Learnable Obfuscation for Temporally Related Video Data

Taro Iyadomi[1] Solomon Xu[2] Rain Chai[2] Landon Mulford[2]
William Chang[2] Nader Sehatbakhsh[3]

[1]Department of Statistics and Data Science, Northwestern University
[2]Department of Mathematics, University of California, Los Angeles
[3]Department of Electrical and Computer Engineering, University of California, Los Angeles

Abstract

Learnable obfuscation enables model training on encoded data while limiting information leakage. Existing formulations assume binary membership inference—the adversary either recovers an exact datapoint or fails entirely. In video, however, even a near-miss prediction reveals privacy-relevant information, yet binary scoring treats it as a random guess.

We introduce *weighted membership inference*, where adversarial success is graded by similarity to the true dataset. We derive closed-form prior success probabilities under $1/m$-resolution weight discretizations and show they strictly exceed the binary baseline ($31\times$ prior-cost inflation on a representative universe, requiring $5.6\times$ more noise). Extending the mutual-information framework of Xiao et al. (2024), we prove that mixing and permutation bounds transfer to the weighted setting and derive noise calibration conditions under arbitrary weight resolution. We further show that temporal correlation among frame embeddings enables an aggregation attack, and that correlated noise injection eliminates this vulnerability. Experiments on UCF-101 validate the predicted noise inflation and report empirical attack scores under both weighted and binary scoring.

1 Introduction

Modern machine learning increasingly relies on outsourced computation, introducing privacy risks, such as membership inference, when the data is sensitive (Shan et al., 2018). Video data in surveillance, autonomous driving, and edge analytics contains rich temporal and semantic information (Zhou et al., 2019; Yu et al., 2018). *Learnable obfuscation* (Christiano et al., 2024) addresses this by transforming data prior to outsourcing so that (i) it remains learnable and (ii) the transformation limits information leakage. Unlike fully homomorphic encryption (Gentry, 2009; Gorantala et al., 2023), learnable obfuscation permits structured randomization (Yala et al., 2021; Xiao et al., 2025) while maintaining statistical learnability (Xiao and Devadas, 2023; Xiao et al., 2024). Xiao *et al.* (Xiao et al., 2024) bound the adversary's inference success as a function of the mutual information $\mathrm{MI}(X; \mathcal{M}(X))$.

Existing analyzes, however, consider *static* i.i.d. datasets and evaluate adversarial success under *binary* membership inference (i.e., whether or not a sample belongs to a known group such as the training dataset). Video data challenges this along two dimensions. First, the **similarity structure of the data universe**: frames that are temporally close are nearly identical, so an adversary who predicts a frame one step away from a true training frame has localized a sensitive event, a privacy breach that binary scoring ignores (Zhu et al., 2017; Jain et al., 2019). Second, **statistical dependence**: consecutive frames are highly correlated, affecting privacy.

This paper addresses the first dimension in depth and provides an initial analysis of the second. Our central contribution is a *weighted membership inference* framework that captures the graded nature of privacy leakage in structured universes, retaining the i.i.d. sampling model of Xiao et al. (2024) for the combinatorial prior analysis and separately examining temporal correlation (Section 7).

Our Key Idea. We generalize the adversary's binary score to a weighted score $\sum_i w_i$, where $w_i \in [0, 1]$ reflects proximity to the nearest true frame. In the *half-integer* model (m=2), each guess earns 1 for an exact match, 1/2 for a frame within a temporal window, and 0 otherwise. The general $1/m$-resolution model refines this to $m + 1$ levels. We characterize the adversary's prior success probability under random guessing using multivariate hypergeometric models.

Significance. On a representative universe (N=1000, n=50, m=100), the prior-cost constant, C, inflates by 31× from integer to half-integer weights, requiring 5.6× more noise to maintain the same MI-based privacy bound. For state-of-the-art Gaussian masking $\mathcal{M}(X) = XW + B$ method, we derive noise calibration conditions and show that the MI bounds from Xiao et al. (2024) carry over unchanged. Experiments on UCF-101 (Soomro et al., 2012) evaluate the predicted noise inflation, report empirical attack scores under both weighted and binary scoring, and measure downstream classifier accuracy.

Contributions.

- We formalize *weighted membership inference* for structured data universes, in which adversarial success is graded by similarity rather than exact recovery.
- We derive closed-form prior success probabilities under $1/m$-resolution weights that strictly exceed the integer-weight baseline (31× inflation at half-integer resolution).
- We extend the MI-based privacy framework (Xiao et al., 2024) to weighted inference with explicit noise calibration conditions; mixing and permutation bounds transfer directly.
- We analyze temporal correlation among frame embeddings, showing it enables an aggregation attack that weakens privacy, and that correlated noise injection eliminates this vulnerability.
- We empirically evaluate the privacy–utility tradeoff on UCF-101, reporting weighted and binary attack scores and measuring real frame-embedding autocorrelation.

Reproducibility. Code and scripts to reproduce all experiments and figures are available at the supplementary material.

2 Related Work

Our work builds on learnable obfuscation, membership inference, differential privacy for correlated data, and cryptographic approaches. We provide a condensed overview here; the full discussion is in Appendix C.

Learnable Obfuscation. Xiao et al. (2024) initiate the formal study of learnable obfuscation, building on PAC Privacy (Xiao and Devadas, 2023). Heuristic systems such as NeuraCrypt (Yala et al., 2021) and Instahide (Huang et al., 2020) lack formal guarantees (Carlini et al., 2021a,b). Membership inference attacks (Shokri et al., 2017; Carlini et al., 2022) use binary criteria; property inference (Ganju et al., 2018) and reconstruction (Balle et al., 2022) show leakage at finer levels, motivating our weighted formulation.

Differential Privacy and Temporal Correlation. DP (Dwork, 2008) provides worst-case guarantees; extensions for correlated data include Pufferfish privacy (Kifer and Machanavajjhala, 2014; Song et al., 2017) and temporal privacy leakage (Cao et al., 2017; Bozkir et al., 2021). FHE (Gentry, 2009; Gorantala et al., 2023) provides computational security but with prohibitive overhead. Our framework adopts the MI perspective of Xiao et al. (2024) and models the adversary's *scoring rule*; the two perspectives are complementary (Section 7). The full discussion is in Appendix C.

3 Preliminary

We build on the learnable obfuscation framework of Xiao et al. (2024), which we summarize here for self-containment before introducing our weighted generalization.

Binary Membership Inference (Prior Work) In the standard membership inference problem (Shokri et al., 2017; Xiao et al., 2024), a dataset $X = \{x_1, \ldots, x_n\}$ is sampled uniformly at random from a finite universe $U = \{u_1, \ldots, u_N\}$. A mechanism $\mathcal{M}$ produces a (possibly randomized) encoding $\mathcal{M}(X)$, and an adversary observes $\mathcal{M}(X)$ and attempts to recover X. The adversary outputs an n-subset $\hat{X} \subseteq U$ as its estimate of the true membership set. In the binary setting, the adversary's score is the number of correctly identified elements:

$$\text{Score}_{\text{binary}}(\hat{X}, X) = |\hat{X} \cap X| = \sum_{i=1}^{n} \mathbf{1}\{x_i \in \hat{X}\}. \tag{1}$$

Xiao et al. (2024) define a mechanism $\mathcal{M}$ to be resistant to $(1-\bar{\delta})$-average membership inference if, for every adversary, $\mathbb{E}_X[|\hat{X} \cap X|] \leq n(1-\bar{\delta})$, and show that $\bar{\delta}$ is controlled by the mutual information $\text{MI}(X; \mathcal{M}(X))$ via an f-divergence bound. Specifically, when X consists of n randomly generated datapoints and the sampling rate is $q = n/N$, the prior (random-guess) success probability for identifying at least j memberships is given by a stratified hypergeometric distribution, and the posterior success is bounded by the mutual information between X and $\mathcal{M}(X)$.

Role of Uniform Sampling. The uniform-sampling assumption from Xiao et al. (2024) yields closed-form priors via the hypergeometric distribution and models a maximum-entropy (conservative) prior; any structure in the data owner's sampling can only help the adversary. We analyze correlated sampling in Section 7.

Weighted Membership Inference (This Work). In video data, the universe U exhibits strong similarity structure: frames that are temporally close depict nearly identical content. An adversary who predicts a frame one index away from a true training frame has not achieved an exact match, yet has localized a sensitive event in time. To capture this, we replace the binary score $\mathbf{1}\{x_i \in \hat{X}\}$ with a weighted score $w_i \in [0, 1]$ that reflects the similarity between the adversary's guess and the nearest true frame.

Weight Selection. The weight function $w : U \times U \to [0, 1]$ encodes the data owner's (or analyst's) judgment about which near-miss predictions are privacy-relevant. We consider two natural families:

1. **Temporal-window weights.** Fix a window half-width Δ. For a frame at index t and a true frame at index t^* in the same clip, $w = 1$ if $t = t^*$, $w = 1/2$ if $0 < |t - t^*| \leq \Delta$, $w = 0$ otherwise. This is the half-integer model (m=2). For finer gradation, one can use $w = 1 - |t - t^*|/(m\Delta)$, yielding $1/m$-resolution weights.

2. **Embedding-similarity weights.** Given a feature extractor ϕ, set $w = \max(0, \cos(\phi(u), \phi(u^*)))$ or a discretized version thereof. This captures semantic rather than purely temporal similarity.

In our experiments, we use temporal-window weights with $\Delta = 2$, which is suitable for the 30 fps clips in UCF-101 and corresponds to a temporal tolerance of ≈ 67 ms.

Definition 1 (Weighted Membership Inference Challenge)**.** *Let $U = \{u_1, u_2, \ldots, u_N\}$ be a finite pool of video frames and let $\mathcal{M}$ be a (possibly randomized) processing mechanism. A dataset $X \subseteq U$ of size n is drawn uniformly at random from U, and the adversary observes the released output $\mathcal{M}(X)$. Based on this observation, the adversary outputs an n-subset $\hat{X} \subseteq U$ as its estimate of the true membership set X.*

We say that $\mathcal{M}$ satisfies a $(1 - \bar{\delta})$ average weighted membership inference challenge if, for every adversary,

$$\mathbb{E}_X \left[\sum_{i=1}^{n} w_i \right] \leq n(1 - \bar{\delta}), \tag{2}$$

where $w_i \in [0, 1]$ denotes the weight assigned to the ith selected frame in $\hat{X}$ with respect to the true subset X, as determined by the weight function defined above. The expectation is taken over the random draw of X (and any randomness in $\mathcal{M}$). Thus, on average, no adversary can accumulate more than $n(1 - \bar{\delta})$ total weight from correctly (or approximately correctly) identifying elements of X.

The average guarantee above controls the adversary's aggregate success across the entire dataset. Such aggregate control is determined by the adversary's performance on individual elements. The individual guarantee, following Definition 7 of Xiao et al. (2024), formalizes a stronger, element-level notion of protection. If a uniform bound $\delta_i \geq \delta$ holds for all $u_i \in U$, then the average parameter satisfies $\bar{\delta} \geq \delta$. Conversely, an average guarantee alone does not preclude the possibility that certain elements are more vulnerable than others. We therefore distinguish between aggregate protection and element-level robustness.

Individual Guarantee. For a fixed element $u_i \in U$, we say that $\mathcal{M}$ is resistant to $(1 - \delta_i)$ individual weighted membership inference for u_i if, for every adversary,

$$\Pr\left(\mathbf{1}_{\{u_i \in X\}} = \mathbf{1}_{\{u_i \in \hat{X}\}}\right) \leq 1 - \delta_i,$$

where $\mathbf{1}_{\{u_i \in X\}}$ and $\mathbf{1}_{\{u_i \in \hat{X}\}}$ are indicator variables denoting whether u_i belongs to the true set X and the adversary's estimate $\hat{X}$, respectively.

Downstream Task. Throughout this paper, the downstream task is *multi-class classification*: given obfuscated data $\mathcal{M}(X)$, a classifier is trained to predict the class label (e.g., action category in video, or object class in images) associated with each sample. The obfuscation mechanism must preserve sufficient structure for the classifier to achieve non-trivial accuracy on the transformed domain, while limiting the adversary's ability to recover membership information from $\mathcal{M}(X)$. In our experiments, the classifier operates on obfuscated clip-level video embeddings and predicts one of c action categories.

Weighted versus Binary Inference. In the classical (binary) membership inference setting, the weights satisfy $w_i \in \{0, 1\}$, so the left-hand side of (2) simply counts the number of correctly identified elements. In our setting, we allow fractional weights $w_i \in [0, 1]$ to reflect the similarity structure of the universe. If a frame returned by the adversary is close to a true frame in X, it receives partial credit rather than being counted as entirely incorrect. Consequently, the quantity on the left-hand side of (2) can only increase relative to the binary case, and satisfying the same upper bound $n(1-\bar{\delta})$ is strictly more demanding in the weighted setting, since the adversary can accumulate success even from near-miss predictions. This captures the additional difficulty of ensuring privacy when the universe exhibits strong similarity structure.

Threat Model and Obfuscation Mechanism We instantiate the learnable obfuscation framework (Xiao et al., 2024) in an outsourced *training* setting, adopting the construction in Algorithm 1 of Xiao et al. (2024). The data owner holds raw video clips X and their ground-truth class labels Y locally, and applies a paired feature/label encoding to its training set:

$$T_X(X) = \Pi_1 M_{\text{mix}} X W + B, \qquad T_Y(Y) = \Pi_1 M_{\text{mix}} Y \Pi_2,$$

where M_{mix} is a class-k mixing matrix that averages k samples from each ordered pair of action classes, Π_1 is a uniformly random row permutation, $W \in \mathbb{R}^{d_0 \times d}$ is a random Gaussian projection, B is i.i.d. Gaussian noise of variance σ^2, and Π_2 is a fixed label-class permutation. The same Π_1 and M_{mix} are applied to features and labels so that mixed feature rows remain aligned with their soft-label targets. We write $\mathcal{M}(X) := T_X(X)$ for the released obfuscated features when the label encoding is not relevant. The data owner ships only the obfuscated training pair $(T_X(X_{\text{train}}), T_Y(Y_{\text{train}}))$ to an external training service.

The parameter σ is the mechanism's primary privacy knob: increasing σ reduces $\text{MI}(X; \mathcal{M}(X))$ but degrades downstream accuracy. The external trainer is honest-but-curious. At inference time, only the ResNet-18 frame encoder and trained classifier are applied (no mixing or noise); the only retained artifact is Π_2^{-1} on logits.

The full obfuscation pipeline (Algorithm 1) and the adversary's inference procedure (Algorithm 2) are detailed in Appendix A. The adversary observes the obfuscated release and a candidate pool, and attempts to identify contributing clips or frames using likelihood-ratio scoring (Carlini et al., 2022). We instantiate two concrete adversaries: *index inference* (Attack 1, informed) and *frame-level membership inference* (Attack 2, uninformed), varying the side information available.

4 Main Results

We now bound the adversary's success in the weighted inference challenge (Definition 1) via mutual information. Following the f-divergence framework of Xiao et al. (2024) (detailed in Appendix D), the adversary's posterior advantage is bounded by $\text{MI}(X; \mathcal{F}(X))$, which in turn admits explicit upper bounds under differential privacy or via the Gram-matrix analysis of Theorem 14.

Lemma 2 (Temporal Membership Inference Hardness)**.** *For an arbitrary processing mechanism $\mathcal{M}$ and a normalized n-random-subset X from the N-universe U, the adversary's posterior success rate to identify the participation of some $u_i \in U$ is upper bounded by $(1 - \delta_i)$, where for sampling rate $q = n/N$ and*

$$1 - \delta_0 = \max\{q, 1 - q\},$$

$$(1 - \delta_i) \log\left(\frac{1 - \delta_i}{1 - \delta_0}\right) + \delta_i \log\left(\frac{\delta_i}{\delta_0}\right) \leq \text{MI}(1_{u_i}, \mathcal{M}(X)) \tag{3}$$

Moreover, the average failure rate $\bar{\delta}$ in Definition 1 satisfies

$$1 - \bar{\delta} \leq \sum_{j=1}^{n} \frac{\mathrm{MI}(X; \mathcal{M}(X)) + \log 2}{-\log(1 - \delta_0^{\geq j})}. \tag{4}$$

Proof. See Appendix D. □

For integer-valued weights, prior work (Xiao et al., 2024) shows

$$1 - \delta_0^{\geq j} = \sum_{l=j}^{n} \sum_{\substack{\{p_{1:c}\}: \\ \sum_z p_z = l}} \prod_{z=1}^{c} \frac{\binom{n_0}{p_z}\binom{N_0 - n_0}{n_0 - p_z}}{\binom{N_0}{n_0}}. \tag{5}$$

Here $\{p_{1:c}\}_l$ denotes all sets of non-negative integers $\{p_1, p_2, \ldots, p_c\}$ such that

$$\sum_{z=1}^{c} p_z = l.$$

This characterization relies on integer-valued weights and exact membership counts. When fractional weights are introduced, the combinatorial structure of the prior changes, and the corresponding success probability must be reformulated to account for weighted overlap rather than exact intersection size. The following sections analyze specific fractional-weight discretization schemes and derive analogous expressions for the prior success probability in each case.

