EXPOSING MEMBERSHIP LEAKAGE IN TEST-TIME ADAPTATION VIA MANIFOLD CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) improves model robustness under distribution shifts by updating pretrained models on unlabeled test data during inference. However, test samples may leave detectable traces during model adaptation, raising a privacy concern that remains largely unexplored. Existing sequential membership inference attacks (MIAs) consider training or model-update trajectories, but do not account for the distribution-driven adaptation dynamics in TTA that can obscure candidate-specific membership traces. To address this gap, we propose the Manifold Calibration Attack (MCA), a label-free, trajectory-based attack designed for sequential TTA that separates candidate-specific update effects from shared adaptation dynamics. Specifically, MCA calibrates each candidate's stepwise response against a local manifold neighborhood to remove shared adaptation motion, localizes candidate-relevant changes along the adaptation trajectory, and uses nuisance-aware Jacobian attribution to measure how strongly the selected updates align with the candidate's functional subspace. Our analysis reveals that membership evidence is often stronger in early adaptation across different TTA methods. Extensive experiments across 13 settings spanning multiple datasets, model architectures, corruption streams, and representative TTA methods show that MCA achieves the best macro-average ROC-AUC and low-FPR detection performance among the evaluated attacks.

1 INTRODUCTION

Test-time adaptation (TTA) enables a pretrained model to adapt to distribution shifts using unlabeled samples encountered during deployment (Schneider et al., 2020; Wang et al., 2020; Yuan et al., 2023; Liang et al., 2025). This capability is particularly valuable in non-stationary environments, where corruptions, sensors, or operating conditions may evolve after deployment and labeled target data are unavailable. Recent methods have expanded TTA beyond normalization-based adaptation to include reliability filtering, sharpness-aware optimization, teacher–student updates, and forward-only prompt optimization (Gong et al., 2022; Niu et al., 2022; 2023; Wang et al., 2022; Döbler et al., 2023; Niu et al., 2024), broadening its applicability across model architectures and deployment scenarios.

However, this adaptivity introduces a new privacy concern. Unlike a conventional inference service, where test samples only receive predictions, TTA uses incoming samples to update the deployed model. A test sample may therefore leave a detectable trace in subsequent model states even though its label is never collected. Such traces expose TTA systems to membership inference attacks (MIAs), in which an adversary uses access to a target model to determine whether a candidate sample contributed to its optimization. Early MIAs exploited differences in prediction confidence or loss between members and non-members (Shokri et al., 2017; Yeom et al., 2018), while subsequent work relaxed the reliance on shadow models, developed white-box attacks, and calibrated membership scores according to sample difficulty (Salem et al., 2018; Nasr et al., 2019; Sablayrolles et al., 2019; Watson et al., 2021). More recent likelihood- and reference-based attacks, including LiRA, RMIA, and BASE, provide increasingly powerful tests of training-set membership (Carlini et al., 2022; Zarifzadeh et al., 2023; Lassila et al., 2026).

Despite this progress, membership inference in TTA differs fundamentally from conventional training-set membership. Most MIAs operate on a final fixed model, although several studies have

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

exploited temporal information from intermediate or updated models. TrajectoryMIA and SeqMIA reconstruct sequences of training-stage metrics through knowledge distillation (Liu et al., 2022; Li et al., 2024), while attacks on multiple updated models combine membership evidence across successive model releases (Jagielski et al., 2023). More recent work directly formulates membership inference over a sequence of model updates(Michel et al., 2026). These studies establish that model evolution can strengthen membership inference, but do not account for the distribution-driven adaptation dynamics of TTA that can obscure candidate-specific membership traces. In TTA, membership evidence is temporally sparse because a candidate directly participates in only one adaptation batch; distribution-wide adaptation moves many samples together and can obscure the candidate-specific response; and different TTA algorithms update heterogeneous functional states, such as normalization statistics, affine parameters, prompts, or teacher–student components. An effective TTA attack must therefore localize when a candidate affects the trajectory, separate this effect from shared adaptation motion, and determine whether the corresponding update is functionally aligned with the candidate.

To address this gap, we propose the Manifold Calibration Attack (MCA), a label-free MIA tailored to sequential TTA. MCA first measures step-wise changes in predictions and representations and calibrates them against nearby reference samples, suppressing motion shared along the local data manifold. It then localizes candidate-relevant change points rather than aggregating the entire trajectory. Finally, MCA projects the residual functional update onto the candidate's Jacobian subspace and softly fuses this attribution evidence with the calibrated temporal response. This design separates candidate-local deformation from distribution-level adaptation while remaining applicable across heterogeneous TTA mechanisms.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to systematically study membership inference over sequential TTA streams. We show that unlabeled test samples can leave temporally localized, candidate-aligned traces through adaptation, revealing a distinct form of deployment-time privacy leakage.
- We introduce MCA, a label-free attack designed for sequential TTA. By suppressing shared manifold drift, localizing candidate-relevant updates, and attributing them through the candidate's functional subspace, MCA addresses a key limitation of existing final-state and sequential MIAs, which can confound transient candidate-specific evidence with global adaptation dynamics.
- We systematically evaluate membership inference across 13 diverse TTA settings, demonstrating the broad effectiveness and stability of MCA. Our analysis further reveals substantial variation in privacy leakage across different TTA mechanisms, providing a more comprehensive basis for privacy assessment in TTA.

2 Related Work

We first review the background and representative paradigms of TTA, followed by recent advances in MIAs.

2.1 Test-time Adaptation

TTA updates a source-trained model using unlabeled samples encountered after deployment. Early approaches adapt batch-normalization statistics to the target distribution (Schneider et al., 2020). TENT instead minimizes prediction entropy while updating normalization affine parameters (Wang et al., 2020); this principle can also be applied to LayerNorm-based backbones (TENT-LN). Subsequent methods improve reliability: EATA filters unreliable or redundant samples and regularizes important parameters (Niu et al., 2022), whereas SAR combines reliable entropy minimization with sharpness-aware optimization (Niu et al., 2023). Under continual shifts, CoTTA uses a teacher model, augmentation-averaged predictions, and stochastic parameter restoration (Wang et al., 2022), while RMT combines mean-teacher consistency with contrastive alignment (Döbler et al., 2023). FOA further enables derivative-free prompt adaptation using only forward passes (Niu et al., 2024). Collectively, these methods span statistics-, gradient-, teacher–student-, and prompt-based adaptation.

2.2 Membership Inference Attacks

MIAs determine whether a record contributed to a released model. Early attacks train shadow-model classifiers, threshold sample losses, or exploit white-box signals (Shokri et al., 2017; Yeom et al., 2018; Salem et al., 2018; Nasr et al., 2019). More powerful hypothesis-testing attacks include LiRA (Carlini et al., 2022), the lower-cost RMIA (Zarifzadeh et al., 2023), and BASE (Lassila et al., 2026). TrajectoryMIA and SeqMIA exploit metric sequences reconstructed from training states (Liu et al., 2022; Li et al., 2024), while update-aware attacks analyze privacy leakage across successive model releases (Salem et al., 2020; Jagielski et al., 2023). More recently, sequential MIA directly considers whether a target is introduced at a particular model-update step (Michel et al., 2026), while trajectory-based attribution methods such as TracIn use gradient alignment across optimization checkpoints (Pruthi et al., 2020). These works exploit training- or update-induced model evolution, but do not address the TTA-specific challenge of disentangling temporally localized candidate effects from distribution-driven adaptation dynamics shared across samples.

