a function of that aggregation almost everywhere.

Proposition B2. Heterogeneous coordinate effects can defeat every fixed linear compression

Flatten E to $e \in (0, \infty)^p$, where $p = KL$. Under nonmembership let the coordinates be independent Gamma variables with shapes $k_i > 0$ and rates $r_i > 0$. Under membership first choose a latent coordinate $J = i$ with probability $\pi_i > 0$, where $\sum_i \pi_i = 1$. In component i, increase only that coordinate's rate by $\eta_i > 0$, leaving all other Gamma laws unchanged.

For every fixed matrix A with $\text{rank}(A) < p$, the likelihood ratio is not measurable with respect to Ae. Hence the full observation has strictly greater optimal power at some interior FPR.

Proof. Division by the nonmember density yields

$$\Lambda(e) = \sum_{i=1}^{p} \pi_i \left(\frac{r_i + \eta_i}{r_i} \right)^{k_i} e^{-\eta_i e_i}. \tag{B4}$$

Its Hessian is diagonal with strictly positive entries

$$\partial_{ii} \Lambda = \pi_i \left(\frac{r_i + \eta_i}{r_i} \right)^{k_i} \eta_i^2 e^{-\eta_i e_i} > 0. \tag{B5}$$

It is therefore strictly convex along every nonzero direction in the positive orthant.

The almost-everywhere issue matters. The nonmember density is strictly positive throughout this orthant, so P_0 and Lebesgue measure have the same null sets there. Suppose $\Lambda(e) = h(Ae)$ almost everywhere. Let $r = \text{rank}(A)$, choose columns of N to span $\ker A$, and choose a complementary matrix B so that $e = Bu + Nv$ is an invertible linear coordinate change, with $u \in \mathbb{R}^r$. For fixed u, the section

$$\Omega_u = \{v : Bu + Nv \in (0, \infty)^p\}$$

is open, and $A(Bu + Nv) = ABu$ does not depend on v. Fubini's theorem implies that for almost every u, $\Lambda(Bu + Nv)$ equals a constant for almost every v in Ω_u. Some such sections are nonempty:

the positive orthant contains an open ball. Continuity then makes the function constant everywhere on a nonempty open section. Along any sufficiently short nonzero kernel direction its second derivative must vanish, contradicting equation B5. The rank-zero case has no u coordinate and the same argument applies. This proves nonmeasurability. Proposition A1 gives the power conclusion. □

With $L > 2$, one global sum and one filter per path position define a map of rank at most $2K < KL$. Thus every fixed choice of these filters, including oracle choices, loses optimal power at an interior FPR that can depend on the chosen compression. The restriction to fixed linear energy summaries is essential: a nonlinear scalar equal to the likelihood ratio itself is sufficient.

Together, B1 and B2 identify two compression regimes. Spatial-frequency effects that enter through a common linear energy statistic admit a compact sufficient summary; heterogeneous alternatives can benefit from retaining frequency structure. This provides a structural interpretation for evaluating learned aggregation alongside the full profile across datasets.

C.7 Rank constraints: optional second-moment interpretation

This section quantifies rank restrictions on population linear scores. Fix a $K \times L$ feature matrix F and global controls g. Take their positive-definite pooled within-class covariance, residualize $f = \mathrm{vec}(F)$ as in equation D1, and reshape the residual mean difference to Δ_R. Suppose the **Schur residual covariance** is separable in column-major coordinates:

$$S = \Omega_f \otimes \Omega_t, \qquad \Omega_f \succ 0, \quad \Omega_t \succ 0.$$

Set $A_R = \Omega_t^{-1/2} \Delta_R \Omega_f^{-1/2}$.

Proposition F1. Separation retained by rank-r band weights

For an integer $1 \le r \le \min(K, L)$, among scores $a^\top g + \langle W, F \rangle_F$ with unrestricted a, nonzero combined coefficients, and $\mathrm{rank}(W) \le r$,

$$D^2_{g+r} = D^2_g + \sum_{j=1}^{r} \sigma_j(A_R)^2. \tag{F1}$$

No Gaussian assumption is required for this Fisher-separation identity.

Proof. The triangular residualization leaves the band coefficient W unchanged and absorbs its global term into the free coefficient a. It block-diagonalizes the within-class covariance. Whiten the two blocks; on the residual block the transformation $V = \Omega_t^{1/2} W \Omega_f^{1/2}$ preserves rank. For $V \neq 0$ the residual separation becomes $\langle V, A_R \rangle_F^2 / \|V\|_F^2$; $V = 0$ contributes no residual separation.

For any rank-at-most-r matrix V, let P project onto its column space. Cauchy–Schwarz gives $\langle V, A_R \rangle_F^2 \le \|V\|_F^2 \|P A_R\|_F^2$. In an eigenbasis of $A_R A_R^\top$, the diagonal entries p_j of P obey $0 \le p_j \le 1$ and $\sum_j p_j \le r$. Therefore $\|P A_R\|_F^2 = \sum_j p_j \sigma_j(A_R)^2 \le \sum_{j=1}^{r} \sigma_j(A_R)^2$. The truncated SVD attains equality.

Let the whitened global mean contrast be ν, so $\|\nu\|^2 = D_g^2$. Cauchy–Schwarz on the concatenated global coefficient and admissible residual coefficient gives the upper bound $\|\nu\|^2 + \sum_{j \le r} \sigma_j^2$. Choosing those coefficients proportional to ν and the truncated SVD attains it. If both contrasts vanish, every admissible nonzero score has zero separation. □

For a fixed nonzero direction across path positions c and $W = c b^\top$, the residual term becomes $\|u^\top A_R\|^2$, with $u = \Omega_t^{1/2} c / \sqrt{c^\top \Omega_t c}$. It is at most $\sigma_1(A_R)^2$, which in turn is at most $\|A_R\|_F^2$. This separates the benefit of selecting a direction across path positions from retaining several joint path-position and spatial-frequency directions.

Equation equation F1 uses the stated separable Schur covariance: general vector whitening need not preserve matrix rank. It therefore provides a population interpretation of rank restrictions, while empirical fitted-score comparisons additionally involve parameterization and regularization. See Dyrholm et al. (2007) for prior bilinear discrimination and Eckart & Young (1936) for the classical matrix approximation background.