Half-Integer Weights. We begin with the simplest nontrivial fractional-weight setting, in which similarity is discretized at resolution $\frac{1}{2}$. This case is the simplest extension of the binary formulation, allowing partial credit for near-miss predictions while preserving tractable combinatorics.

When the fractional weights are restricted to $\frac{1}{2}$, the weights satisfy

$$w_i \in \left\{0, \tfrac{1}{2}, 1\right\},$$

so that exact members of X receive full weight 1, elements that are close to X (for example, temporally adjacent or semantically similar frames) receive partial weight $\frac{1}{2}$, and unrelated elements receive weight 0. Under this model, the adversary's weighted overlap captures both exact membership and coarse similarity.

Under this half-integer weighting scheme, the adversary's prior success probability is determined by the distribution of the weighted overlap between X and a uniformly random guess $\widehat{X}$. As in the integer-weight case, this distribution admits an explicit combinatorial characterization. The following lemma derives the closed-form expression for the prior success probability under half-integer weights.

Lemma 3 (Prior success probability under fuzzy intersection)**.** *Let U be a universe of size N and let $X \subseteq U$ with $|X| = n$. Suppose elements of U are assigned weights in $\{0, \frac{1}{2}, 1\}$ relative to X, with n elements having weight 1, m elements having weight $\frac{1}{2}$, and the remaining $N - n - m$ elements having weight 0.*

Let $\widehat{X}$ be drawn uniformly at random from all n-subsets of U, and define the weighted overlap

$$W(X, \widehat{X}).$$

Then for any threshold $j \in \frac{1}{2}\mathbb{Z}$, the prior probability that the weighted overlap exceeds j is

$$1 - \delta_0^{\geq j} = \sum_{a=0}^{n} \sum_{b=0}^{n-a} \mathbf{1}\left[a + \tfrac{1}{2}b \geq j\right] \cdot \frac{\binom{n}{a}\binom{m}{b}\binom{N-n-m}{n-a-b}}{\binom{N}{n}}. \tag{6}$$

Proof. See Appendix D. □

Remark. When $m = 0$, the fractional-weight class B is empty and $W(X, \widehat{X}) = |\widehat{X} \cap X|$. The double sum collapses to the standard hypergeometric expression for the integer-weight prior, so Lemma 3 recovers the classical setting as a special case.

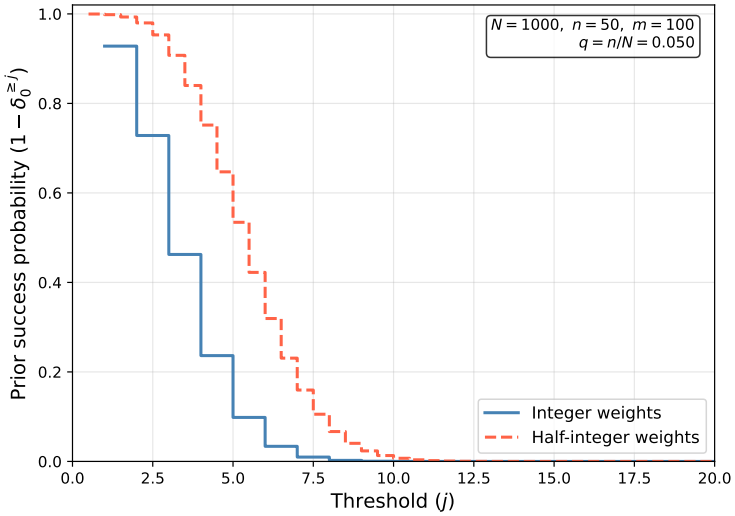


Figure 1: Prior success probability $1 - \delta_0^{\geq j}$ under integer and half-integer weights.

Figure 1 numerically illustrates the gap between integer and half-integer priors implied by Lemma 3, computed for $N = 1000$, $n = 50$, $m = 100$, and sampling rate $q = n/N = 0.05$. At every threshold j, the half-integer curve dominates the integer curve, confirming that for any fixed mechanism, weighted inference is harder to defend against than exact membership recovery.

This characterization shows how fractional weights alter the combinatorial structure of the prior relative to the integer-weight case. The following subsection extends this analysis to a more general $\frac{1}{m}$ resolution discretization, deriving the corresponding prior success probabilities under arbitrary fractional weight increments.

$\frac{1}{m}$ Resolution Weights. The half-integer analysis generalizes to weights discretized in increments of $1/m$, with each element assigned a weight in $\{0, \frac{1}{m}, \ldots, 1\}$. The count vector of selected elements across weight classes follows a multivariate hypergeometric distribution, and the prior success probability admits a closed-form characterization (Lemma 16 in Appendix D) analogous to the half-integer case. The key property is that $C_{1/m} \geq C_{\text{int}}$, with strict inequality whenever any fractional-weight elements exist.

5 Weight-Dependent Noise Calibration

In this section, we connect the weighted temporal membership inference framework to noise calibration rules for learnable obfuscation mechanisms of the form $\mathcal{M}(X) = \Pi M_{\text{mix}} XW + B$, combining mixing, permutation, masking, and additive noise. Our goal is to determine how much additional noise is required when the adversary receives graded (fractional-weight) credit for near-miss predictions.

Mutual Information Bounds under Mixing and Masking. Xiao et al. (2024) establish that for the mechanism $\mathcal{M}(X) = XW + B$ (Theorem 14, stated in full in Appendix D), $\mathrm{MI}(X; XW+B) \leq \frac{d}{2}\min\{(I), (II)\}$, where the Type (I) and (II) bounds are controlled by the Gram matrix $\Sigma_X = XX^T$ and are monotonically decreasing in σ. When mixing M_{mix} and permutation Π are applied, X is replaced by $\tilde{X} = \Pi M_{\mathrm{mix}} X$, reducing the bound by up to an order of magnitude.

Crucially, mutual information depends only on the joint distribution $(X, \mathcal{M}(X))$, not on the adversary's reward structure. All MI bounds from (Xiao et al., 2024) therefore hold unchanged in the weighted setting; the weight resolution affects only the prior success probability C, formalized below.

Noise Calibration under Half-Integer Weights. We first recall the adversary success bound from Lemma 2, which is an adaptation of Xiao et al. (2024, Corollary 1). For any processing mechanism $\mathcal{M}$, the average membership success there is bounded by a single un-factored sum,

$$1 - \bar{\delta} \;\leq\; \sum_{j=1}^{n} \frac{\mathrm{MI}(X; \mathcal{M}(X)) + (\log 2)/n}{D_j},$$

where $D_j := -\log\bigl(1 - \delta_0^{\geq j}\bigr)$ and $\delta_0^{\geq j}$ is the optimal a priori probability that a random guess achieves weighted overlap at least j, obtained in closed form from the stratified-hypergeometric distribution of the overlap between two independent n-subsets of the c-category universe. Since the numerator is independent of j, it factors out of the sum, giving the equivalent form

$$1 - \bar{\delta} \;\leq\; \bigl(\mathrm{MI}(X; \mathcal{M}(X)) + (\log 2)/n\bigr) \cdot C,$$
$$C := \sum_{j=1}^{n} \frac{1}{D_j}. \tag{7}$$

This factoring isolates the dependence of the bound on the adversary's reward structure: the mutual-information term depends only on the mechanism, while C depends only on the weight structure through $\delta_0^{\geq j}$. Changing the weight resolution from integer to $1/m$ modifies D_j, and therefore C, but leaves $\mathrm{MI}(X; \mathcal{M}(X))$ unchanged. This separation reduces the calibration arguments below to comparisons of the ratio $C_{\mathrm{int}}/C_{1/m}$.

Under integer weights, $\delta_0^{\geq j}$ is determined by the standard hypergeometric distribution (Corollary 1), yielding C_{int}. Under half-integer weights, Lemma 3 gives strictly larger prior success probabilities, yielding $C_{\mathrm{half}} > C_{\mathrm{int}}$. The following theorem determines the noise required to neutralize this increase.

Theorem 4 (Half-integer noise calibration)**.** *Consider the mechanism $\mathcal{M}(X) = XW + B$ with W having i.i.d. $\mathcal{N}(0,1)$ entries and B having i.i.d. $\mathcal{N}(0,\sigma)$ entries. Let C_{int} and C_{half} denote the prior constants in (7) under integer and half-integer weights, respectively. Suppose that under integer weights with noise variance σ_{int}, the adversary's success is bounded by*

$$1 - \bar{\delta}_{\mathrm{int}} \;\leq\; \bigl(\mathrm{MI}_{\mathrm{int}} + \log 2\bigr)\, C_{\mathrm{int}},$$

where $\mathrm{MI}_{\mathrm{int}} := \frac{d}{2}\min\{(I_{\sigma_{\mathrm{int}}}), (II_{\sigma_{\mathrm{int}}})\}$ is the mutual information upper bound from Theorem 14.

Then the adversary's success under half-integer weights is no larger than $1 - \bar{\delta}_{\mathrm{int}}$ provided the noise variance $\sigma_{\mathrm{half}} \geq \sigma_{\mathrm{int}}$ satisfies

$$\frac{d}{2}\min\{(I_{\sigma_{\mathrm{half}}}), (II_{\sigma_{\mathrm{half}}})\}$$
$$\leq\; \bigl(\mathrm{MI}_{\mathrm{int}} + \log 2\bigr)\, \frac{C_{\mathrm{int}}}{C_{\mathrm{half}}} - \log 2. \tag{8}$$

Moreover, when mixing and permutation are applied, the same condition holds with the Type (I) and (II) bounds evaluated on $\tilde{X} = \Pi M_{\text{mix}} X$ *in place of* X*.*

Proof. See Appendix D. □

Noise Calibration under $1/m$-Resolution Weights. We now extend the calibration to arbitrary fractional-weight resolution.

Theorem 5 (General $1/m$-resolution noise calibration)**.** *In the same setting as Theorem 4, let* $m \geq 2$ *be an integer specifying the weight resolution, so that weights take values in* $\{0, 1/m, 2/m, \ldots, 1\}$*. Let* $C_{1/m}$ *denote the prior constant in (7) computed using the* $1/m$*-resolution prior success probabilities from Lemma 16. Then:*

(i) $C_{1/m} \geq C_{\text{int}}$*, with equality if and only if* $M_r = 0$ *for all* $1 \leq r \leq m-1$ *(no fractional-weight elements).*

(ii) The adversary's success under $1/m$*-resolution weights is no larger than* $1 - \bar{\delta}_{\text{int}}$ *provided the noise variance* $\sigma_{1/m} \geq \sigma_{\text{int}}$ *satisfies*

$$\frac{d}{2} \min\{(I_{\sigma_{1/m}}), (II_{\sigma_{1/m}})\} \leq \left(\text{MI}_{\text{int}} + (\log 2)/n\right) \frac{C_{\text{int}}}{C_{1/m}} - (\log 2)/n. \quad (9)$$

(iii) Whenever $M_r > 0$ *for some* $1 \leq r \leq m-1$*, the inequality in part (i) is strict, and the required noise variance satisfies* $\sigma_{1/m} > \sigma_{\text{int}}$*: the presence of any fractional-weight elements forces strictly more noise than the integer-weight baseline.*

Proof. See Appendix D. □

Theorems 4 and 5 make precise the privacy cost of weighted inference. The ratio $C_{\text{int}}/C_{1/m} < 1$ quantifies how much the adversary's prior advantage increases at resolution $1/m$; the mechanism must compensate by reducing mutual information by the same factor. When data mixing is applied, the mutual information is already reduced (per the bounds of Xiao et al. (2024)), partially offsetting the increased prior. The privacy–utility relationship thus depends on the interplay between mixing intensity (which reduces MI) and weight resolution (which increases C). A direct consequence is that even modest relaxation of the correctness criterion—from exact membership ($m = 1$) to half-integer proximity ($m = 2$)—can require substantially more noise to maintain equivalent privacy, motivating the experimental comparison in Section 6.

6 Experiments

We empirically evaluate the theoretical results on UCF-101 (Soomro et al., 2012).

Experimental Setup We evaluate on UCF-101 split 1 ($c = 101$ classes, 9,537 training clips). Clips are trimmed to $L = 100$ frames (7,954 retained), and each clip is embedded by drawing $T = 16$ frame indices uniformly without replacement from $[0, L)$, resizing to 224×224, and extracting $d_0 = 512$-dimensional ResNet-18 features (L2-normalized). Random per-clip sampling ensures a meaningful inference task. The universe has $|U| = 7{,}954 \times 16 = 127{,}264$ candidate (clip, index) pairs.

Both the obfuscated classifier and baseline train on 50 clips per class ($n = 5{,}050$) using a Transformer encoder (2 layers, 8 heads). The downstream classifier sees $\mathcal{M}(X) = \Pi_1 M_{\text{mix}} X + B$; the attack uses $T_X(X) = \Pi_1 M_{\text{mix}} XW + B$ on 112×112 flattened pixels ($d = 500$). Full setup details, including a pixel-space visualization of the mechanism, classifier architecture, and reproducibility information, are in Appendix E.

Table 1: Results for Analyses in Theorems 4 and 5. C_{int} and C_{half} denote the prior-dependent factors in the adversary success bound of Lemma 2 under integer and half-integer weights, respectively.

Weight rule	C	C/C_{int}	$\sigma/\sigma_{\text{int}}$
Integer (C_{int})	20.77	1.00	1.00
Half-integer (C_{half})	651.8	31.38	5.60

Numerical Evaluation of Theorems 4 and 5. Table 1 numerically evaluates Theorems 4 and 5 on the universe used in Lemma 3's visualization (Figure 1, $N = 1000$, $n = 50$, $m = 100$, $q = 0.05$). The constant $C := \sum_{j=1}^{n} 1/D_j$ with $D_j = -\log(1 - \delta_0^{\geq j})$ is the prior-dependent factor in the adversary success bound of Lemma 2, and the ratio $C_{\text{half}}/C_{\text{int}}$ controls how much additional noise the mechanism must inject to preserve the same MI-based privacy guarantee under half-integer scoring (Theorem 4).

The numerical evaluation makes two points concrete. First, it confirms that the strict inequality $C_{\text{half}} > C_{\text{int}}$ predicted by Theorem 5 part (i) is not a vacuous one: at half-integer resolution on this universe, C inflates by $31\times$, so even the simplest fractional-weight relaxation incurs an order-of-magnitude penalty on the prior factor. Second, translating the inflation through the MI $\sim 1/\sigma^2$ approximation of Theorem 14 gives a multiplicative noise inflation of $\sqrt{C_{\text{half}}/C_{\text{int}}} \approx 5.6\times$ to maintain the same MI-based bound: the privacy cost of weighted inference is a large constant factor in σ, not a marginal correction.

Empirical Weighted Membership Inference. We instantiate the weighted membership inference attack at multiple operating points and report two attack scenarios against the closed-form random-guess and no-obfuscation baselines. We report scores under both weighted (half-credit) and binary (exact-match-only) scoring rules, allowing direct comparison to the standard membership inference setting of Xiao et al. (2024). The attacks share one likelihood-ratio scoring core, adapted from the Likelihood Ratio Attack (LiRA) of Carlini et al. (2022). LiRA scores each candidate by the log-likelihood ratio of two Gaussian hypotheses—"the candidate was a member" versus "the candidate was a non-member"—under the noise model of the release. We instantiate this idea in our obfuscation setting as follows: the attacker forms a residual $r := o - \mathcal{M}(X_{\text{cf}})$ between the observed release o and the mechanism applied to a same-class counterfactual dataset X_{cf}, and ranks each candidate c by

$$\text{LR}(c) = \frac{r \cdot \Delta_c}{\sigma^2} - \frac{1}{2}\frac{\|\Delta_c\|^2}{\sigma^2},$$

where Δ_c is the candidate-specific predicted difference between the two hypotheses, evaluated at c's mixed-row positions. This is the log-likelihood ratio of the two Gaussian hypotheses under additive noise of variance σ^2, modulo a constant that does not affect ranking. The full procedure is detailed in Algorithm 2.