3 TTA MIA Analysis

TTA produces a sequence of model states from an unlabeled target stream, and differences between successive states can reveal information about the update data (Salem et al., 2020; Jagielski et al., 2023). Because a candidate enters only a particular adaptation batch, its membership signal may be temporally localized even when aggregate model utility changes little.

3.1 Candidate-Aligned Leakage from Sequential Adaptation

Let f_{θ_0} be the source model and $\mathcal{B}_{0:T-1} = \{B_0, \dots, B_{T-1}\}$ the target stream, with

$$\theta_{t+1} = \mathcal{A}(\theta_t, B_t), \qquad m(x) = \mathbb{I}\left[x \in \bigcup_{t=0}^{T-1} B_t\right]. \tag{1}$$

The attacker infers $m(x)$ without observing the batch identities. Let $\mathbf{s}_t$ denote the functional model state and $\Delta\mathbf{s}_t = \mathbf{s}_{t+1} - \mathbf{s}_t$. Under a local first-order approximation,

$$\mathbf{z}_{t+1}(x) - \mathbf{z}_t(x) = \mathbf{J}_t(x)\Delta\mathbf{s}_t + \mathcal{O}(\|\Delta\mathbf{s}_t\|_2^2), \qquad \mathbf{J}_t(x) = \frac{\partial \mathbf{z}_t(x)}{\partial \mathbf{s}_t}. \tag{2}$$

Thus, membership depends not only on update magnitude but also on whether the update follows a functional direction to which x is sensitive.

Let $m_t(x) = \mathbb{I}[x \in B_t]$ and model the update as

$$\Delta\mathbf{s}_t = \boldsymbol{\nu}_t + \zeta_t m_t(x)\boldsymbol{\psi}_t(x), \tag{3}$$

where $\boldsymbol{\nu}_t$ absorbs other samples' contributions and shared drift, and $\boldsymbol{\psi}_t(x)$ is the candidate-specific component. Let $\mathbf{Q}_t$ remove a nuisance subspace and $\mathbf{P}_t(x)$ project onto the row space of the residualized Jacobian $\mathbf{J}_t(x)\mathbf{Q}_t$. Define

$$a_t(x) = \|\mathbf{P}_t(x)\mathbf{Q}_t\Delta\mathbf{s}_t\|_2^2. \tag{4}$$

Conditioned on the candidate and pre-update state, assume that the projected nuisance has equal expected squared norm under $m_t(x) = 0$ and $m_t(x) = 1$ and is mean-orthogonal to the projected candidate component. Then

$$\begin{aligned} &\mathbb{E}[a_t(x) \mid m_t(x) = 1] - \mathbb{E}[a_t(x) \mid m_t(x) = 0] \\ &\quad = \zeta_t^2 \|\mathbf{P}_t(x)\mathbf{Q}_t\boldsymbol{\psi}_t(x)\|_2^2 \geq 0. \end{aligned} \tag{5}$$

The gap is strictly positive whenever the residualized candidate contribution has a nonzero projection onto the candidate's Jacobian row space, motivating directional attribution of model updates.

3.2 Manifold-Shared and Temporally Sparse Evidence

The update-space signal above is difficult to isolate because nearby samples share adaptation motion and the candidate-specific component occurs at only a few unknown steps. We model a scalar response as

$$\rho_t(x) = \mu_t(\mathbf{h}_t(x)) + m_t(x)\delta_t(x) + \varepsilon_t(x), \tag{6}$$

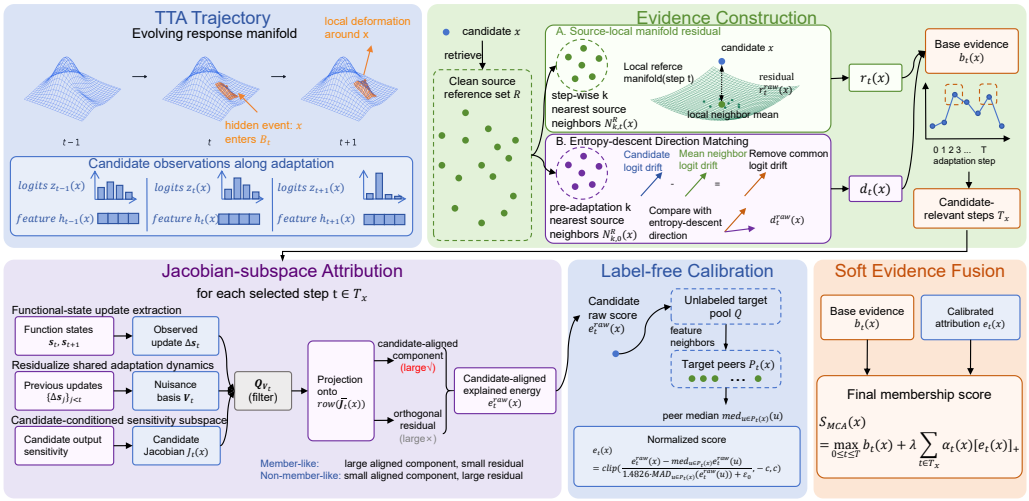


Figure 1: Overview of the MCA pipeline. Zoom in for detailed components and operations.

where μ_t is motion shared along the local data manifold, $\delta_t(x)$ is the candidate-specific response, and $\varepsilon_t(x)$ is zero-mean residual noise. If μ_t is L-Lipschitz and a disjoint reference neighborhood lies within radius ϵ_x of $\mathbf{h}_t(x)$, its average estimates the shared motion with bias at most $L\epsilon_x$, up to finite-reference error. Local subtraction therefore suppresses shared motion while retaining the candidate-specific response.

Privacy leakage need not track TTA utility because aggregate accuracy may remain unchanged under small candidate-conditioned changes in logits, features, or functional state. Moreover, final-state scores lose temporal information, global trajectory norms confound membership with shared drift, and uniform aggregation dilutes sparse evidence. MCA therefore combines local manifold calibration with candidate-specific temporal localization before applying directional attribution at the selected steps.

4 Attack Methodology

The Manifold Calibration Attack (MCA) is a label-free MIA designed for sequential TTA, with its scoring pipeline summarized in Figure 1. We consider a passive, honest-but-curious model recipient, motivated by collaborative adaptation systems that exchange model updates while keeping input streams local. The attacker observes the intermediate model states $\{\theta_t\}_{t=0}^{T}$ and knows the architecture and adaptation procedure. Given an unlabeled candidate, it infers whether the candidate participated in adaptation, without accessing private inputs, batch identities, or candidate occurrence steps.

Under this threat model, MCA proceeds in three stages. First, it compares each candidate's prediction and representation changes with a local reference neighborhood to remove shared manifold motion. Second, it combines the calibrated response with candidate-conditioned directional evidence and selects plausible change points along the trajectory. Third, it applies nuisance-residualized Jacobian attribution at the selected steps and fuses the result with the temporal evidence. None of these operations uses membership labels or the true occurrence step. The source code and implementation details are available at: `https://anonymous.4open.science/r/MCA-4BD3/`.

4.1 Manifold-Temporal Evidence

Local response calibration. For each $u \in \mathcal{Q} \cup \mathcal{R}$, MCA measures its response to update t using entropy reduction and feature displacement:

$$\rho_t(u) = \mathcal{Z}_t([\mathsf{H}(\mathbf{p}_t(u)) - \mathsf{H}(\mathbf{p}_{t+1}(u))]_+) + \mathcal{Z}_t(\|\mathbf{h}_{t+1}(u) - \mathbf{h}_t(u)\|_2), \tag{7}$$

where $[a]_+ = \max(a, 0)$ and $\mathcal{Z}_t$ denotes within-step standardization over the corresponding probe population. This response captures update-induced changes without assuming the objective of a particular TTA method. Let $\mathcal{N}^{\mathcal{R}}_{k,t}(x)$ denote the k nearest source references to x in normalized feature space. MCA removes their shared response through

$$r_t(x) = \rho_t(x) - \frac{1}{k} \sum_{u \in \mathcal{N}^{\mathcal{R}}_{k,t}(x)} \rho_t(u). \tag{8}$$

Thus, $r_t(x)$ retains candidate-local variation relative to the evolving local manifold.

For any candidate score $a_t(x)$, define robust within-step calibration over $\mathcal{Q}$ as

$$\mathrm{RZ}_t(a_t(x)) = \mathrm{clip}\left(\frac{a_t(x) - \mathrm{med}_{u \in \mathcal{Q}}\, a_t(u)}{1.4826\, \mathrm{MAD}_{u \in \mathcal{Q}}(a_t(u)) + \varepsilon_0}, -c, c \right), \tag{9}$$

where $\varepsilon_0 > 0$ is a numerical stabilizer.

Candidate-conditioned directional evidence. The residual magnitude alone does not indicate whether the update direction is specific to x. MCA therefore uses the entropy-descent direction

$$\mathbf{v}_t(x) = -\nabla_{\mathbf{z}_t(x)} \mathsf{H}(\mathbf{p}_t(x)) \tag{10}$$

as a method-agnostic, confidence-increasing template. For this branch, reference neighbors are fixed at $t = 0$ so that their identities are not conditioned on subsequent adaptation. The corresponding local logit residual is

$$\Delta \mathbf{z}^{\mathrm{loc}}_t(x) = \mathbf{z}_{t+1}(x) - \mathbf{z}_t(x) - \frac{1}{k} \sum_{u \in \mathcal{N}^{\mathcal{R}}_{k,0}(x)} \big(\mathbf{z}_{t+1}(u) - \mathbf{z}_t(u) \big). \tag{11}$$

Let $\mathbf{U}$ be an orthonormal basis of common source-reference logit drifts, including the softmax-invariant constant-logit direction, and define $\mathbf{Q}_{\mathbf{U}} = \mathbf{I} - \mathbf{U}^\top \mathbf{U}$. We remove these directions from both the observed residual and the template:

$$\bar{\mathbf{z}}_t(x) = \mathbf{Q}_{\mathbf{U}} \Delta \mathbf{z}^{\mathrm{loc}}_t(x), \qquad \bar{\mathbf{v}}_t(x) = \mathbf{Q}_{\mathbf{U}} \mathbf{v}_t(x). \tag{12}$$

Their gated positive agreement is

$$d_t(x) = g_t(x) \frac{[\langle \bar{\mathbf{z}}_t(x), \bar{\mathbf{v}}_t(x) \rangle]^2_+}{(\|\bar{\mathbf{z}}_t(x)\|_2^2 + \varepsilon_0)(\|\bar{\mathbf{v}}_t(x)\|_2^2 + \varepsilon_0)}, \tag{13}$$

where

$$g_t(x) = 0.25 + 0.75 \sigma(\mathrm{RZ}_t(\log(\|\bar{\mathbf{z}}_t(x)\|_2 + \varepsilon_0))) \tag{14}$$

downweights negligible residuals, for which directional agreement is unstable.

Temporal localization. MCA combines the locally calibrated magnitude and directional evidence as

$$\phi_t(x) = (1 - \beta)\, \mathrm{RZ}_t(r_t(x)) + \beta\, \mathrm{RZ}_t(d_t(x)). \tag{15}$$

To identify candidate-relevant changes, it compares $\phi_t(x)$ with the candidate's recent history and rewards short-term persistence:

$$\begin{aligned} \chi_t(x) &= \phi_t(x) - \mathrm{med}_{\max(0, t-W) \le j < t}\, \phi_j(x) \\ &\quad + \kappa \sum_{j=0}^{\min(F-1, T-1-t)} w_{t,j} \phi_{t+j}(x), \\ b_t(x) &= \mathrm{RZ}_t(\chi_t(x)), \qquad \mathcal{T}_x = \mathrm{Top}_{K_{\mathrm{step}}} \{b_t(x)\}_{t=0}^{T-1}. \end{aligned} \tag{16}$$

Here, the history term is zero at $t = 0$, and $w_{t,j} \propto \varrho^j$ are normalized over the available future horizon. The set $\mathcal{T}_x$ does not estimate the true batch index directly; it restricts expensive attribution to the K_{step} most plausible updates for candidate x.

4.2 JACOBIAN ATTRIBUTION AND INFERENCE

Nuisance-residualized attribution. For each $t \in \mathcal{T}_x$, MCA vectorizes the relevant functional state, including normalization parameters and buffers, prompt state, and the terminal head, as $\mathbf{s}_t$. Let $\mathbf{V}_t$ be an orthonormal basis of preceding functional updates and define $\mathbf{Q}_{\mathbf{V}_t} = \mathbf{I} - \mathbf{V}_t^\top \mathbf{V}_t$ to remove directions repeatedly shared across steps.

MCA constructs L_{out} orthonormal output probes $\mathbf{a}_{t,\ell}(x)$ without using labels, including a top-class contrast, a centered-logit direction, and deterministically generated orthogonal directions. For each probe, define

$$\mathbf{j}_{t,\ell}(x) = \nabla_{\mathbf{s}_t} \langle \mathbf{a}_{t,\ell}(x), \mathbf{z}_t(x) \rangle. \tag{17}$$

Let $\bar{\mathbf{J}}_t(x)$ stack the residualized rows $\mathbf{j}_{t,\ell}(x)^\top \mathbf{Q}_{\mathbf{V}_t}$ and let $\bar{\Delta}\mathbf{s}_t = \mathbf{Q}_{\mathbf{V}_t} \Delta \mathbf{s}_t$. With $\mathbf{u}_t(x) = \bar{\mathbf{J}}_t(x) \bar{\Delta}\mathbf{s}_t$, MCA computes

$$e_t^{\text{raw}}(x) = \frac{\mathbf{u}_t(x)^\top \left(\bar{\mathbf{J}}_t(x) \bar{\mathbf{J}}_t(x)^\top + \xi \mathbf{I} \right)^{-1} \mathbf{u}_t(x)}{\|\bar{\Delta}\mathbf{s}_t\|_2^2 + \varepsilon_0}. \tag{18}$$

This ratio measures the fraction of residual update energy explained by functional directions that affect the candidate, rather than the raw update magnitude.

Label-free calibration and inference. Because the scale of $e_t^{\text{raw}}(x)$ varies across architectures, state types, and steps, MCA calibrates it against an unlabeled peer set $\mathcal{P}_t(x)$:

$$e_t(x) = \text{clip}\left(\frac{e_t^{\text{raw}}(x) - \text{med}_{u \in \mathcal{P}_t(x)} \, e_t^{\text{raw}}(u)}{1.4826 \, \text{MAD}_{u \in \mathcal{P}_t(x)}(e_t^{\text{raw}}(u)) + \varepsilon_0}, -c, c \right). \tag{19}$$

Dense terminal-head scores use target neighbors fixed at the pre-adaptation state, whereas sparse normalization or prompt probes use other candidates selected at the same step, with source/null probes as a fallback.