The two attacks differ only in which set of candidates is ranked and how the top-ranked guesses are scored. *Attack 1 (index inference)* models an *informed* adversary who knows the target clip and is asked to recover its frame indices: it ranks the $L = 100$ candidate frame indices for the target clip

Table 2: Empirical membership inference scores at multiple operating points. "Weighted" uses the ±2-window half-credit rule; "Binary" scores only exact matches. Standard errors (std/$\sqrt{50}$) in parentheses where logged. The No-obf. and Chance columns report the *Attack 1* baselines; Attack 2's no-obfuscation baseline is 0.835 and its chance baseline is 0.008. Attack 2's weighted and binary scores coincide because its subsampled candidate pool contains only exact target rows, with no near-miss frames to receive partial credit. Full (k, σ) sweep in Appendix F.1.

(k, σ)	Scoring	Attack 1	Attack 2	No-obf.	Chance
$(0, 0.05)$	Weighted	0.441 (0.017)	0.751 (0.049)	0.460	0.373
	Binary	0.182	0.751 (0.049)	0.235	0.160
$(1, 0.05)$	Weighted	0.509 (0.013)	0.657 (0.055)	0.460	0.373
	Binary	0.230	0.657 (0.055)	0.235	0.160
$(1, 0.10)$	Weighted	0.454 (0.020)	0.439 (0.050)	0.460	0.373
	Binary	0.205	0.439 (0.050)	0.235	0.160
$(5, 0.10)$	Weighted	0.482 (0.019)	0.647 (0.042)	0.460	0.373
	Binary	0.233	0.647 (0.042)	0.235	0.160

and emits the top-T, scored under a half-credit rule with a ±2-frame temporal-proximity window: an exact (clip, index) match earns 1, a same-clip guess within ±2 of any true frame earns 0.5, otherwise 0. Under binary scoring, only exact matches earn credit. *Attack 2 (frame-level membership inference)* models an *uninformed* adversary with no side information about clip or class: it rank-orders a uniformly random subsample U_2 of $|U_2| = 2{,}016$ candidate (clip, index) pairs drawn from the full universe $|U| = 127{,}264$, with the $T = 16$ true target rows always included in U_2, and emits the top-T guesses under the same scoring rules as Attack 1. The subsample is a tractability concession—scoring all 127,264 candidates dominates cost—and does not give the adversary side information about which pairs are members: the truth is buried in U_2 alongside random distractors, and the chance baseline below is computed against the actual pool the adversary scored.

Chance baselines are computed in closed form: Attack 1's expected score under the ±2-window rule is ≈ 0.373; Attack 2's is $T/|U_2| \approx 0.008$. Table 2 reports scores over $n_{\text{targets}} = 10$ clips with $n_{\text{trials}} = 5$ trials ($n = 50$ samples). Weighted and binary scores are computed on the *same* ranked guesses. For Attack 1, weighted scoring reports 0.44–0.51 while binary reports only 0.18–0.23: measuring leakage by exact recovery alone misses roughly half of what temporal-proximity detects. For Attack 2, weighted and binary coincide (no near-miss frames in the subsampled pool). The mechanism reduces leakage relative to no obfuscation while still leaking non-trivially above chance; obfuscation is substantially more effective for the uninformed adversary. Additional analysis (ROC, privacy–utility frontier) is in Appendix E.

Utility. For $\sigma \leq 0.10$, every obfuscated configuration exceeds the no-obfuscation baseline of 60.19%, with the peak (+12.16pp at $k = 0, \sigma = 0.05$); at $\sigma = 0.50$ utility collapses, defining $\sigma \leq 0.10$ as the operable noise envelope. The full accuracy table (Table 3) is in Appendix E. The calibration-required inflation $\sigma_{\text{half}} \approx 5.6 \cdot \sigma_{\text{int}}$ would push $\sigma_{\text{int}} = 0.05$ to $\sigma_{\text{half}} \approx 0.28$—outside this band—though the inflation factor depends on the deployment-specific universe. Additional results (MI bound evaluation, ROC analysis, privacy–utility frontier, autocorrelation measurements, aggregation attack) are in Appendix E.

7 Effect of Temporal Correlation on Privacy Guarantees

The preceding analysis retains i.i.d. sampling. We additionally develop a formal treatment of temporal correlation among frame embeddings; full definitions, theorem statements with proofs, and empirical validation are in Appendix B.

Correlation model (Definition 6). We model frame embeddings within a clip as following a stationary temporal correlation with autocorrelation $\rho(\tau)$. Under block-correlated sampling (b contiguous frames from each of n/b clips), the correlation matrix R is block-diagonal with Toeplitz blocks. On UCF-101, ResNet-18 embeddings exhibit $\rho(1) \approx 0.78$ with AR(1) fit $\alpha \approx 0.84$.

MI under Correlation (Theorem 8). Under a Gaussian embedding model (Assumption 7), temporal correlation *reduces* the MI upper bound: $I_R \leq I_{I_n}$, with strict inequality when $R \neq I_n$. This follows from Jensen's inequality applied to the eigenvalues of R, which spread while preserving trace.

Prior-Cost under Block Sampling (Theorem 10). Block-correlated sampling yields $C_{\text{int}} \leq C_{\text{half}}^{\text{block}} \leq C_{\text{half}}^{\text{scatter}}$: block sampling concentrates members into clusters, reducing half-weight boundary frames relative to scattered frames. Combined noise calibration (Theorem 11) gives a net inflation factor $\sqrt{C_{\text{half}}^{\text{block}}/C_{\text{int}}}$ strictly between 1 and the scattered-frame factor.

Noise Averaging Vulnerability (Theorem 12). When an adversary averages b correlated frames, the SNR amplifies by autocorrelation factor $\bar{R}_b$. For AR(1) with $\alpha = 0.9$, $b = 16$: $\bar{R}_b \approx 9.8$, reducing effective noise by $\sqrt{9.8} \approx 3.1\times$. This is confirmed empirically: on UCF-101, $\bar{R}_{16} \approx 7.4$.

Correlated Noise Defense (Theorem 13). Matching noise correlation to data correlation ($\Sigma_B = \sigma^2 R$) eliminates the averaging vulnerability: $\text{SNR}(\bar{y}) = \text{SNR}(y_i)$ for all block sizes, and the MI bound reduces to the i.i.d. isotropic-noise bound. On synthetic AR(1) clips, moving from independent to fully clip-correlated noise reduces the aggregation attack's success from 0.29 to 0.17.

8 Conclusion

This work extends the state-of-the-art learnable obfuscation framework (Xiao et al., 2024) to structured data universes through *weighted membership inference*. We derived closed-form prior success probabilities under $1/m$-resolution weights (Lemmas 3 and 16) that strictly dominate the integer-weight baseline, and obtained noise calibration conditions (Theorems 4–5) showing that even half-integer proximity requires $5.6\times$ noise inflation on a representative universe. The correlation analysis (Section 7) reveals that temporal correlation reduces the MI bound but enables an averaging attack (Theorem 12), which correlated noise injection (Theorem 13) eliminates. UCF-101 experiments confirm these predictions and show the no-obfuscation point is Pareto-dominated at $(k, \sigma) = (1, 0.10)$.

Open directions include extending weighted inference to event-level or trajectory-level proximity, a full downstream evaluation of correlated noise, direct estimation of $\text{MI}(X; \mathcal{M}(X))$, and multi-seed evaluation with stronger adversaries.

References

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40 (12):2897–2905, 2018.

Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. Order preserving encryption for numeric data. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 563–574, 2004.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.

Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 439–448. IEEE, 2013.

Eitan Borgnia, Jonas Geiping, Valeriia Cherepanova, Liam Fowl, Arjun Gupta, Amin Ghiasi, Furong Huang, Micah Goldblum, and Tom Goldstein. Dp-instahide: Provably defusing poisoning and backdoor attacks with differentially private data augmentations. *arXiv preprint arXiv:2103.02079*, 2021.

Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F Schaefer, and Enkelejda Kasneci. Differential privacy for eye tracking with temporal correlations. *Plos one*, 16(8):e0255979, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. Quantifying differential privacy under temporal correlations. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 821–832. IEEE, 2017.

Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Abhradeep Thakurta, and Florian Tramèr. Is private learning possible with instance encoding? In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 410–427. IEEE, 2021a.

Nicholas Carlini, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Florian Tramer. Neuracrypt is not private. *arXiv preprint arXiv:2108.07256*, 2021b.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.

Ankur Chattopadhyay and Terrance E Boult. Privacycam: a privacy preserving camera using uclinux on the blackfin dsp. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pages 1964–1974. PMLR, 2021.

Paul Christiano, Jacob Hilton, Victor Lecomte, and Mark Xu. Backdoor defense, learnability and obfuscation. *arXiv preprint arXiv:2409.03077*, 2024.

A Adam Ding, Guanhong Miao, and Samuel S Wu. On the privacy and utility properties of triple matrix-masking. *The Journal of privacy and confidentiality*, 10(2):10–29012, 2020.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th annual symposium on foundations of computer science*, pages 429–438. IEEE, 2013.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. A new approach to adaptive data analysis and learning via maximal leakage. *arXiv preprint arXiv:1903.01777*, 2019.

Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 619–633, 2018.

Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178, 2009.

Shruthi Gorantala, Rob Springer, and Bryant Gipson. Unlocking the potential of fully homomorphic encryption. *Communications of the ACM*, 66(5):72–81, 2023.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens Van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pages 8056–8071. PMLR, 2022.

Qijian He, Wei Yang, Bingren Chen, Yangyang Geng, and Liusheng Huang. Transnet: Training privacy-preserving neural network over transformed layer. *Proceedings of the VLDB Endowment*, 13(12):1849–1862, 2020.

Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019.

Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instance-hiding schemes for private distributed learning. In *International conference on machine learning*, pages 4507–4518. PMLR, 2020.

Piotr Indyk, Rajeev Motwani, Prabhakar Raghavan, and Santosh Vempala. Locality-preserving hashing in multidimensional spaces. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 618–625, 1997.

Samvit Jain, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, and Joseph Gonzalez. Scaling video analytics systems to large camera deployments. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, pages 9–14, 2019.

Seny Kamara and Charalampos Papamanthou. Parallel and dynamic searchable symmetric encryption. In *International conference on financial cryptography and data security*, pages 258–274. Springer, 2013.

Daniel Kang, John Guibas, Peter D Bailis, Tatsunori Hashimoto, and Matei Zaharia. Tasti: Semantic indexes for machine learning-based queries over unstructured data. In *Proceedings of the 2022 International Conference on Management of Data*, pages 1934–1947, 2022.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th annual symposium on foundations of computer science (FOCS)*, pages 633–638. IEEE, 2017.

Kevin Lewi and David J Wu. Order-revealing encryption: New constructions, applications, and lower bounds. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1167–1178, 2016.

Zhijian Liu, Zhanghao Wu, Chuang Gan, Ligeng Zhu, and Song Han. Datamix: Efficient privacy-preserving edge-cloud inference. In *European Conference on Computer Vision*, pages 578–595. Springer, 2020.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *Advances in neural information processing systems*, 30, 2017.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

Raluca Ada Popa, Frank H Li, and Nickolai Zeldovich. An ideal-security protocol for order-preserving encoding. In *2013 IEEE symposium on security and privacy*, pages 463–477. IEEE, 2013.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30. IEEE, 2016.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

Zihao Shan, Kui Ren, Marina Blanton, and Cong Wang. Practical secure computation outsourcing: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–40, 2018.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017. S&P '17.

Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1291–1306, 2017.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Emil Stefanov, Charalampos Papamanthou, and Elaine Shi. Practical dynamic searchable encryption with small leakage. *Cryptology ePrint Archive*, 2013.

Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. Bounding membership inference. *arXiv preprint*, abs/2202.12232, 2022. Derives bounds on MI adversary accuracy from DP assumptions.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, 17(183):1–40, 2016.

Hao Wu, Xuejin Tian, Minghao Li, Yunxin Liu, Ganesh Ananthanarayanan, Fengyuan Xu, and Sheng Zhong. Pecam: Privacy-enhanced video streaming and analytics via securely-reversible transformation. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 229–241, 2021.

Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pages 611–644. Springer, 2023.

Hanshen Xiao, G Edward Suh, and Srinivas Devadas. Formal privacy proof of data encoding: The possibility and impossibility of learnable encryption. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1834–1848, 2024.

Hanshen Xiao, Jun Wan, Elaine Shi, and Srinivas Devadas. One-sided bounded noise: Theory, optimization algorithms and applications. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 4214–4228, 2025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Adam Yala, Homa Esfahanizadeh, Rafael GL D' Oliveira, Ken R Duffy, Manya Ghobadi, Tommi S Jaakkola, Vinod Vaikuntanathan, Regina Barzilay, and Muriel Medard. Neuracrypt: Hiding private health data via random neural networks for public training. *arXiv preprint arXiv:2106.02484*, 2021.

Adam Yala, Victor Quach, Homa Esfahanizadeh, Rafael GL D'Oliveira, Ken R Duffy, Muriel Médard, Tommi S Jaakkola, and Regina Barzilay. Syfer: Neural obfuscation for private data release. *arXiv preprint arXiv:2201.12406*, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

Hyunwoo Yu, Jaemin Lim, Kiyeon Kim, and Suk-Bok Lee. Pinto: enabling video privacy for commodity iot cameras. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1089–1101, 2018.

Liang Zhao and Liqun Chen. Sparse matrix masking-based non-interactive verifiable (outsourced) computation, revisited. *IEEE transactions on dependable and secure computing*, 17(6):1188–1206, 2018.

Wei Zhou, Yan Jia, Yao Yao, Lipeng Zhu, Le Guan, Yuhang Mao, Peng Liu, and Yuqing Zhang. Discovering and understanding the security hazards in the interactions between {IoT} devices, mobile apps, and clouds on smart home platforms. In *28th USENIX security symposium (USENIX security 19)*, pages 1133–1150, 2019.

Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017.

Algorithm 1 Obfuscation Mechanism (Data Owner)

Require: Training features $X \in \mathbb{R}^{n \times d_0}$ (frame embeddings), labels Y; parameters: mixing intensity $k \geq 0$, noise std. dev. $\sigma \geq 0$, projection dim. d, number of classes c
Ensure: Obfuscated features $T_X(X)$, obfuscated labels $T_Y(Y)$
// **Step 1: Class-k mixing**
if $k = 0$ **then**
  $\tilde{X} \leftarrow X$, $\tilde{Y} \leftarrow Y$
else
  for each ordered class pair $(i, j) \in [c] \times [c]$ **do**
    Sample k rows from class i, k from class j; average frame-by-frame
    Soft label: one-hot on i if i=j; mass $1/2$ on each otherwise
  end for
  Stack all c^2 mixed rows into $\tilde{X}$, $\tilde{Y}$
end if
// **Step 2: Row permutation**
$\Pi_1 \leftarrow$ uniform random permutation of rows
// **Step 3: Random projection**
Draw $W \in \mathbb{R}^{d_0 \times d}$, $W_{ij} \sim \mathcal{N}(0, 1/d)$
// **Step 4: Additive noise**
Draw $B \in \mathbb{R}^{|\tilde{X}| \times d}$, $B_{ij} \sim \mathcal{N}(0, \sigma^2)$
// **Step 5: Label permutation**
Draw fixed label permutation Π_2 on $[c]$
$T_X(X) \leftarrow \Pi_1 \tilde{X} W + B$
$T_Y(Y) \leftarrow \Pi_1 \tilde{Y} \Pi_2$
return $(T_X(X),\ T_Y(Y))$

A Algorithm Details

This appendix gives the two procedures referenced from the main paper: the data owner's obfuscation mechanism (Algorithm 1) and the adversary's weighted membership inference attack (Algorithm 2).