Finally, the temporal score softly weights attribution evidence from the selected steps:

$$\begin{aligned} \alpha_t(x) &= \frac{\exp(b_t(x)/\tau)}{\sum_{j \in \mathcal{T}_x} \exp(b_j(x)/\tau)}, \\ S_{\text{MCA}}(x) &= \max_{0 \le t < T} b_t(x) + \lambda \sum_{t \in \mathcal{T}_x} \alpha_t(x) [e_t(x)]_+, \\ \widehat{m}(x) &= \mathbb{I}[S_{\text{MCA}}(x) > \gamma]. \end{aligned} \tag{20}$$

The positive part prevents negative attribution evidence from suppressing the more robust temporal score. All MCA hyperparameters are fixed across datasets and TTA methods without case-specific tuning; their default values are reported in Sec. 5.1. Restricting Jacobian computation to $K_{\text{step}} \ll T$ candidate-relevant steps keeps MCA practical across the evaluated TTA settings.

5 EXPERIMENTS

We evaluate whether MCA reliably exposes membership leakage across diverse TTA algorithms and architectures. We further study which components create the signal, how robust the attack is to deployment choices and reference mismatch, and whether differential privacy mitigates the leakage.

5.1 EXPERIMENTAL SETUP

Datasets, models, and TTA methods. Our benchmark contains 13 cases spanning CIFAR-10-C, CIFAR-100-C, ImageNet-C, and Mixed ImageNet-CKrizhevsky et al. (2009); Deng et al. (2009); Hendrycks & Dietterich (2019). M1–M5 use ResNet-18 on CIFAR-10-C with BN Adapt, TENT, EATA, SAR, and CoTTA; M6–M7 use WRN-28-10 on CIFAR-100-C with TENT and SAR; M8–M9 use ResNet-50 on ImageNet-C with TENT and SAR; and M10–M11 use ViT-B/16 with TENT-LN and FOA. Finally, M12–M13 evaluate RMT on mixed streams using ResNet-50 and Swin-T. The single-corruption cases use Gaussian noise at a severity of 3, whereas the mixed cases interleave all 15 ImageNet-C corruptions. This suite covers statistics-, gradient-, sharpness-, teacher–student-, prompt-, and full-model adaptation.

Membership and attack protocol. The adaptation stream is processed once. The adaptation stream consists of 4,096 samples and is processed once. For membership inference evaluation, we select 256 members from this stream and 256 non-members from the same target distribution. For M1–M7, the stream is divided with a batch size of 64; for M8–M13, it is divided with a batch size of 32. The non-members are disjoint from the adaptation stream and are never used for adaptation. Membership labels and member occurrence steps are used only for evaluation and are unavailable to the attacker. We reserve a clean, disjoint source reference set for manifold calibration, containing 3,000 examples for CIFAR and 512 for ImageNet. Ground-truth class labels are used only to report post-hoc model utility. Unless otherwise stated, MCA uses $k = 30$, $K_{\text{step}} = 2$, $\lambda = 2.0$, $\tau = 0.5$, $\beta = 0.5$, $W = 6$, $F = 3$, $\kappa = 0.35$, and $\varrho = 0.6$. These values are fixed across all datasets and TTA methods without case-specific tuning.

Baselines and metrics. We compare MCA with four direct TTA signals (final entropy, entropy drop, feature drift, and a global trajectory score) and established MIAs: LiRA (Carlini et al., 2022), online/offline RMIA (Zarifzadeh et al., 2023), SeqMIA (Li et al., 2024), online/offline BASE (Lassila et al., 2026), and SeMI (Michel et al., 2026). We adapt these MIAs to the TTA setting by replacing their conventional target, shadow, or reference training procedures with matched TTA streams initialized from the same source checkpoint. LiRA, RMIA, and BASE retain their original likelihood-, ratio-, and posterior-based scoring rules, respectively. For SeqMIA, the target and shadow teachers are first adapted through TTA, after which its original distillation-based metric sequence and attention-LSTM classifier are applied. We adapt SeMI using loss drops (black-box) and approximate likelihood ratios based on entropy gradients (white-box), maximizing scores over updates for unknown stream positions. The primary metric is ROC-AUC; we additionally report TPR at 5% FPR, which emphasizes the low-false-positive regime, and attack accuracy where relevant. Main results are reported as means and sample standard deviations over seeds 7, 17, and 27. "Avg." denotes the unweighted macro-average over the 13 cases.

5.2 Main Results

Table 1: Membership inference performance across all TTA cases over three independent seeds. We report mean ± standard deviation for ROC-AUC and TPR at 5% FPR. MCA denotes our proposed attack. Best and second-best mean results in each column are shown in **bold** and underlined, respectively. Ties after rounding receive the same formatting.

(a) ROC-AUC

	CIFAR-10-C					CIFAR-100-C		ImageNet-C / Mixed ImageNet-C						Avg.
Attack	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	
Final entropy	.491±.027	.513±.010	.496±.023	.499±.017	.564±.017	.502±.014	.497±.012	.591±.015	.494±.031	.541±.022	.494±.018	.507±.005	.515±.026	.516±.008
Entropy drop	.501 ± .031	.504±.022	.490±.009	.505±.017	.559±.010	.497±.009	.490±.013	.603±.014	.492±.013	.584±.008	.494±.016	.515±.019	.507±.021	.519±.007
Feature drift	.480±.035	**.540 ± .030**	.516±.027	.526 ± .029	.616±.010	.509±.003	.516 ± .029	.511±.028	.519 ± .038	.506±.030	.497±.024	.489±.010	.493±.022	.517±.018
Trajectory	.480±.050	.527±.019	.526 ± .032	**.531 ± .043**	**.678 ± .016**	.508±.005	.512±.014	.585±.023	**.520 ± .038**	.559±.027	.500±.049	.480±.041	.510±.034	.532±.023
LiRA	.475±.011	.488±.020	**.590 ± .018**	.512±.054	.647 ± .024	.512±.035	.494±.023	.595±.035	.517±.039	**.657 ± .025**	.504 ± .042	.516±.011	.564±.008	.544 ± .012
RMIA-Online	.500±.000	.497±.022	.504±.007	.508±.027	.485±.040	.494±.037	.511±.029	.506±.018	.505±.013	.507±.028	.497±.010	.521±.010	.596±.026	.510±.004
RMIA-Offline	.492±.026	.508±.018	.511±.021	.512±.017	.498±.012	.478±.014	.472±.008	.509±.016	.507±.019	.512±.009	.490±.015	.526±.014	.594±.024	.508±.001
SeqMIA	.500±.042	.492±.028	.496±.028	.505±.007	.601±.042	.504±.006	.481±.019	.555±.029	.498±.008	.524±.042	.482±.037	.509±.024	.499±.038	.511±.006
BASE-Online	.496±.039	.471±.014	.507±.040	.490±.019	.429±.018	.484±.023	.515±.007	.525±.034	.503±.013	.503±.005	.477±.020	.491±.003	.605 ± .022	.500±.002
BASE-Offline	.484±.020	.488±.018	.510±.044	.503±.035	.466±.025	.519 ± .027	**.523 ± .002**	.522±.047	.483±.039	.525±.022	.476±.034	.495±.007	.577±.016	.505±.008
SeMI (black-box)	.490±.019	.498±.006	.499±.027	.500±.000	.517±.023	.490±.055	.497±.068	.527±.011	.504±.007	.486±.007	.499±.012	.533 ± .049	.499±.014	.503±.013
SeMI (white-box)	.490±.028	.531±.013	.499±.030	.486±.025	.519±.038	.515±.024	.506±.016	.647 ± .011	.508±.036	.635 ± .022	.488±.024	.521±.015	.547±.023	.530±.006
MCA (ours)	**.507 ± .048**	.539 ± .028	.516±.001	.498±.004	.622±.017	**.531 ± .006**	.509±.016	**.684 ± .026**	.513±.037	.593±.020	**.511 ± .004**	**.622 ± .031**	**.649 ± .025**	**.561 ± .006**