Obfuscation mechanism (Algorithm 1). The mechanism composes five steps, each contributing a distinct protection. *Class-k mixing* (Step 1) replaces the raw rows by averages of k samples from class i and k from class j, one such average for every ordered class pair, giving $m = c^2$ mixed rows carrying soft labels; at $k = 0$ the step is the identity. This is the lever that dilutes any single clip's contribution, and empirically the one that moves realized attack success most (Section E.5). *Row permutation* (Step 2) destroys the correspondence between released rows and their source records. *Random projection* (Step 3) maps d_0-dimensional features to $d = 500$ dimensions through W with i.i.d. $\mathcal{N}(0, 1/d)$ entries. Together with the *additive Gaussian noise* B of Step 4 this forms the mechanism $\mathcal{M}(X) = \tilde{X}W + B$ whose mutual information the calibration of Theorem 4 controls, and whose Gram-matrix bound is Theorem 14. *Label permutation* (Step 5) applies a fixed permutation Π_2 to the class alphabet, so the released labels are usable for training but not semantically interpretable. The output $(T_X(X), T_Y(Y))$ is what the downstream classifier trains on and what the adversary observes; no other information is released.

Algorithm 2 Weighted Membership Inference Attack (Adversary)

Require: Observed release $o = T_X(X)$, candidate pool U', mechanism parameters $(\sigma, k, W, \Pi_1, M_{\text{mix}})$; number of guesses T, temporal window half-width Δ
Ensure: Top-T candidates with weighted score

// **Step 1: Counterfactual construction**
Construct same-class counterfactual dataset X_{cf} excluding target
$o_{\text{cf}} \leftarrow \mathcal{M}(X_{\text{cf}})$
$r \leftarrow o - o_{\text{cf}}$ {residual}
// **Step 2: Likelihood-ratio scoring**
for each candidate $c \in U'$ **do**
  Compute predicted difference Δ_c between member/non-member hypotheses
  $\text{LR}(c) \leftarrow \frac{r \cdot \Delta_c}{\sigma^2} - \frac{\|\Delta_c\|^2}{2\sigma^2}$
end for
// **Step 3: Rank and emit top-T**
$\hat{X} \leftarrow$ top-T candidates by $\text{LR}(c)$
// **Step 4: Weighted scoring**
for each guess $\hat{x}_i \in \hat{X}$ **do**
  if exact match with a true frame **then**
    $w_i \leftarrow 1$
  else if same clip, $|\text{idx}(\hat{x}_i) - \text{idx}(x^*)| \leq \Delta$ **then**
    $w_i \leftarrow 1/2$
  else
    $w_i \leftarrow 0$
  end if
end for
return $\hat{X}$, Score $= \frac{1}{T}\sum_{i=1}^{T} w_i$

Adversary (Algorithm 2). The adversary knows the mechanism and its parameters and must identify which (clip, frame-index) pairs entered the released dataset. It proceeds in four steps. *Counterfactual construction* (Step 1) builds X_{cf} by replacing the target's rows with same-class rows from other clips and pushes it through the same mechanism draw, so the residual $r = o - o_{\text{cf}}$ isolates the target's contribution while holding the projection, permutation, and mixing fixed. *Likelihood-ratio scoring* (Step 2) scores each candidate by the Gaussian log-likelihood ratio between the member and non-member hypotheses, the LiRA statistic of Carlini et al. (2022) specialized to this mechanism. The $1/\sigma^2$ factor is a positive rescaling common to all candidates within a call, so it does not affect the ranking. Step 3 emits the top-T candidates. *Weighted scoring* (Step 4) applies the weighted formulation of Definition 1. An exact (clip, index) match earns weight 1, a guess in the same clip within $\pm\Delta$ indices of a true frame earns 1/2, and everything else earns 0, and the reported score is the mean weight over the T guesses. Setting $\Delta = 0$ recovers the binary rule reported in the "Binary" column of Table 2, and the reported experiments use $\Delta = 2$ on $L = 100$-frame clips. Both algorithms, and both weight rules, are implemented in the accompanying code supplement.

B Temporal Correlation Analysis: Full Results

This appendix contains the full definitions, assumptions, theorem statements, and empirical validation for the temporal correlation analysis summarized in Section 7.

B.1 Temporal Correlation Model

Definition 6 (Temporally correlated frame model)**.** *Let $U = \{u_1, \ldots, u_N\}$ be a universe of video frames arranged in temporal order within N/L clips of L frames each. Each frame u_t has an embedding $\phi(u_t) \in \mathbb{R}^{d_0}$. We say the embeddings follow a* stationary temporal correlation model *with autocorrelation function $\rho : \mathbb{Z}_{\geq 0} \to [0, 1]$ if, for frames u_s, u_t within the same clip,*

$$\frac{\mathbb{E}[(\phi(u_s) - \bar{\phi})^\top (\phi(u_t) - \bar{\phi})]}{\mathbb{E}[\|\phi(u_s) - \bar{\phi}\|^2]} = \rho(|s - t|),$$

where $\bar{\phi}$ is the clip mean, and frames from different clips are independent. We assume $\rho(0) = 1$, ρ is non-increasing, and $\rho(\tau) \geq 0$ for all τ.

Two concrete sampling strategies are relevant:

1. **Independent sampling.** Select n frames uniformly at random from U, independently.
2. **Block-correlated sampling.** Select n/b clips uniformly at random, then from each selected clip draw b contiguous frames starting from a uniformly random position.

We write $R \in \mathbb{R}^{n \times n}$ for the temporal correlation matrix induced by the sampling strategy. Under independent sampling, $R \approx I_n$. Under block-correlated sampling, R is block-diagonal with n/b blocks, each a $b \times b$ Toeplitz matrix with entries $\rho(|s - t|)$.

B.2 Mutual Information under Temporal Correlation

Assumption 7 (Gaussian embedding model)**.** *The centered frame embeddings $\tilde{\phi}(u_t) := \phi(u_t) - \bar{\phi}$ are jointly Gaussian with $\mathrm{vec}(\tilde{X}) \sim \mathcal{N}(0, R \otimes \Sigma_0)$, where $\Sigma_0 \in \mathbb{R}^{d_0 \times d_0}$ is a positive definite feature covariance and R is the temporal correlation matrix from Definition 6.*

Theorem 8 (MI reduction under temporal correlation)**.** *Under Assumption 7, consider the mechanism $\mathcal{M}(X) = XW + B$ with $W \in \mathbb{R}^{d_0 \times d}$ having i.i.d. $\mathcal{N}(0, 1)$ entries and $B \in \mathbb{R}^{n \times d}$ having i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, independent of X and W. Let $\lambda_1(R) \geq \cdots \geq \lambda_n(R)$ denote the eigenvalues of the temporal correlation matrix R. Then:*

(i) The mutual information satisfies

$$I(X; \mathcal{M}(X)) \;\leq\; \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{d} \mathbb{E}_W \left[\log\Big(1 + \sigma^{-2}\, \lambda_i(R)\, \lambda_j(W^\top \Sigma_0 W)\Big) \right]. \quad (10)$$

(ii) Let I_R and I_{I_n} denote the right-hand side of (10) evaluated at correlation matrix R and I_n (independent sampling), respectively. Then

$$I_R \;\leq\; I_{I_n}, \quad (11)$$

with equality if and only if $R = I_n$.

(iii) For the AR(1) autocorrelation $\rho(\tau) = \alpha^{|\tau|}$ with $\alpha \in [0, 1)$ and block size b, the MI gap satisfies

$$I_{I_n} - I_R \;\geq\; \frac{n}{2b} \sum_{j=1}^{d} \mathbb{E}_W\left[b \log(1 + \sigma^{-2}\lambda_j) - \sum_{i=1}^{b} \log(1 + \sigma^{-2}\mu_i \lambda_j) \right], \quad (12)$$

where $\mu_1 \geq \cdots \geq \mu_b$ are the eigenvalues of the $b \times b$ Toeplitz correlation matrix $T_b = (\alpha^{|s-t|})_{s,t=1}^{b}$ and $\lambda_j = \lambda_j(W^\top \Sigma_0 W)$.

Proof sketch. Under the Gaussian model the MI bound is a sum of $\log(1+\lambda_i/\sigma^2)$ over the eigenvalues λ_i of the correlated Gram matrix. Correlation spreads these eigenvalues while preserving their trace, and $\lambda \mapsto \log(1 + \lambda/\sigma^2)$ is concave, so Jensen's inequality gives $I_R \leq I_{I_n}$, strictly whenever $R \neq I_n$. Full proof in Appendix D. □

B.3 Adversary Success under Combined Correlation and Weighted Scoring

Definition 9 (Block-weighted prior)**.** *Under block-correlated sampling with block size b and $1/m$-resolution weighted scoring with window half-width Δ, the weight structure of the universe is determined by the block geometry. Each of the n/b sampled blocks occupies b contiguous frame positions. The weight classes are:*

- *$A_m = X$: the n true members (weight 1), consisting of n/b blocks of b contiguous frames;*
- *A_r for $1 \leq r \leq m-1$: frames at temporal distance τ from the nearest block boundary, where $r = m - \lceil m\tau/\Delta \rceil$, yielding weight r/m;*
- *A_0: all remaining frames (weight 0).*

Let M_r^{block} denote the cardinality of A_r under the block geometry. In the half-integer case ($m = 2$), the half-weight class has size

$$M_1^{\text{block}} = \min\left(\frac{2n\Delta}{b},\; N - n \right). \quad (13)$$

Theorem 10 (Prior-cost ordering under block vs. scattered sampling)**.** *Let $C_{\text{half}}^{\text{block}}$ and $C_{\text{half}}^{\text{scatter}}$ denote the prior-cost constants from (7) computed under half-integer weights with the block and scattered weight-class layouts of Definition 9, respectively. Then for block size $b \geq 2$,*

$$C_{\text{int}} \;\leq\; C_{\text{half}}^{\text{block}} \;\leq\; C_{\text{half}}^{\text{scatter}}, \quad (14)$$

with the first inequality strict whenever $\Delta \geq 1$, and the second strict whenever $b \geq 2$ and $n\Delta < (N-n)/2$.

Proof sketch. Block sampling draws b contiguous frames per clip, so members form clusters rather than scattering across the universe. Clustering reduces the number of distinct half-weight boundary frames adjacent to the member set, lowering the prior success probability at every threshold relative to scattered sampling while still exceeding the integer-weight case; summing over thresholds gives $C_{\text{int}} \leq C_{\text{half}}^{\text{block}} \leq C_{\text{half}}^{\text{scatter}}$. Full proof in Appendix D. □

Theorem 11 (Combined noise calibration under correlation and weighted scoring)**.** *Under the setting of Theorems 8 and 10, the adversary's success under block-correlated sampling with half-integer weighted scoring is bounded by*

$$1 - \bar{\delta} \;\le\; \big(I_R + (\log 2)/n\big) \cdot C_{\text{half}}^{\text{block}}, \tag{15}$$

where I_R is the MI bound of Theorem 8. To match the privacy guarantee $1 - \bar{\delta}_{\text{int}}$ of the i.i.d., integer-weight baseline, the noise must satisfy

$$I_R(\sigma_{\text{comb}}) \;\le\; \big(I_{I_n}(\sigma_{\text{int}}) + (\log 2)/n\big) \cdot \frac{C_{\text{int}}}{C_{\text{half}}^{\text{block}}} \;-\; (\log 2)/n. \tag{16}$$

The required noise satisfies

$$\sigma_{\text{comb}} \le \sigma_{\text{int}} \cdot \sqrt{C_{\text{half}}^{\text{block}}/C_{\text{int}}}. \tag{17}$$

Proof sketch. Substitute the correlated MI bound I_R and the block prior constant $C_{\text{half}}^{\text{block}}$ into the success bound and require it not to exceed the i.i.d. integer-weight guarantee; rearranging isolates the admissible mutual information. Because $I_R \le I_{I_n}$, the resulting constraint is no harder to satisfy than the pure weighted-scoring one, and yields the stated upper bound on σ_{comb}. Full proof in Appendix D. □

B.4 Signal-to-Noise Degradation under Temporal Aggregation

Theorem 12 (Noise averaging under temporal aggregation)**.** *Let $y_1, \dots, y_b \in \mathbb{R}^d$ be b rows of the mechanism's output $\mathcal{M}(X) = XW + B$, corresponding to b frames from the same temporal block. Write $y_i = s_i + z_i$ where $s_i = (XW)_i$ is the signal and $z_i \sim \mathcal{N}(0, \sigma^2 I_d)$ is the independent additive noise. The signal component of the block average $\bar{y} = \frac{1}{b}\sum_{i=1}^{b} y_i$ has expected squared norm*

$$\mathbb{E}[\|\bar{s}\|^2] = \frac{\mathbb{E}[\|s\|^2]}{b}\,\bar{R}_b, \tag{18}$$

where $\bar{R}_b := 1 + \frac{2}{b}\sum_{\tau=1}^{b-1}(b-\tau)\rho(\tau)$ satisfies $1 \le \bar{R}_b \le b$. The noise variance is σ^2/b, so the SNR satisfies

$$\text{SNR}(\bar{y}) = \text{SNR}(y_i) \cdot \bar{R}_b, \tag{19}$$

where $\bar{R}_b := 1 + \frac{2}{b}\sum_{\tau=1}^{b-1}(b-\tau)\rho(\tau)$ satisfies $1 \le \bar{R}_b \le b$. For the AR(1) model $\rho(\tau) = \alpha^{|\tau|}$,

$$\bar{R}_b = \frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^b)}{b(1-\alpha)^2}. \tag{20}$$

Proof sketch. Averaging b frames leaves the signal term intact up to the autocorrelation sum while the independent noise variance falls as $1/b$. Collecting the cross-terms of the signal covariance gives the factor $\bar{R}_b = 1 + \frac{2}{b}\sum_{\tau=1}^{b-1}(b-\tau)\rho(\tau)$, so the SNR is amplified by $\bar{R}_b$, with $1 \le \bar{R}_b \le b$. Full proof in Appendix D. □

B.5 Correlated Noise Injection

Theorem 13 (Privacy-preserving correlated noise)**.** *Consider the mechanism $\mathcal{M}(X) = XW + B$ where B has row covariance $\Sigma_B \in \mathbb{R}^{n\times n}$. If $\Sigma_B = \sigma^2 R$ (noise correlation matches data correlation), then:*

(i) $\text{SNR}(\bar{y}) = \text{SNR}(y_i)$ *for all block sizes b: temporal aggregation does not amplify the SNR.*

(ii) The Type-(II) MI bound generalizes to

$$\begin{aligned}\mathrm{MI}(X; XW{+}B) \le \tfrac{d}{2}\big[\log\det\big(\mathbb{E}_X\big[I_n{+}\sigma^{-2}R^{-1}\Sigma_X\big]\big)\\ -\,\mathbb{E}_X\big[\log\det\big(I_n{+}\sigma^{-2}R^{-1}\Sigma_X\big)\big]\big].\end{aligned} \tag{21}$$

(iii) Under Assumption 7, this evaluates to

$$\mathrm{MI} \;\le\; \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{E}_W\Big[\log\Big(1+\sigma^{-2}\lambda_j(W^\top\Sigma_0 W)\Big)\Big], \tag{22}$$

which is identical to the i.i.d. isotropic-noise bound $I_{I_n}(\sigma)$.

Proof sketch. With $\Sigma_B = \sigma^2 R$ the noise inherits the data's correlation, so under block averaging signal and noise are amplified by the same factor $\bar{R}_b$ and the ratio is unchanged: $\mathrm{SNR}(\bar{y}) = \mathrm{SNR}(y_i)$ for every block size. Whitening by $R^{-1/2}$ then reduces the MI bound to the i.i.d. isotropic case, so the defense costs no additional noise budget. Full proof in Appendix D. □

C Extended Related Work

Early systems such as PrivacyCam (Chattopadhyay and Boult, 2007) demonstrated that selective encryption of regions of interest in video streams can provide invertible privacy protection at the camera level; however, these approaches did not provide formal information-theoretic guarantees relating privacy leakage to downstream learning performance. More recently, PECAM (Wu et al., 2021) proposes a securely reversible transformation for video streaming that preserves downstream analytics accuracy while empirically suppressing sensitive attributes, and Kang *et al.* (Kang et al., 2022) develop learned semantic indexing mechanisms to accelerate machine learning-based queries over large unstructured datasets, including video corpora. These approaches focus on system efficiency and empirical evaluation rather than on formal information-theoretic guarantees that characterize adversarial inference success as a function of the mutual information between original and transformed video streams.

Our work builds on learnable obfuscation, membership inference, heuristic data transformations, and formal privacy guarantees. We situate our contribution within these lines of research.

C.1 Learnable Obfuscation and Property-Preserving Transformations

A recent line of work formalizes the notion of property-preserving data transformations under explicit learnability constraints. Xiao *et al.* (Xiao et al., 2024) initiate a formal study of *learnable obfuscation*, investigating whether data encodings can preserve PAC learnability (Valiant, 1984) while providing provable privacy guarantees. They show that non-trivial property-preserving transformations cannot achieve cryptographic Indistinguishability under Chosen Plaintext Attack (IND-CPA) security in the worst case, establishing a fundamental impossibility result. On the positive side, they develop information-theoretic tools under PAC Privacy (Xiao and Devadas, 2023) to quantify leakage for heuristic obfuscations such as matrix masking, data mixing, and permutation, and derive mutual-information-based bounds against reconstruction and membership inference attacks.

Parallel to this formal line of work, heuristic property-preserving transformations have appeared in several systems, including Trans Net (He et al., 2020), NeuraCrypt (Yala et al., 2021), and Syfer (Yala et al., 2022), which rely on random linear projections or random neural networks. Data mixing approaches such as Instahide (Huang et al., 2020) and its variants (Borgnia et al., 2021;

Liu et al., 2020) apply convex combinations of samples to obscure individual data points. Matrix masking techniques have also been studied in privacy and statistical disclosure contexts (Ding et al., 2020; Zhao and Chen, 2018), often motivated by the Johnson–Lindenstrauss lemma (Larsen and Nelson, 2017) and locality-preserving embeddings (Indyk et al., 1997).

Several of these constructions have subsequently been shown to lack formal security guarantees. Specifically, Carlini *et al.* (Carlini et al., 2021a,b) demonstrate reconstruction and identification attacks against instance encoding schemes such as Instahide and NeuraCrypt, illustrating the gap between empirical obfuscation and provable privacy.

Related ideas also arise in privacy-aware representation learning, where the goal is to construct feature mappings that retain task-relevant information while suppressing sensitive attributes. Adversarial representation learning frameworks (Edwards and Storkey, 2015; Louppe et al., 2017) formalize this tradeoff by training encoders to minimize predictive power about protected variables, while information bottleneck–based approaches (Alemi et al., 2016) explicitly regularize mutual information between inputs and latent representations. Formal analyses of invariant and minimal sufficient representations (Achille and Soatto, 2018) connect compression and invariance to generalization and privacy. Although these works are primarily motivated by fairness or robustness rather than outsourced learning, they share the central objective of constructing property-preserving transformations that discard extraneous information.

Our work extends the formal prior line of work (Xiao et al., 2024) by introducing graded similarity-based inference models for structured data universes, requiring new prior success characterizations beyond binary correctness. In contrast to prior representation-learning approaches, we analyze such transformations under explicit weighted membership inference guarantees and derive prior-dependent noise calibration conditions within a formal learnable obfuscation framework.

C.2 PAC Privacy and Formal Learnable Obfuscation Foundations

Our analysis builds on recent information-theoretic approaches to instance-aware privacy, particularly the PAC-privacy framework and its formalization of learnable obfuscation. Xiao and Devadas introduced *PAC Privacy* as a distribution-dependent privacy notion that quantifies the posterior inference advantage of an adversary observing the output of an arbitrary data-processing mechanism (Xiao and Devadas, 2023). Unlike differential privacy, which provides input-independent worst-case guarantees, PAC Privacy characterizes instance-level inference hardness and admits automatic Monte Carlo estimation for black-box mechanisms.

Building on this framework, Xiao *et al.* formalized *learnable obfuscation* and established both possibility and impossibility results for property-preserving encodings (Xiao et al., 2024). They show that non-trivial encodings supporting learning cannot, in general, achieve cryptographic indistinguishability but can provide information-theoretic protection against membership and reconstruction attacks when combined with calibrated random perturbations. These results provide the formal basis for analyzing mechanisms such as matrix masking, data mixing, and permutation under explicit inference-hardness guarantees.

More broadly, PAC Privacy belongs to a line of work characterizing privacy leakage through information-theoretic dependence between inputs and released outputs. Mutual-information-based analyses show that controlling $I(X; A(X))$ limits overfitting, reconstruction risk, and inference advantage (Russo and Zou, 2016; Raginsky et al., 2016; Xu and Raginsky, 2017). Related notions such as maximal leakage and coupled-worlds privacy similarly quantify adversarial success via divergence-based dependence measures rather than worst-case indistinguishability (Esposito et al., 2019; Bassily et al., 2013).

Connections between privacy guarantees and statistical learnability have also been studied in

differentially private learning, where stability characterizes when learning remains feasible under privacy constraints (Wang et al., 2016).

These results reinforce the central theme of learnable obfuscation: privacy-preserving transformations must control inference leakage while preserving task-relevant structure.

Our work extends this information-theoretic framework to structured data universes with similarity-aware inference objectives. In contrast to prior analyses focusing on binary membership recovery in static i.i.d. datasets, we introduce weighted inference models that capture graded temporal and semantic alignment between frames and derive the corresponding prior-success and noise-calibration consequences within the PAC-privacy framework.

C.3 Membership Inference Attacks

Membership inference attacks were first demonstrated in practice by Shokri *et al.* (Shokri et al., 2017). Subsequent work strengthened the theoretical understanding of reconstruction and inference hardness, including bounds on training data reconstruction (Balle et al., 2022; Guo et al., 2022) and analyses of membership inference advantage (Carlini et al., 2022; Thudi et al., 2022; Hilprecht et al., 2019).

While most prior membership inference work evaluates adversarial success under binary criteria, i.e., whether a specific record belongs to the training set, related lines of research show that privacy leakage can occur at more granular or aggregate levels. Property inference attacks show that adversaries may recover dataset-level attributes without identifying specific instances (Ganju et al., 2018; Melis et al., 2019), and reconstruction analyses quantify partial recovery risk beyond exact membership inclusion (Balle et al., 2022; Guo et al., 2022). These results indicate that inference risk is not inherently binary, motivating our graded, similarity-based formulation in which adversarial success is weighted according to partial alignment rather than exact recovery.

Beyond instance-level attacks, recent work has explored more powerful and adaptive membership inference strategies, including label-only attacks that operate without confidence scores (Choquette-Choo et al., 2021), and auditing-style attacks that assess whether entire datasets or distributions were used during training (Sablayrolles et al., 2019). Other analyses characterize membership inference through hypothesis testing and likelihood-ratio formulations, providing a statistical perspective on adversarial advantage (Yeom et al., 2018). These developments further indicate that inference risk extends beyond simple binary inclusion tests and can manifest through distributional, aggregate, or confidence-based signals, complementing the graded formulation adopted in our setting.

C.4 Differential Privacy

Differential privacy (DP) provides worst-case indistinguishability guarantees (Dwork, 2008; Duchi et al., 2013; Bun and Steinke, 2016), and has been extended to subsampling amplification (Balle et al., 2018). DP-based analysis has also been used to derive bounds on membership inference hardness (Thudi et al., 2022) and reconstruction risk (Balle et al., 2022).

While DP provides worst-case indistinguishability guarantees that hold uniformly over all neighboring datasets (Dwork, 2008; Duchi et al., 2013; Bun and Steinke, 2016), its input-independent nature can make tight, data-dependent analysis of complex transformations difficult. Information-theoretic leakage measures such as mutual information have been used to quantify inference advantage in a distribution-aware manner (Russo and Zou, 2016; Raginsky et al., 2016; Xiao and Devadas, 2023). Our framework adopts this instance-aware perspective, allowing us to directly characterize how relaxing the correctness criterion increases the adversarial prior success probability.

C.5 Differential Privacy for Temporally Correlated Data

Standard differential privacy assumes that dataset entries are independent, and its guarantees can degrade when applied to correlated data. Cao et al. (Cao et al., 2017) model temporal correlations as a Markov chain and show that event-level privacy loss under a standard DP mechanism can accumulate over time, a phenomenon they call *temporal privacy leakage*. Kifer and Machanavajjhala (Kifer and Machanavajjhala, 2014) introduce the *Pufferfish* framework, generalizing DP by specifying explicit sets of secrets and data-generating distributions, allowing privacy guarantees that account for known correlations. Song et al. (Song et al., 2017) provide the first general-purpose mechanism (the Wasserstein mechanism) for Pufferfish privacy and demonstrate it on temporally correlated physical-activity measurements. Bozkir et al. (Bozkir et al., 2021) apply DP to eye-tracking data with temporal correlations, a setting closely analogous to video frames.

These approaches modify the privacy *definition* or *mechanism* to account for dependence in the data-generating process. Our work differs in two respects. First, we operate within the mutual-information framework of learnable obfuscation (Xiao et al., 2024) rather than the (ε, δ) indistinguishability paradigm. Second, while the Pufferfish line models correlation through the data distribution, our weighted inference framework models the *adversary's scoring rule*: the data may be sampled independently, but the universe has similarity structure that makes near-miss predictions privacy-relevant. The two perspectives are complementary: Pufferfish-style analysis quantifies how correlation changes the mechanism's privacy guarantee, while weighted inference quantifies how the evaluation metric itself changes the baseline difficulty of the inference task. In Section 7, we additionally analyze how temporal correlation affects the mutual information bounds within our framework, connecting to the concerns raised by the Pufferfish literature.

C.6 Cryptographic and Secure Computation Approaches

Cryptographic solutions such as Fully Homomorphic Encryption (FHE) provide computational security for outsourced learning (Gentry, 2009; Gorantala et al., 2023), but incur significant computational overhead. Other approaches include searchable encryption (Kamara and Papamanthou, 2013; Stefanov et al., 2013) and order-preserving or order-revealing encryption (Agrawal et al., 2004; Popa et al., 2013; Lewi and Wu, 2016), enabling limited functionality without preserving statistical learnability.

Prior work (Xiao et al., 2024) show that non-trivial learnable obfuscation cannot achieve cryptographic indistinguishability in general. Our work operates within this information-theoretic setting, focusing on quantifying leakage rather than achieving computational secrecy.

Summary. Prior work studies heuristic property-preserving transformations, membership inference and reconstruction attacks, differential privacy (including extensions for correlated data), and cryptographic approaches to outsourced learning. We extend the formal learnable obfuscation framework to structured data universes, introducing graded similarity-based inference and deriving weight-dependent noise calibration rules, and provide an initial analysis of how temporal correlation among frames affects the mutual information privacy bounds.

D Proofs and Deferred Formal Statements

This appendix collects all proofs deferred from the main text, together with theorem and equation statements referenced by the proofs.

D.1 Deferred Formal Statements

The following f-divergence bound (from (Xiao et al., 2024)) underlies Lemma 2:

$$\Delta_f^\delta = \mathcal{D}_f\Big(\mathbf{1}_\delta \,\Big\|\, \mathbf{1}_{\delta_0^\rho}\Big) \;\le\; \inf_{P_W} \mathcal{D}_f\big(P_{X,\mathcal{F}(X)} \,\big\|\, P_X \otimes P_W\big)\,. \tag{23}$$

Specializing to KL divergence yields the mutual information bound:

$$\Delta_{\mathrm{KL}}^\delta \;\le\; \mathrm{MI}(X;\mathcal{F}(X))\,. \tag{24}$$

Theorem 14 (Inference Regarding Entire Set X; (Xiao et al., 2024, Theorem 2))**.** *When $W \in \mathbb{R}^{d_0\times d}$ and $B \in \mathbb{R}^{n\times d}$ are two independent random matrices, where each entry of W is i.i.d. $\mathcal{N}(0,1)$ and each column of B is i.i.d. $\mathcal{N}(0,\Sigma_B)$ for some non-singular Σ_B, then for X consisting of n randomly generated data points,*

$$\mathrm{MI}(X;XW+B) \le \frac{d}{2}\min\{(\mathrm{I}),(\mathrm{II})\}, \tag{25}$$

where $\Sigma_X = XX^T$, $\Sigma_{X,B} = XX^T + \Sigma_B$, X' is an independent copy of X, and

$$(\mathrm{I}) := \mathbb{E}_{X,X'}\Big[\mathrm{Tr}\Big(\Sigma_{X,B}^{-1}\Sigma_{X',B}\Big) - n\Big]\,,$$
$$(\mathrm{II}) := \log\det\Big(\mathbb{E}_X\big[I_n + \Sigma_B^{-1}\Sigma_X\big]\Big) - \mathbb{E}_X\Big[\log\det\Big(I_n + \Sigma_B^{-1}\Sigma_X\Big)\Big].$$

Corollary 15 (Isotropic noise)**.** *Under the setting of Theorem 14 with $\Sigma_B = \sigma I_n$, $\mathrm{MI}(X;XW+B) \le \frac{d}{2}\min\{(\mathrm{I}),(\mathrm{II})\}$, where $(\mathrm{I}) := \mathbb{E}_{X,X'}\big[\mathrm{Tr}\big((XX^\top+\sigma I_n)^{-1}(X'X'^\top+\sigma I_n)\big) - n\big]$ and $(\mathrm{II}) := \log\det\big(\mathbb{E}_X\big[I_n + \frac{1}{\sigma}\Sigma_X\big]\big) - \mathbb{E}_X\big[\log\det\big(I_n + \frac{1}{\sigma}\Sigma_X\big)\big]$.*

Lemma 16 (Prior success probability under discretized fuzzy intersection)**.** *Let U be a universe of size N and let $X \subseteq U$ with $|X| = n$. Suppose each element $u \in U$ is assigned a weight $w(u;X) \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$, with exact membership corresponding to weight 1. Let M_r denote the number of elements in U with weight r/m, so that $\sum_{r=0}^m M_r = N$ and $M_m = n$. Let $\widehat{X}$ be drawn uniformly at random from all n-subsets of U, and define the weighted overlap $W(X,\widehat{X}) = \sum_{u\in\widehat{X}} w(u;X)$. Then for any threshold $j \in \{0, \frac{1}{m}, \ldots, 1\}$,*

$$1 - \delta_0^{\ge j} = \sum_{\substack{k_0,\ldots,k_m\ge 0\\ \sum_{r=0}^m k_r = n}} \mathbf{1}\left[\sum_{r=0}^m \frac{r}{m}k_r \ge j\right] \cdot \frac{\prod_{r=0}^m \binom{M_r}{k_r}}{\binom{N}{n}}. \tag{26}$$

D.2 Proofs

Proof of Lemma 2. Equation (3) immediately follows from (24), given that under random guessing on X, the probability of positively identifying the participation of an individual is $q = n/N$. Let $\delta_{\ge j}$ represent the posterior chance that the adversary correctly identifies at least j memberships of selected X, and let $\delta_0^{\ge j}$ be the optimal *a priori* success rate before observing $\mathcal{M}(X)$. Under the normalized sampling strategy, where we randomly select n_0 samples from each class,

$$1 - \delta_0^{\ge j} = \sum_{l=j}^{n} \sum_{\substack{\{p_{1:c}\}:\\ \sum_z p_z = l}} \prod_{z=1}^{c} \frac{\binom{n_0}{p_z}\binom{N_0-n_0}{n_0-p_z}}{\binom{N_0}{n_0}}.$$

By (23),

$$\begin{aligned} & D_{\mathrm{KL}}\left(1_{\delta_{\geq j}} \,\middle\|\, 1_{\delta_0^{\geq j}}\right) \\ &= \delta_{\geq j} \log \frac{\delta_{\geq j}}{\delta_0^{\geq j}} + (1-\delta_{\geq j}) \log \frac{1-\delta_{\geq j}}{1-\delta_0^{\geq j}} \\ &\geq -(1-\delta_{\geq j}) \log(1-\delta_0^{\geq j}) - \log 2 - \delta_{\geq j} \log(\delta_0^{\geq j}), \end{aligned}$$

where we use the fact that

$$-\delta_{\geq j} \log(\delta_{\geq j}) - (1-\delta_{\geq j}) \log(1-\delta_{\geq j}) \leq \log(2).$$

Therefore, given $\delta_{\geq j} \log(\delta_0^{\geq j}) \leq 0$, combining with (23), we obtain

$$1 - \delta_{\geq j} \leq \frac{\mathrm{MI}(X; \mathcal{M}(X)) + \log(2)}{-\log(1-\delta_0^{\geq j})}. \tag{27}$$

By Fubini's theorem on expectation, the expected number of memberships that an adversary recovers in X equals

$$\sum_{j=1}^{n} (1-\delta_{\geq j}),$$

which is upper bounded by (27). Substituting the expression for $\delta_0^{\geq j}$ yields the claim.