(b) TPR at 5% FPR (%)

	CIFAR-10-C					CIFAR-100-C		ImageNet-C / Mixed ImageNet-C						Avg.
Attack	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	
Final entropy	3.906±1.033	6.641±2.734	5.599±2.151	5.339±1.193	8.073±3.181	7.943 ± 2.355	5.339±1.761	8.333±5.202	6.771 ± 3.322	3.516±1.353	4.427±1.927	7.812 ± 2.734	6.641±2.344	6.180±.705
Entropy drop	3.776±.983	4.818±3.501	6.510 ± 1.579	2.083±1.966	6.380±2.932	2.995±.597	4.036±2.151	10.026±3.630	5.859±3.405	9.375±1.562	4.036±2.255	7.422±1.408	5.208±2.005	5.579±.346
Feature drift	4.427±.597	**7.812 ± 1.953**	5.469±1.703	3.906±3.726	**23.958 ± 4.693**	5.078±1.033	6.380±.451	5.469±1.033	5.990±.983	3.516±.391	3.646±.983	6.641±1.408	4.818±1.479	6.701±.138
Trajectory	4.948 ± 3.181	5.729±1.761	6.510 ± 1.479	4.818±.983	20.703 ± 4.800	5.990±.813	6.380±1.372	9.245±.597	5.990±1.579	8.203±1.033	5.078±1.562	4.557±1.579	3.646±.983	7.061±.693
LiRA	3.125±1.033	4.818±1.579	4.688±1.033	**7.422 ± 4.297**	18.229±.813	3.906±2.175	4.818±1.256	11.458±2.151	6.771 ± 2.771	**21.484 ± 2.439**	**5.990 ± 2.744**	4.948±1.256	9.766±1.703	8.263 ± .599
RMIA-Online	.000±.000	4.688±.781	1.562±2.706	1.953±2.439	3.255±1.579	5.469±2.949	6.771 ± 1.927	7.031±4.225	4.167±2.932	3.385±2.481	.521±.226	7.161±2.151	10.417±3.523	4.337±.571
RMIA-Offline	4.427±.902	.000±.000	6.510 ± 1.966	5.859 ± 1.353	6.510±1.479	3.255±1.761	2.474±.226	7.943±6.266	4.297±1.172	5.339±2.542	.781±.391	6.641±2.706	11.979 ± 1.846	5.078±.601
SeqMIA	4.427±.451	4.557±.813	4.557±1.256	5.339±1.804	16.927±5.158	5.469±1.790	6.250±1.562	6.901±1.846	4.818±2.005	4.688±3.405	4.688±1.562	5.339±3.133	2.604±.902	5.889±.472
BASE-Online	4.297±.677	4.948±1.193	4.948±1.626	4.818±1.761	4.818±1.626	6.641±1.033	**7.682 ± 2.005**	6.771±3.754	6.510±1.927	2.995±1.256	5.208 ± 2.542	3.385±.597	11.849±2.221	5.759±.558
BASE-Offline	2.474±.451	6.120±1.761	**7.682 ± 3.501**	4.948±1.579	11.328±1.562	5.339±1.372	5.078±1.353	13.021±2.659	5.599±1.579	13.151 ± 4.660	4.297±1.408	4.427±1.256	9.505±.813	7.151±.797
SeMI (black-box)	4.167±.597	2.604±1.579	5.990±.597	.000±.000	9.766±4.078	4.688±4.134	3.516±2.067	6.641±2.175	5.599±1.372	5.078±2.439	4.297±1.033	**9.766 ± 4.225**	4.297±1.562	5.108±.830
SeMI (white-box)	**5.078 ± 1.562**	4.688±1.790	4.297±1.033	4.948±1.193	5.208±1.579	**8.464 ± 2.221**	6.120±1.579	14.974 ± 1.256	4.167±.813	8.203±.391	4.427±1.579	4.297±.677	5.859±1.703	6.210±.483
MCA (ours)	4.948 ± 1.479	7.682 ± 4.774	6.120±2.826	1.432±2.481	17.708±1.193	4.818±1.966	6.510±3.413	**23.047 ± 5.078**	**7.422 ± 2.734**	12.370±3.322	5.208 ± .597	5.339±2.826	**21.875 ± 6.454**	**9.575 ± .800**

Notes. Case definitions are provided in Sec. 5.1. "Avg." denotes the unweighted macro-average over all 13 cases. Values are reported as mean ± sample standard deviation over seeds 7, 17, and 27.

MCA improves overall attack performance. Table 1 compares MCA with direct TTA signals and established MIAs. MCA achieves the highest macro-average ROC-AUC of 0.561, outperforming the strongest baseline, LiRA (0.544), by 0.017 in absolute terms. It also obtains the highest average TPR at 5% FPR, 9.575%, compared with 8.263% for LiRA. Across the 13 cases, MCA

achieves the best mean AUC in six cases, demonstrating that its advantage extends across different architectures and adaptation mechanisms.

The gains are particularly pronounced in the ImageNet-scale and mixed-corruption settings. MCA reaches AUCs of 0.684, 0.622, and 0.649 on M8, M12, and M13, exceeding the corresponding runners-up by 0.037, 0.089, and 0.044 in the table. White-box SeMI ranks second in AUC on M8 and M10 and exceeds MCA on M10, while black-box SeMI remains near chance on average. RMIA and BASE use prediction-based ratio and posterior scores, where sample-specific effects can be diluted by later updates, while SeqMIA models a distillation-derived metric sequence. Although effective in their original settings, these methods are less well matched to the transient and sequential nature of TTA leakage. MCA retains such transient evidence through manifold calibration, temporal localization, and Jacobian attribution.

Performance nevertheless remains TTA-dependent. LiRA achieves the highest mean AUC on M3 and M10, while the trajectory baseline leads on M4 and M5; MCA is near chance on M4. Reliability filtering in EATA, sharpness-aware updates in SAR, and teacher–student adaptation with stochastic restoration in CoTTA may weaken or reshape sample-local traces, making them less compatible with a single attack signal. Despite this variation, MCA is the only evaluated attack that ranks first on both macro-average AUC and low-FPR TPR, supporting its broad effectiveness.

5.3 Ablation and Robustness Analysis

Table 2: Component ablation of MCA, macro-averaged across all 13 TTA cases using seed 7. Δ AUC is measured relative to the full method and computed from unrounded values. Best and second-best results in ROC-AUC, attack accuracy, and TPR@5% FPR are shown in **bold** and underlined, respectively.