Proof of Lemma 3 (half-integer prior). Partition the universe U into three disjoint subsets:

$$A := X, \qquad B := P(X), \qquad C := U \setminus (A \cup B),$$

where $|A| = n$, $|B| = m$, and $|C| = N - n - m$. Elements of A have weight 1, elements of B have weight $\frac{1}{2}$, and elements of C have weight 0.

For a uniformly random n-subset $\widehat{X}$, define

$$a := |\widehat{X} \cap A|, \qquad b := |\widehat{X} \cap B|.$$

Then $n - a - b = |\widehat{X} \cap C|$, and the weighted overlap satisfies

$$W(X, \widehat{X}) = a + \tfrac{1}{2} b.$$

Because $\widehat{X}$ is sampled uniformly from all n-subsets of U, the vector

$$(a, b, n - a - b)$$

follows a multivariate hypergeometric distribution with population sizes $(n, m, N - n - m)$ and sample size n. Hence

$$\Pr[a, b] = \frac{\binom{n}{a}\binom{m}{b}\binom{N-n-m}{n-a-b}}{\binom{N}{n}}.$$

The event $\{W(X, \widehat{X}) \geq j\}$ is equivalent to $\{a + \frac{1}{2} b \geq j\}$. Summing the above probability over all (a, b) satisfying this condition yields (6).

Proof of Lemma 16 ($1/m$-resolution prior). Partition U into weight classes

$$A_r := \{u \in U : w(u; X) = r/m\}, \qquad |A_r| = M_r,$$

for $r = 0, 1, \ldots, m$. These sets form a disjoint partition of U, and in particular $A_m = X$.

For a uniformly random n-subset $\widehat{X}$, define

$$k_r := |\widehat{X} \cap A_r|, \qquad r = 0, \ldots, m,$$

so that $\sum_{r=0}^{m} k_r = n$. Then

$$W(X, \widehat{X}) = \sum_{r=0}^{m} \frac{r}{m} k_r.$$

Because $\widehat{X}$ is sampled uniformly from all n-subsets of U, the count vector

$$(k_0, \ldots, k_m)$$

follows a multivariate hypergeometric distribution with category sizes $(M_0, \ldots, M_m)$ and sample size n. Hence

$$\Pr[k_0, \ldots, k_m] = \frac{\prod_{r=0}^{m} \binom{M_r}{k_r}}{\binom{N}{n}}$$

for all nonnegative vectors satisfying $\sum_{r=0}^{m} k_r = n$.

The event $\{W(X, \widehat{X}) \geq j\}$ is equivalent to

$$\sum_{r=0}^{m} \frac{r}{m} k_r \geq j,$$

or equivalently $\sum_{r=0}^{m} r\, k_r \geq mj$. Summing the above probability over all admissible count vectors satisfying this condition yields (26).

Proof of Theorem 4. Under half-integer weights, the adversary's success bound (7) becomes

$$1 - \bar{\delta}_{\text{half}} \;\leq\; \big(\text{MI}_{\text{half}} + \log 2\big)\, C_{\text{half}},$$

where MI_{half} is the mutual information upper bound at noise level σ_{half}. We require $1 - \bar{\delta}_{\text{half}} \leq 1 - \bar{\delta}_{\text{int}}$, i.e.,

$$\big(\text{MI}_{\text{half}} + \log 2\big)\, C_{\text{half}} \;\leq\; \big(\text{MI}_{\text{int}} + \log 2\big)\, C_{\text{int}}.$$

Rearranging for MI_{half}:

$$\text{MI}_{\text{half}} \;\leq\; \big(\text{MI}_{\text{int}} + \log 2\big)\, \frac{C_{\text{int}}}{C_{\text{half}}} \;-\; \log 2.$$

Since $C_{\text{half}} > C_{\text{int}}$ (because $\delta_{0,\text{half}}^{\geq j} \geq \delta_{0,\text{int}}^{\geq j}$ for all j by Lemma 3, giving $D_{j,\text{half}} \leq D_{j,\text{int}}$), the right-hand side is strictly less than MI_{int}. By Theorem 14, the mutual information bound $\frac{d}{2} \min\{(I_\sigma), (II_\sigma)\}$ is monotonically decreasing in σ. Therefore, there exists $\sigma_{\text{half}} \geq \sigma_{\text{int}}$ satisfying (8), and any such σ_{half} ensures $1 - \bar{\delta}_{\text{half}} \leq 1 - \bar{\delta}_{\text{int}}$.

For the mixing extension, mutual information is a divergence between the joint distribution $(X, \mathcal{M}(X))$ and the product of marginals, and hence depends only on the mechanism and data distribution, not on the adversary's reward structure. The Type (I) and (II) bounds from Theorems 4–5 of (Xiao et al., 2024) therefore apply without modification under half-integer weights, with X replaced by $\tilde{X}$.

Proof of Theorem 5. Part (i). Fix a threshold $j \in \{1, \dots, n\}$. For any realization of the count vector $(k_0, \dots, k_m)$ drawn from the $1/m$-resolution multivariate hypergeometric of Lemma 16, the integer overlap on the same $\widehat{X}$ equals $W_{\text{int}} = k_m$, whereas the $1/m$-resolution overlap is

$$W_{1/m} \;=\; k_m + \sum_{r=1}^{m-1} \frac{r}{m} k_r \;\geq\; k_m \;=\; W_{\text{int}},$$

with strict inequality whenever some $k_r > 0$ for $1 \leq r \leq m-1$. Hence $\{W_{\text{int}} \geq j\} \subseteq \{W_{1/m} \geq j\}$ as events on the same probability space, so

$$1 - \delta_{0,1/m}^{\geq j} \;\geq\; 1 - \delta_{0,\text{int}}^{\geq j},$$

which gives $D_{j,1/m} \leq D_{j,\text{int}}$ for every j, and summing yields $C_{1/m} \geq C_{\text{int}}$. Equality holds throughout iff $k_r = 0$ almost surely for all $1 \leq r \leq m-1$, i.e. iff $M_r = 0$ for those r.

Part (ii). By (7) the adversary's success under $1/m$-resolution weights satisfies

$$1 - \bar{\delta}_{1/m} \;\leq\; \big(\text{MI}_{1/m} + (\log 2)/n\big)\, C_{1/m}.$$

Requiring $1 - \bar{\delta}_{1/m} \leq 1 - \bar{\delta}_{\text{int}} \leq (\text{MI}_{\text{int}} + (\log 2)/n)\, C_{\text{int}}$ and rearranging yields

$$\text{MI}_{1/m} \;\leq\; \big(\text{MI}_{\text{int}} + (\log 2)/n\big)\, \frac{C_{\text{int}}}{C_{1/m}} \;-\; (\log 2)/n.$$

Since $C_{1/m} \geq C_{\text{int}}$ by part (i), the right-hand side is at most MI_{int}. By Theorem 14 the mutual information bound is monotonically decreasing in σ, so there exists $\sigma_{1/m} \geq \sigma_{\text{int}}$ satisfying (9).

Part (iii). Under the hypothesis $M_r > 0$ for some $1 \leq r \leq m-1$, the event-inclusion in part (i) is strict with positive probability, so $C_{1/m} > C_{\text{int}}$. The right-hand side of (9) is therefore strictly less than MI_{int}, and strict monotonicity of the mutual information bound in σ (Theorem 14) forces $\sigma_{1/m} > \sigma_{\text{int}}$.

Proof of Theorem 11 (combined calibration). Inequality (15) follows directly from the success bound (7) applied to the block-correlated, half-integer setting. The MI term uses the upper bound I_R from Theorem 8 (i), and the prior-cost constant is $C_{\text{half}}^{\text{block}}$ from Theorem 10.

For (16): we require $1-\bar{\delta} \leq 1-\bar{\delta}_{\text{int}}$, where the baseline satisfies $1-\bar{\delta}_{\text{int}} \leq (I_{I_n}(\sigma_{\text{int}})+(\log 2)/n)\cdot C_{\text{int}}$. Substituting into (15) and rearranging for I_R yields (16).

For (17): by Theorem 8 (ii), $I_R(\sigma) \leq I_{I_n}(\sigma)$ for all σ, so the constraint (16) is no harder to satisfy than the pure weighted-scoring constraint of (8) applied with $C_{\text{half}}^{\text{block}}$ in place of C_{half}. The MI $\sim 1/\sigma^2$ scaling then gives the square-root bound. The two-sided bound on the ratio follows from Theorem 10.

Proof of Theorem 8 (MI under correlation). Part (i). By the data processing inequality, $I(X; \mathcal{M}(X)) \leq I(X; \mathcal{M}(X) \mid W)$, since W is independent of X. Conditioned on W, the channel is $Y = XW + B$ with $XW \mid W$ Gaussian (by Assumption 7) having row covariance $W^\top \Sigma_0 W$ and row-correlation R, so $\text{Cov}(\text{vec}(XW) \mid W) = R \otimes (W^\top \Sigma_0 W)$. The noise B is independent with $\text{Cov}(\text{vec}(B)) = \sigma^2 I_{nd}$. For Gaussian inputs and additive Gaussian noise,

$$I(X; XW+B \mid W) \\ = \tfrac{1}{2} \log \det\big(I_{nd} + \sigma^{-2}\, R \otimes (W^\top \Sigma_0 W)\big).$$

Using $\det(I + A \otimes B) = \prod_{i,j}(1 + \lambda_i(A)\lambda_j(B))$ for Kronecker products and taking the expectation over W yields (10).

Part (ii). Fix W and any index j. The function $f_j(x) = \log(1 + \sigma^{-2} x\, \lambda_j)$ is concave in x for $\lambda_j > 0$. The eigenvalues of R satisfy $\lambda_i(R) \geq 0$ and $\sum_i \lambda_i(R) = \operatorname{tr}(R) = n$ (since the diagonal entries of a correlation matrix are 1). By Jensen's inequality applied to the uniform distribution on $\{\lambda_1(R), \ldots, \lambda_n(R)\}$:

$$\frac{1}{n}\sum_{i=1}^{n} \log\bigl(1 + \sigma^{-2}\lambda_i(R)\lambda_j\bigr) \;\leq\; \log\bigl(1 + \sigma^{-2}\lambda_j\bigr),$$

with equality iff all $\lambda_i(R)$ are equal to 1, i.e. $R = I_n$. Multiplying by n, summing over j, and taking the expectation over W gives (11).

Part (iii). Under block-correlated sampling with n/b independent blocks, $R = I_{n/b} \otimes T_b$. The eigenvalues of R are the eigenvalues of T_b each with multiplicity n/b. Substituting into (10) and comparing term by term with the $R = I_n$ case yields (12). The gap is non-negative by concavity of log and strict positivity of λ_j with probability one.

Proof of Theorem 10 (block vs. scattered prior). The first inequality follows from Theorem 5 part (i): any nonzero fractional-weight class ($M_1^{\text{block}} > 0$, which holds for $\Delta \geq 1$) gives $C_{\text{half}}^{\text{block}} > C_{\text{int}}$.

For the second inequality, observe that $M_1^{\text{block}} \leq M_1^{\text{scatter}}$ whenever $b \geq 2$ (by (13)). For any pair of weight-class layouts with n full-weight elements, $M_1 \leq M_1'$ implies $\delta_0^{\geq j,\text{block}} \leq \delta_0^{\geq j,\text{scatter}}$ for every threshold j, because enlarging the half-weight class stochastically increases the weighted overlap. This follows from the coupling argument of Lemma 3, in which we sample $\hat{X}$ uniformly; the weighted overlap under M_1' dominates the overlap under M_1 pathwise whenever $M_1' \geq M_1$ and $M_0' \leq M_0$. The dominance is strict for at least one j under the stated conditions, giving strict inequality for the sum C.

Proof of Theorem 12 (noise averaging). Part (i). $\bar{s} = \frac{1}{b}\sum_i s_i$, so

$$\mathbb{E}[\|\bar{s}\|^2] = \frac{1}{b^2}\sum_{i,j}\mathbb{E}[s_i^\top s_j] = \frac{\mathbb{E}[\|s\|^2]}{b^2}\sum_{i,j}\rho(|i-j|).$$

Writing $\sum_{i,j}\rho(|i-j|) = b + 2\sum_{\tau=1}^{b-1}(b-\tau)\rho(\tau)$ gives (18). The bounds $1 \leq \bar{R}_b \leq b$ follow from $0 \leq \rho(\tau) \leq 1$.

Part (ii). $\bar{z} = \frac{1}{b}\sum_i z_i$ with z_i i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$, so $\operatorname{Var}(\bar{z}) = \sigma^2/b$.

Part (iii). Division of (i) by (ii) gives (19). For AR(1), $\sum_{\tau=1}^{b-1}(b-\tau)\alpha^\tau = b\,\frac{\alpha}{1-\alpha} - \frac{\alpha(1-\alpha^b)}{(1-\alpha)^2}$, and substitution yields (20).

Proof of Theorem 13 (correlated noise). Part (i). With B having row covariance $\sigma^2 R$, the noise within a block of b frames has covariance $\sigma^2 T_b$ (the same Toeplitz structure as the signal). The block-averaged noise has variance $\sigma^2\bar{R}_b/b$, while the signal has expected squared norm $\mathbb{E}[\|s\|^2]\bar{R}_b/b$ by (18). Their ratio is $\mathbb{E}[\|s\|^2]/\sigma^2 = \text{SNR}(y_i)$, independent of b.

Part (ii). Theorem 14 with $\Sigma_B = \sigma^2 R$ gives the bound with $\Sigma_B^{-1} = \sigma^{-2}R^{-1}$.

Part (iii). Under Assumption 7, $\mathbb{E}[\Sigma_X] = \operatorname{tr}(\Sigma_0)\cdot R$ (up to centering), so $R^{-1}\mathbb{E}[\Sigma_X] = \operatorname{tr}(\Sigma_0)\, I_n$. The Kronecker factorization used in the proof of Theorem 8 yields the MI through the channel $R^{-1/2}XW + R^{-1/2}B'$, where $R^{-1/2}X$ has i.i.d. rows (by the whitening property) and $R^{-1/2}B'$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ rows. The resulting MI is the i.i.d. expression (22).

E Additional Experimental Results

E.1 Implementation and Compute Environment

This subsection records the implementation details required to reproduce every number reported in the paper.

Dataset choice. All experiments use UCF-101 split 1 (Soomro et al., 2012), selected for three reasons specific to what this paper measures. UCF-101 is densely temporally structured. Clips are recorded at 30 fps, so adjacent frames are separated by ≈ 33 ms and the intra-clip correlation the weighted formulation is built around is present rather than assumed. Section E.6 verifies this directly, measuring $\rho(1) \approx 0.78$ on real frame embeddings. That frame rate also makes the $\Delta = 2$ proximity window meaningful, since a ± 2-index window spans ≈ 67 ms, a near-miss rather than an arbitrary tolerance. The class count is large ($c = 101$), and class-k mixing constructs one mixed row per ordered class pair, so $c = 101$ yields $m = c^2 = 10{,}201$ mixed samples, enough for the mixing lever to be exercised at a realistic scale, whereas a ten-class benchmark would make the mixed dataset degenerately small. UCF-101 is also a public, widely used action-recognition benchmark with a standard split, so both the utility numbers and the membership-inference setup are comparable to prior work and reproducible without a data-access request. We use no other dataset; the controlled study of Section E.7 uses synthetic AR(1) clips generated at runtime by the included code rather than any external corpus.

Classifier architecture. The downstream model is a Transformer encoder applied to the (T, d_0) sequence of per-frame embeddings: 2 encoder layers, 8 attention heads, feed-forward width 512, dropout 0.1, `batch_first` ordering, with a learnable `[CLS]` token and learned positional embeddings; the class logits are read from the `[CLS]` position. The frame encoder is an ImageNet-pretrained ResNet-18 producing $d_0 = 512$-dimensional embeddings, cached before obfuscation. Frames are grayscale 112×112; each clip contributes $T = 16$ frames sampled from its first $L = 100$.

Training hyperparameters. All obfuscated and no-obfuscation runs share an identical training configuration: Adam, learning rate 10^{-3} (constant, no schedule), batch size 64, 50 epochs, no weight decay, soft-label cross-entropy. The training set is a stratified subset of 50 clips per class ($n = 5{,}050$); class-k mixing produces $m = c^2 = 10{,}201$ mixed samples. The random projection is $W \in \mathbb{R}^{d_0 \times d}$ with $d = 500$. No hyperparameter search was performed: these values were fixed once and held constant across every (k, σ) cell, so the comparison across cells varies only the mechanism parameters. The swept mechanism parameters are $k \in \{0, 1, 5\}$ and $\sigma \in \{0, 0.01, 0.05, 0.10, 0.50\}$ (plus $\sigma \in \{2.0, 5.0\}$ at $k = 5$), with the operable band $\sigma \leq 0.10$ selected by downstream accuracy.

Randomness and seeding. A single integer seed (default 42) determines every stochastic component. It is applied through `torch.manual_seed`, `torch.cuda.manual_seed_all`, and NumPy's `default_rng`. Derived streams are seeded deterministically from it so that runs are reproducible and independent across units: the per-clip frame-index draw uses `default_rng(seed · 2^20 + clip_index)`, each attack trial uses `default_rng(seed · 10^6 + target_index · 10^3 + trial)`, and each noise level in the sweep uses seed $+ i$. Within a trial, the mixing records M_{mix}, the permutation Π_1, the projection W, and the noise B are drawn from that trial's generator. Attack cells report $n_{\text{targets}} \times n_{\text{trials}}$ Monte-Carlo samples (10×5 for the operating-point table, 15×6 for the ROC evaluation); the utility numbers are a single training run per (k, σ) cell.

Compute environment. All experiments were run on a single Apple M2 Pro workstation (10 CPU cores, 16 GB unified memory) under macOS 26.0.1 (arm64), using the PyTorch MPS backend; no CUDA GPU was used. Software: Python 3.11.14, PyTorch 2.13.0, torchvision 0.28.0, NumPy 2.4.6, SciPy 1.17.1, pandas 3.0.3, Matplotlib 3.11.1, OpenCV 5.0.0. Wall-clock cost is dominated by the 127,264 × 12,544 projection applied per Monte-Carlo trial in the membership-inference attacks (on the order of a minute per trial); a full downstream training run of 50 epochs completes in a few minutes on this hardware.

E.2 Utility

Table 3 reports downstream test accuracy across the full (k, σ) sweep, and Figure 2 shows what the mechanism does to the data in pixel space at a representative operating point.

Two features of the sweep matter for the rest of the appendix. The operable band is $\sigma \leq 0.10$, and every cell in that band exceeds the no-obfuscation baseline of 60.19%, by between +4.62 and +12.16 percentage points, whereas $\sigma = 0.50$ costs 16–30 points and falls far below the baseline. That band delimits the settings under which the privacy results are informative, since a setting that destroys the downstream task is not a privacy–utility trade-off, and Section E.5 restricts attention to that envelope.

The accuracy gain at low σ is not the paper's claim and should not be read as one. Mild additive noise and class-k mixing act as regularizers on a 5,050-clip training set, and the mixed dataset is also twice the size ($m = c^2 = 10{,}201$ rows against $n = 5{,}050$), so a stronger baseline, or more training data, would narrow or remove the gap. The table establishes the weaker and more robust statement that the mechanism is not costly in utility within the operable band, so the leakage reductions reported in Sections E.4 and E.5 are not being bought with accuracy.

Each cell is a single training run at seed 42 under the shared configuration of Section E.1; see Section E.5 for what that does and does not support.

Table 3: UCF-101 split-1 test accuracy across the (k, σ) sweep.

k	σ	Test Acc. (%)	Δ vs. baseline (pp)
No obfuscation		60.19	—
0	0.01	64.81	+4.62
0	0.05	72.35	+12.16
0	0.10	68.24	+8.05
0	0.50	30.54	−29.64
1	0.01	70.00	+9.82
1	0.05	68.37	+8.18
1	0.10	66.60	+6.42
1	0.50	42.83	−17.36
5	0.01	66.92	+6.74
5	0.05	66.67	+6.48
5	0.10	64.93	+4.75
5	0.50	43.95	−16.23

E.3 Direct Evaluation of the Mutual-Information Bound on Real Data

The calibration above controls privacy through $\mathrm{MI}(X; \mathcal{M}(X))$ but evaluates only the *prior* factor C; the mutual-information side is never examined on real data. As a computable diagnostic we track



(a) Original Basketball frame (b) After M: mix with Skiing (k=5) (c) After $+B$: Gaussian noise (σ=0.1)

Figure 2: Pixel-space visualization of the obfuscation pipeline at $(k, \sigma) = (5, 0.1)$.

the log-determinant functional of the released Gram matrix,

$$L(\sigma) \;:=\; \log\det\!\big(I_n + \sigma^{-2}\,\tilde{X}\tilde{X}^\top\big),$$

which shares the $\log\det(I + \Sigma_X/\sigma^2)$ structure that appears in the mutual-information analysis of Corollary 15 and measures how much of the data's second-order structure survives the additive noise. We state the caveats up front. $L(\sigma)$ is not the Corollary 15 Type-(II) bound itself. That term is a Jensen-gap expectation $\log\det\mathbb{E}_X[\,\cdot\,] - \mathbb{E}_X\log\det[\,\cdot\,]$, whereas L is a single-realization log-determinant, and we do not claim it certifies a numerical value of, or a rigorous upper bound on, $\mathrm{MI}(X;\mathcal{M}(X))$. We use it only as a proxy whose qualitative behavior in (k, σ) matches what the calibration assumes. We evaluate $L(\sigma)$ on 512-dimensional clip-level R3D-18 embeddings of a stratified UCF-101 subset ($n_0 = 50$ clips per class), a feature space distinct from the frame-level ResNet-18 encoder the attacks and downstream classifier use. The evaluation is therefore an independent probe of the mechanism on real video features rather than a measurement on the attacked representation, and it uses the Sylvester identity

$$\log\det\!\big(I_n + \sigma^{-2}\tilde{X}\tilde{X}^\top\big) = \log\det\!\big(I_{d_0} + \sigma^{-2}\tilde{X}^\top\tilde{X}\big)$$

to reduce the $n \times n$ determinant to a 512×512 one. Table 4 and Figure 3 report $\frac{d}{2}L(\sigma)$ per datapoint across (k, σ) (with $d = 500$).

Three properties emerge, each matching what the calibration assumes qualitatively. First, L is finite and well-behaved on real embeddings, falling from 240 nats/point at $\sigma = 0.01$ to 26 at $\sigma = 1.0$ for $k = 0$, rather than taking the effectively unbounded values the raw-pixel Gram matrix produces. Second, it decreases monotonically in σ and is roughly halved by class-1 mixing ($240 \to 118 \to 101$ nats/point at $\sigma = 0.01$ for $k = 0, 1, 5$), so both of the mechanism's levers (additive noise and mixing) reduce it, consistent with the roles they play in Theorem 14 and the mixing analysis of (Xiao et al., 2024). Third, within the operable band L decays like $\sim 1/\sigma$ rather than the $1/\sigma^2$ that the $\sqrt{\cdot}$ calibration of Theorem 4 assumes (the measured ratio across the $\sigma : 0.5 \to 1.0$ doubling is 0.54, versus 0.25 for a quadratic law), and the quadratic decay appears only in the large-σ tail. We flag this as a caution rather than a re-derivation. If the true mutual information shared this local scaling, the noise needed to reach a target would grow faster than $\sqrt{\cdot}$, so the $\sqrt{C_{\mathrm{half}}/C_{\mathrm{int}}} \approx 5.6\times$ calibration would be optimistic. That observation concerns a proxy's local slope, not an established bound, and we do not restate it as fact.

To be explicit, $L(\sigma)$ is a proxy, not a direct estimate of $\mathrm{MI}(X;\mathcal{M}(X))$ nor a certified upper bound on it, and a Monte-Carlo estimator of the true mutual information is left to future work.

Table 4: Log-determinant functional $\frac{d}{2}L(\sigma)$ (nats per datapoint; d=500) of the released Gram matrix, evaluated on clip-level R3D-18 UCF-101 embeddings across mixing k and noise σ. A computable proxy for the mutual-information analysis, not the Corollary 15 bound itself and not a certified MI bound. The functional is finite, monotone decreasing in σ, and reduced by mixing.

σ	$k=0$	$k=1$	$k=5$
0.01	239.7	118.5	101.0
0.05	158.2	78.1	60.8
0.10	123.3	60.9	44.1
0.25	78.8	38.9	24.6
0.50	49.0	24.2	13.8
1.00	26.3	13.0	7.0

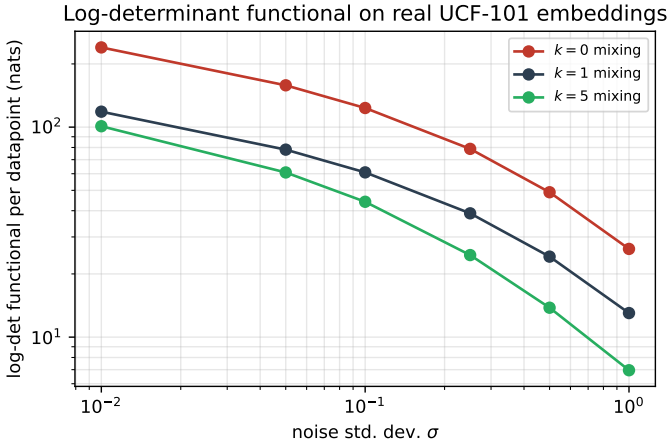


Figure 3: Log-determinant functional (a computable proxy for the mutual-information analysis, not a certified bound) per datapoint on real UCF-101 R3D-18 embeddings, log-log. Finite, decreasing in the noise σ, and reduced by class-k mixing; its $\sim 1/\sigma$ decay in the operable band is shallower than the $1/\sigma^2$ the calibration assumes.

What this establishes is modest but real. The log-determinant diagnostic the calibration implicitly relies on is finite and correctly ordered in both mechanism parameters on real video features.

E.4 Modern-Metric Evaluation: ROC and TPR at Low FPR

The weighted and binary scores of Section 6 are average-case summaries. Carlini *et al.* (Carlini et al., 2022) argue that an average score can hide the property that actually matters for privacy: whether the attack is *confident* on a subset of examples, i.e. whether it attains a high true-positive rate (TPR) at a low false-positive rate (FPR). We therefore re-express the *same* per-candidate LiRA scores as a membership-inference ROC and report AUC together with TPR at FPR=1% and 0.1%. To make scores comparable across queries we standardize each candidate's score by the mean and standard deviation of that query's non-member scores (an offline-LiRA null), then pool across all 90 queries (15 targets $\times$ 6 trials) and sweep one global threshold; the reported AUC agrees with the per-query mean AUC to within 0.02.

We evaluate three adversaries that differ only in the non-member set they must reject, an axis that isolates the effect of temporal and semantic similarity:

- **Informed, within-clip (index inference).** Members are the T=16 frames of the target clip

that entered U; non-members are the other 84 frames of the *same* clip. This is the hardest, most on-point test: the adversary knows the clip and must separate members from their own temporal neighbors.

- **Uninformed, same-class.** Non-members are same-class frames from *other* clips (the class confound removed).
- **Uninformed, cross-class.** Non-members are drawn from the broad universe (mostly other classes), the standard frame-identification setting.

Table 5 and Figure 4 report the result at the operating point σ=0.10, and three findings follow. First, identification is confident but within-clip localization is not. The uninformed cross-class adversary reaches AUC 0.90–0.98 and TPR=0.74 at FPR=1% (a 74× lift over the 1% random rate), so the mechanism does not hide which frames are members when the distractors are unrelated. The informed within-clip adversary, in contrast, attains AUC 0.55–0.61 with TPR $\approx$0.02–0.03 at FPR=1%, barely above the 0.01 random rate, and TPR <0.01 at FPR=0.1%. Knowing the clip and being asked to localize the exact member frames among their temporal neighbors, the adversary is essentially guessing at low false-positive rates. Second, the monotone gradient (informed $\ll$ same-class $\lesssim$ cross-class) shows attack success is governed by how similar the non-members are to the members. The more the distractors resemble near-miss frames, the closer the attack falls to chance. This is the temporal-correlation effect analyzed in Section 7, now quantified in the metric membership-inference practice trusts, and the weighted-scoring formulation of Section 6 was introduced to credit this behavior. Third, mixing helps and noise (within the operable band) does not. Increasing mixing from k=0 to k=1 lowers the cross-class AUC from 0.98 to 0.90 (TPR@1% 0.74 $\rightarrow$ 0.60), echoing the mutual-information reduction from mixing in Section E.3, though the effect is not strictly monotone in k. The rank-based metrics, by contrast, are nearly invariant to σ across $[0, 2.0]$ (e.g. cross-class AUC at k=0 moves only 0.981 $\rightarrow$ 0.981 $\rightarrow$ 0.951 from σ=0 to 2.0), echoing the Appendix F.1 finding that the attacks stay above chance even at large σ. Within the operable band, additive noise is not the operative empirical privacy lever, since it barely moves rank-based attack success. The privacy value these experiments demonstrate comes from mixing (which lowers the attack AUC, as reported above) and from the structural result that weighted scoring raises the adversary's prior (Section 6). The additive-noise calibration of Section 6 is a statement about the formal MI bound, whose empirical force in this band is limited, and its numerical looseness is discussed in Section E.3.

E.5 Privacy–Utility Trade-off

The main paper reports leakage (Table 2) and accuracy (Table 3) separately; here we place them on a common axis to ask the operational question a practitioner faces: is there a mechanism setting that is simultaneously more private *and* at least as useful as no obfuscation? Table 6 and Figure 5 plot the uninformed-adversary leakage (Attack 2 of Table 2) against downstream test accuracy across the operable band $\sigma \leq 0.10$, taking the no-obfuscation point (leakage 0.835, accuracy 60.19%) as the reference.

The no-obfuscation point is Pareto-dominated. The configuration $(k, \sigma) = (1, 0.10)$ cuts uninformed-adversary leakage by 47.5% (0.835 $\rightarrow$ 0.439) and raises accuracy by 6.4 percentage points (60.19 $\rightarrow$ 66.60%), improving on both axes. The lever is mixing rather than noise. At the fixed noise $\sigma = 0.10$, increasing mixing from $k = 0$ to $k = 1$ deepens the leakage reduction from 14% to 47% (0.715 $\rightarrow$ 0.439) at essentially unchanged accuracy, whereas raising σ within the band trades accuracy for privacy only weakly. Section E.3 gives the empirical counterpart, in which mixing rather

Table 5: LiRA ROC metrics at the operating point $\sigma = 0.10$, across mixing k, for the three adversaries. TPR is reported at FPR=1% and 0.1%; "Adv." is $\max_t(\text{TPR}-\text{FPR})$. Random guessing gives AUC 0.5, TPR 0.01/0.001. Pooled over 90 queries (15 targets × 6 trials).

k	AUC	TPR@1%	TPR@0.1%	Adv.
Informed (within-clip)				
0	0.610	0.024	0.004	0.175
1	0.554	0.032	0.009	0.093
5	0.584	0.026	0.004	0.130
Uninformed (same-class)				
0	0.981	0.387	0.081	0.895
1	0.908	0.371	0.128	0.692
5	0.979	0.391	0.116	0.903
Uninformed (cross-class)				
0	0.981	0.744	0.584	0.848
1	0.904	0.599	0.413	0.698
5	0.979	0.735	0.454	0.847

Table 6: Privacy–utility trade-off across the operable band. "Leakage" is the uninformed adversary (Attack 2 of Table 2); "Δ leak" is the relative reduction versus the no-obfuscation leakage 0.835; accuracy and its gap are from Table 3. Bold marks the Pareto-dominant point.

k	σ	Leakage	Δ leak	Δ acc (pp)
No obfuscation		0.835	—	—
0	0.05	0.751	−10.0%	+12.16
0	0.10	0.715	−14.4%	+8.05
1	0.01	0.565	−32.3%	+9.81
1	**0.10**	**0.439**	**−47.5%**	**+6.41**
5	0.10	0.648	−22.5%	+4.74

than noise produced the largest reduction in the mutual-information bound. Mixing lowers both the information bound and the realized attack at little utility cost.