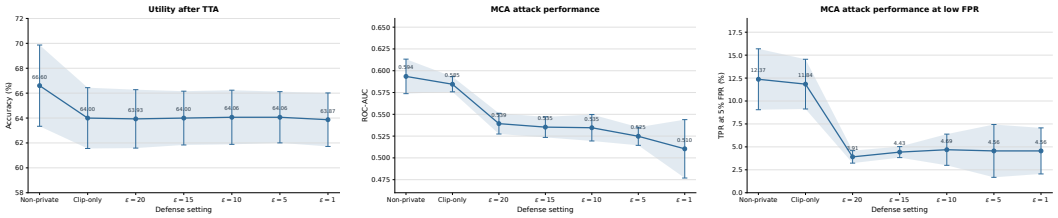
Variant	ROC-AUC	Δ AUC	Attack accuracy	TPR@5% FPR (%)
w/o Jacobian branch	.5517	$-.0156$	.5559	6.70
w/o manifold calibration	.5477	$-.0196$	.5517	**8.77**
w/o temporal localization	.5440	$-.0233$	.5496	6.97
w/o manifold and temporal	.5371	$-.0302$	.5440	7.54
w/o nuisance projection	.5513	$-.0160$	.5563	7.00
w/o target-peer calibration for sparse probes	<u>.5657</u>	$-.0016$	<u>.5658</u>	8.50
Full MCA	**.5673**	.0000	**.5679**	<u>8.65</u>

Component ablation. Table 2 examines each MCA component using seed 7. Removing manifold calibration or temporal localization reduces macro-AUC by 0.0196 and 0.0233, respectively, while removing both causes the largest drop of 0.0302. This confirms that the two mechanisms provide complementary evidence, with temporal localization making the larger individual contribution. Removing Jacobian attribution or nuisance projection also produces substantial drops of 0.0156 and 0.0160, whereas removing target-peer calibration for sparse probes causes only a small drop of 0.0016. Removing manifold calibration slightly improves TPR at 5% FPR despite reducing AUC, indicating a ranking–tail trade-off.

Table 3: Reference-mismatch robustness of MCA on M10. Each reference set contains 480 samples disjoint from the candidate pool. Δ_{match} is relative to the matched-reference result for the same target corruption.

Reference type	Configurations	MCA AUC	Mean Δ_{match}	Max. $\lvert\Delta_{\mathrm{match}}\rvert$
Matched	3	$.6204 \pm .0076$	$+.0000 \pm .0000$	.0000
Cross-corruption	6	$.6175 \pm .0136$	$-.0030 \pm .0169$	.0285
Mixed unseen	3	$.6144 \pm .0222$	$-.0060 \pm .0272$	.0375
Clean reference	3	$.6188 \pm .0134$	$-.0016 \pm .0159$	.0194

Sensitivity to TTA and attack parameters. Appendix Table 4 examines the utility-aware hyperparameter sensitivity of MCA on M10 using seed 7. Increasing the TTA learning rate from 10^{-4} to 10^{-3} raises MCA AUC from 0.5242 to 0.5766 and post-TTA target accuracy from 0.6152 to



(a) Utility after TTA. (b) MCA ROC-AUC. (c) MCA TPR at 5% FPR.

Figure 2: Utility and privacy performance of DP-TENT on M10. Results are reported as mean ± sample standard deviation over three random seeds.

0.6309. Thus, stronger adaptation improves utility while also increasing membership distinguishability in this setting. Increasing the batch size from 16 to 64 reduces MCA AUC from 0.5838 to 0.5422, indicating weaker membership distinguishability with larger batches in this setting. By contrast, varying the manifold size from $k = 10$ to $k = 50$ produces AUC values between 0.5728 and 0.5806, suggesting limited AUC sensitivity to the neighborhood size over the tested range.

Reference mismatch. On M10, MCA remains effective without a corruption-matched reference set. Table 3 reports mean AUCs of 0.6204 with matched references, 0.6175 with cross-corruption references, 0.6144 with unseen mixtures, and 0.6188 with clean references. The corresponding average penalties are only 0.0030, 0.0060, and 0.0016, and some mismatched configurations even improve performance. These results are consistent with the reference set serving as a local estimator of shared manifold motion, rather than requiring it to replicate the target stream exactly.

When is leakage strongest? Membership leakage is often more pronounced in the early stages of adaptation. As shown in Appendix Table 7, the early-half AUC exceeds the late-half AUC in 10 of the 13 settings, with macro-averages of 0.5415 and 0.5262, respectively. This pattern is consistent with the source model being most misaligned with the target distribution at deployment: the resulting early updates can induce larger functional changes, making individual-sample effects easier to distinguish. The trend is not universal, however; M2 and M6 show higher late-half AUC, motivating MCA's candidate-specific temporal localization instead of a fixed early attack window.

5.4 Differential Privacy Defense

Following the DP-TENT formulation of Li et al. Li et al. (2026), we apply per-sample gradient clipping and calibrated Gaussian noise to entropy-minimization updates. Figure 2 compares the privacy–utility performance on M10 under the original non-private setting, clipping alone, and DP-TENT. Tightening ϵ from 20 to 1 reduces MCA AUC from 0.5393 to 0.5104, compared with 0.5935 in the non-private setting. At $\epsilon = 10$, MCA AUC decreases to 0.5346 and TPR at 5% FPR falls from 12.37% to 4.69%. Meanwhile, accuracy decreases from 66.60% under non-private adaptation to 64.06% under DP-TENT. These results demonstrate that DP-TENT effectively mitigates membership leakage while incurring a slight utility trade-off.

6 Conclusion

This work investigates membership leakage in sequential test-time adaptation (TTA), where unlabeled deployment samples can leave candidate-specific temporal traces in the evolving model. We introduce MCA, a label-free attack that uses local manifold calibration, temporal localization, and Jacobian-subspace attribution to separate these traces from shared adaptation dynamics. Across 13 diverse TTA settings, MCA achieves the best average ROC-AUC and low-FPR detection among the evaluated attacks, while remaining robust to reference mismatch. Our analysis further shows that leakage varies across TTA methods and is often stronger during early adaptation, highlighting the importance of adaptation-specific privacy assessment. Differentially private adaptation substantially reduces leakage with slight utility loss. These results highlight the importance of privacy protection in deploying TTA systems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing parts of the manuscript, improving writing clarity and organization, and formatting LaTeX tables. After an initial manual literature survey, AI tools were also used to suggest additional potentially relevant works, which were subsequently checked and further investigated by the authors. We further used AI assistance for the presentation and LaTeX formatting of mathematical expressions, as well as for developing basic code modules and routine implementation components. All AI-assisted writing, mathematical content, references, and code were manually reviewed and revised by the authors, with the manuscript additionally checked by a third-party proofreader and all code verified before use in the experiments.

REFERENCES

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)*, pp. 1897–1914. IEEE, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7704–7714. IEEE, 2023.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Matthew Jagielski, Stanley Wu, Alina Oprea, Jonathan Ullman, and Roxana Geambasu. How to combine membership-inference attacks on multiple updated machine learning models. *Proceedings on Privacy Enhancing Technologies*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Marcus Lassila, Johan Östman, Khac-Hoang Ngo, and Alexandre Graell i Amat. Practical bayes-optimal membership inference attacks. *Advances in Neural Information Processing Systems*, 38: 34278–34312, 2026.

Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. Seqmia: Sequential-metric based membership inference attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3496–3510, 2024.

Zefeng Li, Qiaoyue Tang, Mathias Lécuyer, and Evan Shelhamer. Private and stable test-time adaptation with differential privacy. *arXiv preprint arXiv:2606.01908*, 2026.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 2085–2098, 2022.

Thomas Michel, Debabrota Basu, and Emilie Kaufmann. Sequential membership inference attacks. *arXiv preprint arXiv:2602.16596*, 2026.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. *arXiv preprint arXiv:2404.01650*, 2024.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in neural information processing systems*, 33: 19920–19930, 2020.

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*, pp. 5558–5567. PMLR, 2019.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1291–1308, 2020.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*, 2021.

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)*, pp. 268–282. IEEE, 2018.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15922–15932. IEEE, 2023.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

A Appendix

This appendix provides additional theoretical analysis and experimental results supporting the main paper.

A.1 Justification of Local Manifold Calibration

We provide a justification for the local subtraction in Eq. (8). Recall from Eq. (6) that

$$\rho_t(x) = \mu_t(h_t(x)) + m_t(x)\delta_t(x) + \epsilon_t(x), \tag{21}$$

where μ_t denotes the response shared along the local data manifold and $\epsilon_t(x)$ is zero-mean residual noise.

Proposition. Assume that μ_t is L-Lipschitz and that every $u \in \mathcal{N}^R_{k,t}(x)$ satisfies

$$\|h_t(u) - h_t(x)\|_2 \le \epsilon_x. \tag{22}$$

Since the reference set is disjoint from the adaptation stream, $m_t(u) = 0$ for all $u \in \mathcal{N}^R_{k,t}(x)$. Then

$$\left| \mathbb{E}\left[\frac{1}{k} \sum_{u \in \mathcal{N}^R_{k,t}(x)} \rho_t(u) \right] - \mu_t(h_t(x)) \right| \le L\epsilon_x. \tag{23}$$

Consequently, the locally calibrated response in Eq. (8) satisfies

$$|\mathbb{E}[r_t(x) \mid m_t(x)] - m_t(x)\delta_t(x)| \le L\epsilon_x. \tag{24}$$

Proof. Because $m_t(u) = 0$ and $\mathbb{E}[\epsilon_t(u)] = 0$ for each reference neighbor,

$$\mathbb{E}\left[\frac{1}{k} \sum_{u \in \mathcal{N}^R_{k,t}(x)} \rho_t(u) \right] = \frac{1}{k} \sum_{u \in \mathcal{N}^R_{k,t}(x)} \mu_t(h_t(u)). \tag{25}$$

By the L-Lipschitz continuity of μ_t,

$$\left| \frac{1}{k} \sum_{u \in \mathcal{N}^R_{k,t}(x)} \mu_t(h_t(u)) - \mu_t(h_t(x)) \right| \le \frac{1}{k} \sum_{u \in \mathcal{N}^R_{k,t}(x)} L\|h_t(u) - h_t(x)\|_2 \le L\epsilon_x. \tag{26}$$

Substituting Eq. equation 21 into Eq. (8) and using $\mathbb{E}[\epsilon_t(x)] = 0$ gives Eq. equation 24. Thus, local averaging removes the manifold-shared response up to a bias controlled by the neighborhood radius, while preserving the candidate-specific term $m_t(x)\delta_t(x)$. □

A.2 Additional Experimental Results

This subsection provides the complete configurations underlying the robustness results summarized in Sec. 5. It reports sensitivity to adaptation and attack hyperparameters, attack-wise results under differential privacy, all target–reference pairs, and per-case temporal diagnostics.

A.2.1 Sensitivity to Adaptation and Manifold Parameters

Table 4 examines the effects of learning rate, batch size, and manifold size on M10. MCA is most sensitive to the adaptation configuration: reducing the learning rate lowers AUC from 0.5766 to 0.5242, while batch size 16 produces the strongest attack performance. In contrast, varying the manifold size yields relatively stable AUC values between 0.5728 and 0.5806. Utility remains between 0.6152 and 0.6309 across all configurations.

A.2.2 Attack-wise Results under Differential Privacy

Table 5 complements the privacy–utility sweep in Figure 2. At $\epsilon = 10$, DP strongly reduces LiRA and MCA, while simpler response statistics remain close to chance.

Table 4: Utility-aware sensitivity analysis of MCA on M10 (ImageNet-C/ViT-B/16/TENT-LN) over seed 7. The default configuration uses learning rate 10^{-3}, batch size 32, corruption severity 3, manifold size $k = 30$. Panel (a) varies the TTA learning rate and batch size. Panel (b) varies the manifold size. Utility denotes target accuracy after TTA, and ΔU denotes its change relative to pre-TTA accuracy. Best and second-best attack results in each panel are shown in **bold** and underlined, respectively.

(a) TTA learning rate and batch size

Factor	Setting	Batch	TTA LR	Utility before	Utility after	ΔU	MCA AUC	Attack acc.	TPR5% FPR (%)
Default	–	32	1.0×10^{-3}	.6191	.6309	$+.0117$	.5766	.5645	13.67
TTA learning rate	5.0×10^{-4}	32	5.0×10^{-4}	.6191	.6230	$+.0039$	.5700	.5625	14.42
TTA learning rate	2.5×10^{-4}	32	2.5×10^{-4}	.6191	.6191	$+.0000$	.5588	.5625	14.81
TTA learning rate	1.0×10^{-4}	32	1.0×10^{-4}	.6191	.6152	$-.0039$	.5242	.5332	12.08
Batch size	16	16	1.0×10^{-3}	.6191	.6270	$+.0078$	**.5838**	**.5762**	**15.23**
Batch size	64	64	1.0×10^{-3}	.6191	.6250	$+.0059$	.5422	.5488	10.91

(b) Manifold size

Factor	Setting	k	Utility before	Utility after	ΔU	MCA AUC	Attack acc.	TPR5% FPR (%)
Default	–	30	.6191	.6309	$+.0117$	.5766	.5645	13.67
Manifold size	$k = 10$	10	.6191	.6309	$+.0117$	.5728	**.5742**	14.59
Manifold size	$k = 50$	50	.6191	.6309	$+.0117$	**.5806**	.5645	**15.38**

Table 5: Attack comparison under DP-TENT at $\epsilon = 10$ on M10 over three random seeds (7, 17, and 27). Results are reported as mean $\pm$ sample standard deviation.

Attack	Non-private AUC	DP AUC	Δ AUC	Non-private TPR@5% FPR (%)	DP TPR@5% FPR (%)
Final entropy	$.5415 \pm .0219$	$.5078 \pm .0258$	$-.0336 \pm .0140$	3.52 ± 1.35	3.78 ± 1.37
Entropy drop	$.5841 \pm .0078$	$.5311 \pm .0193$	$-.0529 \pm .0268$	9.38 ± 1.56	4.17 ± 1.26
Feature drift	$.5065 \pm .0302$	$.5133 \pm .0285$	$+.0069 \pm .0276$	$3.52 \pm .39$	$\mathbf{7.16 \pm .23}$
Trajectory	$.5588 \pm .0274$	$.5251 \pm .0283$	$-.0337 \pm .0298$	8.20 ± 1.03	3.65 ± 1.97
LiRA	$\mathbf{.6574 \pm .0252}$	$.5002 \pm .0410$	$-.1572 \pm .0544$	$\mathbf{21.48 \pm 2.44}$	6.12 ± 2.66
RMIA-Online	$.5072 \pm .0279$	$.5000 \pm .0000$	$-.0072 \pm .0279$	3.39 ± 2.48	0.00 ± 0.00
RMIA-Offline	$.5116 \pm .0089$	$.5000 \pm .0000$	$-.0116 \pm .0089$	5.34 ± 2.54	0.00 ± 0.00
SeqMIA	$.5240 \pm .0425$	$.4996 \pm .0087$	$-.0243 \pm .0451$	4.69 ± 3.41	$6.25 \pm .78$
BASE-Online	$.5031 \pm .0054$	$.5141 \pm .0153$	$+.0110 \pm .0204$	2.99 ± 1.26	4.56 ± 1.26
BASE-Offline	$.5246 \pm .0216$	$.5080 \pm .0486$	$-.0166 \pm .0294$	13.15 ± 4.66	7.03 ± 3.20
SeMI (black-box)	$.4858 \pm .0072$	$.4910 \pm .0091$	$+.0053 \pm .0063$	5.08 ± 2.44	3.52 ± 1.41
SeMI (white-box)	$.6346 \pm .0221$	$\mathbf{.5458 \pm .0316}$	$-.0888 \pm .0240$	$8.20 \pm .39$	$6.51 \pm .81$
MCA (ours)	$.5935 \pm .0198$	$.5346 \pm .0152$	$-.0588 \pm .0267$	12.37 ± 3.32	4.69 ± 1.70