We frame the result as a Pareto improvement, not a claim of strong absolute privacy. Even at $(1, 0.10)$, the residual leakage 0.439 is still $\approx 55\times$ the closed-form chance baseline of 0.008 (Section 6); a substantial reduction in leakage accompanied by a utility gain is the relevant comparison against the deploy-or-not baseline, not evidence that the mechanism drives leakage to chance within the operable band.

The two axes also carry unequal evidential weight. The leakage side is estimated from $n = 50$ Monte-Carlo samples per cell (10 targets × 5 trials) and the $0.835 \to 0.439$ gap is large relative to the reported standard errors (0.039 and 0.050). The accuracy side is not: each (k, σ) cell is a *single* training run at seed 42 (Section E.1), so the +6.4 pp gap carries no estimate of seed-to-seed variance and we have not subjected either axis to a significance test. The Pareto claim should therefore be read as holding firmly on the privacy axis and as a single-run observation on the utility axis. Multi-seed utility runs are supported by the released code and are the natural next measurement.

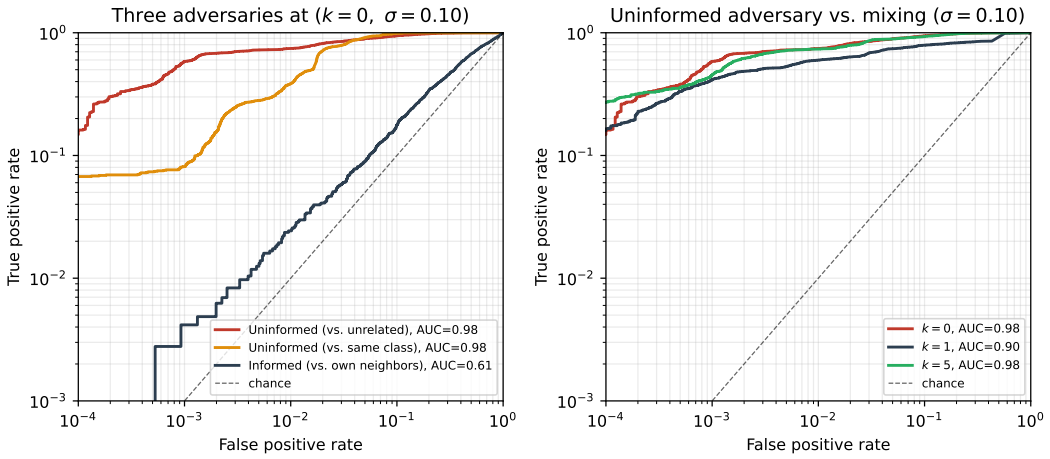


Figure 4: Membership-inference ROC (log-log) of the LiRA attacks. **Left:** at (k=0, σ=0.10), the informed within-clip adversary (separating members from their own temporal neighbors) tracks the chance diagonal (AUC 0.61), while the uninformed identification adversaries are confident (AUC 0.98). Attack success is set by distractor similarity rather than by the noise. **Right:** mixing reduces the uninformed adversary's confidence (k=0 $\to$ 1: AUC 0.98 $\to$ 0.90), non-monotonically in k.

E.6 Empirical Autocorrelation of Frame Embeddings

To ground the preceding analysis, we measure the temporal autocorrelation of ResNet-18 embeddings on UCF-101. For each clip with frame embeddings $z_0, z_1, \ldots, z_{L-1} \in \mathbb{R}^{d_0}$, we compute the average cosine similarity at lag τ:

$$\rho_{\text{clip}}(\tau) = \frac{1}{L-\tau} \sum_{t=0}^{L-\tau-1} \frac{z_t^\top z_{t+\tau}}{\|z_t\| \, \|z_{t+\tau}\|},$$

and average over 200 retained clips (first $L = 100$ frames each) to obtain $\rho(\tau)$.

The embeddings are correlated and decay slowly (Figure 6). We find $\rho(1) \approx 0.78$, decaying to $\rho(10) \approx 0.21$ and crossing zero near $\tau \approx 20$. Fitting an AR(1) model $\rho(\tau) \approx \alpha^\tau$ gives $\alpha \approx 0.84$, confirming that the AR(1) instantiation used throughout this section is an accurate model of real frame embeddings rather than a mathematical convenience. Substituting the measured ρ into the aggregation factor of Theorem 12, $\bar{R}_b = 1 + \frac{2}{b}\sum_{\tau=1}^{b-1}(b-\tau)\rho(\tau)$, yields $\bar{R}_{16} \approx 7.4$ for a $b = 16$-frame clip: averaging 16 frames amplifies the signal-to-noise ratio by that factor, so independent per-frame noise is effectively reduced by $\sqrt{\bar{R}_{16}} \approx 2.7\times$ under temporal aggregation. This is a real-embedding instance of the degradation Theorem 12 predicts, and it quantifies the noise inflation a deployment must apply (or the correlated-noise mechanism of Theorem 13 it must adopt) to remain private against an adversary that aggregates across frames. The same measured ρ also confirms Theorem 8 numerically: forming the block-Toeplitz correlation matrix R from the autocorrelation and evaluating the mutual-information bound, the correlated-sampling bound I_R is 19% below the independent-sampling bound I_{I_n}, the reduction driven by the spread of R's eigenvalues (which sum to the block size, preserving the trace as the theorem's Jensen argument requires).

E.7 Empirical Illustration: Aggregation Attack and Correlated Noise

The preceding subsections establish that the correlation Theorems 8–13 assume is present in real embeddings. We now instantiate the aggregation adversary of Theorem 12 and the correlated-noise

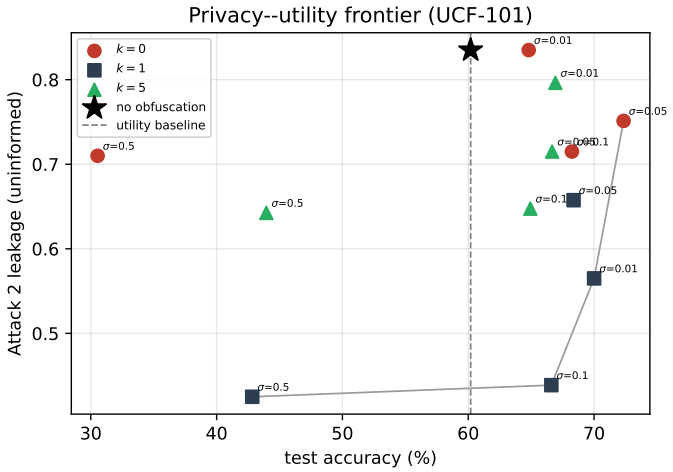


Figure 5: Privacy–utility plane: uninformed-adversary leakage (Attack 2, lower is more private) versus UCF-101 test accuracy. The no-obfuscation point (leakage 0.835, accuracy 60.19%) is Pareto-dominated by $(k, \sigma) = (1, 0.10)$, a setting that is both more private and more accurate. Mixing k, not noise σ, is the privacy lever within the operable band.

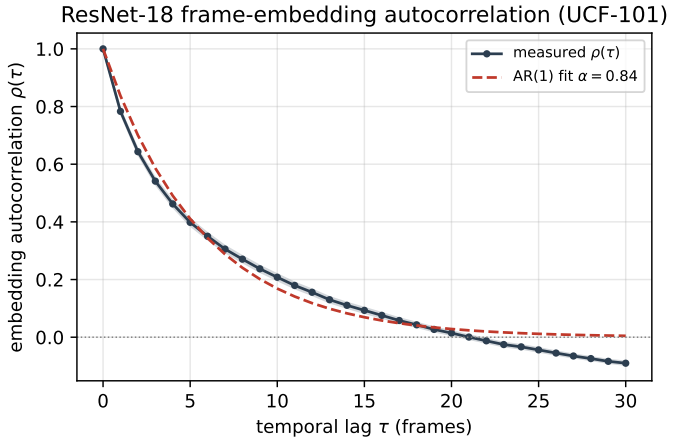


Figure 6: Measured temporal autocorrelation $\rho(\tau)$ of ResNet-18 frame embeddings on UCF-101 (mean over 200 clips, shaded ±SE), with the fitted AR(1) law. The slow decay confirms the correlation model of Definition 6.

defense of Theorem 13 directly, both on UCF-101 and on a controlled synthetic model in which the correlation is varied directly.

The aggregation adversary (Theorem 12). Theorem 12 shows that pooling evidence across correlated frames cancels independent noise. We realize such an adversary and compare it to the frame-independent attack the mechanism is calibrated against. Both use the same per-frame likelihood-ratio scores (LiRA; Carlini et al., 2022) on the released output and differ only in the decision rule: the *blind* adversary ranks all candidate rows individually and emits the top-T (the i.i.d. attack the calibration assumes), while the *correlation-aware* adversary aggregates the per-frame scores within each clip ($s_c = \sum_{r \in c} \ell_r$), identifies the highest-scoring clip, and emits its top-T frames, the empirical form of the temporal aggregation in Theorem 12. On UCF-101 (Table 7), the

Table 7: Blind vs. correlation-aware (Theorem 12 aggregation) attack on UCF-101, half-integer weighted recovery, and the correlation-aware source-clip top-1 out of 400 candidate clips (1/400 = 0.0025 chance). Mean over 75 (15 targets × 5 trials).

σ	Blind (half)	Aware (half)	Aware clip top-1
0.05	0.552	0.573	0.573
0.10	0.446	0.467	0.467
0.25	0.483	0.480	0.480
0.50	0.493	0.533	0.533

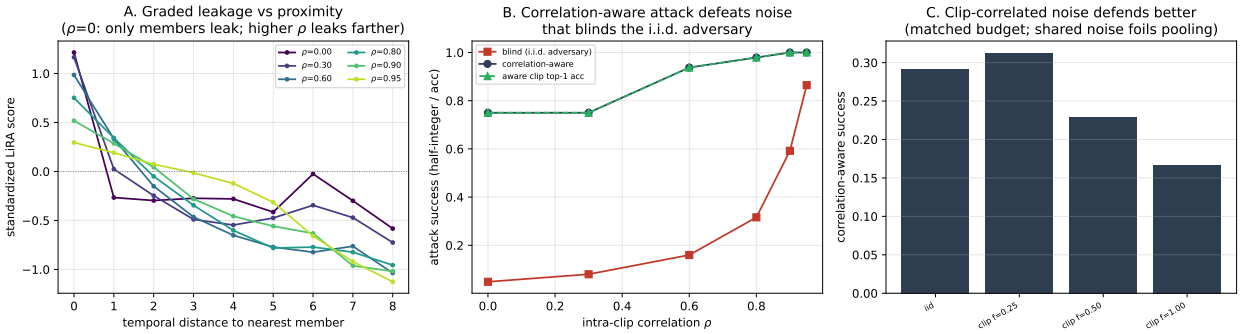


Figure 7: Controlled-correlation study on synthetic AR(1) clips. Left: graded leakage, likelihood-ratio score against temporal distance to the nearest member, flat at ρ=0 and reaching farther as ρ grows. Middle: the aggregation attack of Theorem 12 defeats a noise level that blinds the i.i.d. adversary, with the advantage growing in ρ. Right: at matched budget, the clip-correlated noise of Theorem 13 defends the aggregation adversary better than i.i.d. noise.

correlation-aware adversary identifies the source clip well above chance, reaching top-1 accuracy 0.47–0.57 out of 400 candidate clips against a 1/400 random baseline, and improves the weighted recovery over the blind attack at three of the four noise levels. The gain is modest on real data because the random per-clip frame sampling of Section 6 scatters the $T = 16$ members across the $L = 100$-frame clip, so the aggregation of Theorem 12 acts on a weakened intra-clip correlation; we isolate the full effect synthetically below.

Controlled study and correlated-noise defense (Theorem 13). To isolate the effect the real-data sampling mutes, we generate synthetic clips whose intra-clip correlation is set by a single parameter $\rho \in [0, 1)$. Each clip is an AR(1) sequence with $\mathrm{Corr}(\mathrm{frame}_s, \mathrm{frame}_t) = \rho^{|s-t|}$, matching Definition 6, and we pass them through the same mechanism and scoring (Figure 7). Three effects emerge, each matching the theory. First, *graded leakage is real*: the likelihood-ratio score of a non-member frame rises as its temporal distance to the nearest member shrinks, with reach growing in ρ; at ρ=0 only members score above baseline, so the partial credit the weighted formulation assigns tracks measurable leakage. Second, the aggregation attack's source-clip identification rises from 0.75 at ρ=0 to 1.00 at $\rho \geq 0.9$ while the blind adversary lags ($0.05 \to 0.87$), so the aggregation exploits the correlation the calibration treats as absent, confirming Theorem 12. Third, and validating Theorem 13, at matched noise budget and ρ=0.8 we vary the fraction f of the per-frame noise that is shared within a clip; moving from independent noise (f=0) to fully clip-correlated noise (f=1) reduces the aggregation attack's success from 0.29 to 0.17 while leaving the blind adversary unchanged. The shared component sums coherently under the adversary's pooling and cannot be averaged out, consistent with Theorem 13.

Together, these experiments turn Section 7 from prediction into measurement: the correlation the theory assumes is present and slowly-decaying in real embeddings (§E.6), the aggregation vulnerability of Theorem 12 is realized by a concrete attack, and the correlated-noise remedy of Theorem 13 measurably closes it.

F Supplementary Experimental Results

This appendix supplements the operating-point results of Section 6 with the full noise sweep and a closed-form analysis of mixing intensity. None of the material here is required for the paper's central claims; we include it for completeness and as sanity checks.

F.1 Full (k, σ) sweep

Table 8 reports the two temporal membership inference scores across $k \in \{0, 1, 5\}$ and $\sigma \in \{0, 0.01, 0.05, 0.10, 0.50\}$, plus two σ-extension cells at $k = 5$ and $\sigma \in \{2.00, 5.00\}$. Section 6 identifies $\sigma \leq 0.10$ as the mechanism's operable noise range on the downstream task; $\sigma = 0.50$ reduces utility to chance level, and the σ-extension cells lie far beyond any utility-preserving setting.

Table 8: Empirical temporal membership inference scores across the full (k, σ) sweep.

k	σ	Attack 1	Attack 2
0	0.00	0.460 (0.019)	0.835 (0.039)
0	0.01	0.460 (0.019)	0.835 (0.039)
0	0.05	0.441 (0.017)	0.751 (0.049)
0	0.10	0.484 (0.018)	0.715 (0.036)
0	0.50	0.489 (0.016)	0.710 (0.051)
1	0.00	0.403 (0.021)	0.565 (0.060)
1	0.01	0.403 (0.021)	0.565 (0.060)
1	0.05	0.509 (0.013)	0.657 (0.055)
1	0.10	0.454 (0.020)	0.439 (0.050)
1	0.50	0.408 (0.015)	0.425 (0.061)
5	0.00	0.436 (0.019)	0.797 (0.045)
5	0.01	0.439 (0.019)	0.796 (0.045)
5	0.05	0.469 (0.018)	0.715 (0.052)
5	0.10	0.482 (0.019)	0.648 (0.042)
5	0.50	0.445 (0.018)	0.643 (0.056)
5	2.00	0.376 (0.017)	0.423 (0.058)
5	5.00	0.404 (0.017)	0.214 (0.047)

Each cell is averaged over $n_{\text{targets}} = 10$ targets with $n_{\text{trials}} = 5$ Monte Carlo trials (50 samples); standard errors std/$\sqrt{50}$ are shown in parentheses.

F.2 Frame-survival closed-form analysis

For mixing intensity $k \geq 1$ and n_0 samples per class, each target clip appears in expectation $\mathbb{E}[\text{appearances}] = 2ck/n_0$ class-pair mixed samples, each carrying signal weight $1/(2k)$ from the k-fold averaging. The expected total signal mass is therefore

$$\mathbb{E}[\text{appearances}] \cdot \tfrac{1}{2k} \;=\; \frac{c}{n_0},$$

independent of k for $k \geq 1$. With $c = 101$ and $n_0 = 50$, this gives $c/n_0 = 2.02$ across all $k \geq 1$ (Table 9).

Table 9: Closed-form signal mass under class-k mixing.

k	$\mathbb{E}$[appearances]	weight	signal mass
0	1.00	1	1.00
1	4.04	1/2	2.02
2	8.08	1/4	2.02
3	12.12	1/6	2.02
5	20.20	1/10	2.02
10	40.40	1/20	2.02

A first-moment argument therefore predicts no variation in attack difficulty as k varies among $k \geq 1$: each target's total signal is preserved in aggregate.