Notes. For each seed s, we compute $\Delta_s = \mathrm{AUC}_{\mathrm{DP},s} - \mathrm{AUC}_{\text{non-private},s}$ and report the mean and sample standard deviation of Δ_s across the three seeds. Negative values indicate reduced attack AUC under DP. All statistics are computed from unrounded per-seed results. Bold and underlined entries denote the highest and second-highest mean attack performance, respectively, within each AUC or TPR column; ties receive the same formatting. The Δ AUC column is not ranked.

A.2.3 Complete Reference-Mismatch Results

Table 6 expands the summary in Table 3. MCA remains relatively robust to reference mismatch, with AUC changes ranging from -0.0375 to $+0.0148$ relative to matched references. Mismatched references moderately reduce AUC for Gaussian-noise targets, whereas they achieve comparable or slightly higher AUC for motion-blur and JPEG-compression targets. These results show that MCA does not require an exactly matched reference corruption, although the effect of mismatch remains target-dependent.

A.2.4 Early- and Late-window Diagnostics

Table 7 reports every case. Leakage is stronger on average in the first half of adaptation, but the direction is not universal: M2 and M6 have higher late-window AUC, and several cases have higher late-window TPR. This heterogeneity supports candidate-specific temporal localization.

Table 6: All target–reference configurations for the reference-mismatch study on M10. Δ_{match} is the difference in MCA AUC relative to the matched reference for the same target corruption.

Target corruption	Reference data	Reference type	MCA AUC	Δ_{match}	TPR5% FPR (%)
Gaussian noise	Gaussian noise	Matched	.6269	$+.0000$	15.63
Gaussian noise	Motion blur	Cross-corruption	.5984	$-.0285$	14.06
Gaussian noise	JPEG compression	Cross-corruption	.6080	$-.0189$	12.89
Gaussian noise	Motion blur + JPEG	Mixed unseen	.5894	$-.0375$	12.89
Gaussian noise	Clean images	Clean	.6075	$-.0194$	16.02
Motion blur	Motion blur	Matched	.6224	$+.0000$	23.05
Motion blur	Gaussian noise	Cross-corruption	.6372	$+.0148$	23.05
Motion blur	JPEG compression	Cross-corruption	.6250	$+.0026$	23.83
Motion blur	Gaussian noise + JPEG	Mixed unseen	.6320	$+.0096$	22.27
Motion blur	Clean images	Clean	.6336	$+.0112$	22.66
JPEG compression	JPEG compression	Matched	.6120	$+.0000$	21.48
JPEG compression	Gaussian noise	Cross-corruption	.6152	$+.0032$	23.44
JPEG compression	Motion blur	Cross-corruption	.6210	$+.0090$	21.88
JPEG compression	Gaussian noise + Motion blur	Mixed unseen	.6218	$+.0098$	21.88
JPEG compression	Clean images	Clean	.6154	$+.0034$	19.53

Notes. "Mixed unseen" uses an equal mixture of the other two corruption families, excluding the target corruption. Positive Δ_{match} indicates that the mismatched reference achieves a higher MCA AUC than the matched reference for the same target corruption.

Table 7: Early- versus late-window diagnostics for MCA across all 13 TTA cases over three random seeds. Early and late denote the first and second halves of each adaptation trajectory, respectively. For each candidate, the selected update steps are partitioned between the two windows without additional step selection. Both windows are evaluated on the same member/non-member candidate pool, using only the evidence assigned to the respective window; candidates are not grouped by their actual occurrence times. The resulting window-specific AUCs need not average to the full-trajectory AUC and may both be lower. Δ AUC is the paired late-minus-early difference. "Average" denotes the unweighted macro-average over the 13 cases. Results are reported as mean $\pm$ standard deviation.

	ROC-AUC			Attack accuracy		TPR@5% FPR (%)	
Case	Early	Late	Δ	Early	Late	Early	Late
M1	$\mathbf{.5058 \pm .0460}$	$.5013 \pm .0333$	$-.0045 \pm .0130$	$\mathbf{.5339 \pm .0157}$	$.5273 \pm .0098$	$\mathbf{5.34 \pm 1.93}$	4.95 ± 2.15
M2	$.5184 \pm .0510$	$\mathbf{.5220 \pm .0464}$	$+.0036 \pm .0946$	$.5384 \pm .0262$	$\mathbf{.5443 \pm .0198}$	$\mathbf{7.68 \pm 1.48}$	7.16 ± 2.96
M3	$\mathbf{.5208 \pm .0095}$	$.4883 \pm .0167$	$-.0325 \pm .0260$	$\mathbf{.5378 \pm .0081}$	$.5189 \pm .0098$	5.08 ± 1.41	$\mathbf{5.73 \pm 1.48}$
M4	$\mathbf{.5000 \pm .0000}$	$\mathbf{.5000 \pm .0000}$	$+.0000 \pm .0000$	$\mathbf{.5000 \pm .0000}$	$\mathbf{.5000 \pm .0000}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$
M5	$\mathbf{.5984 \pm .0218}$	$.5510 \pm .0329$	$-.0474 \pm .0377$	$\mathbf{.5846 \pm .0214}$	$.5625 \pm .0186$	$\mathbf{15.23 \pm 1.03}$	11.07 ± 1.19
M6	$.5149 \pm .0249$	$\mathbf{.5264 \pm .0162}$	$+.0115 \pm .0173$	$.5312 \pm .0122$	$\mathbf{.5462 \pm .0108}$	3.91 ± 2.07	$\mathbf{6.90 \pm 2.15}$
M7	$\mathbf{.5087 \pm .0134}$	$.5086 \pm .0365$	$-.0001 \pm .0415$	$.5254 \pm .0052$	$\mathbf{.5312 \pm .0186}$	5.21 ± 1.93	$\mathbf{5.99 \pm 2.15}$
M8	$\mathbf{.6172 \pm .0283}$	$.5788 \pm .0210$	$-.0383 \pm .0226$	$\mathbf{.6061 \pm .0215}$	$.5846 \pm .0164$	$\mathbf{19.27 \pm 0.23}$	13.41 ± 4.73
M9	$\mathbf{.5201 \pm .0197}$	$.5082 \pm .0236$	$-.0119 \pm .0058$	$\mathbf{.5345 \pm .0133}$	$.5319 \pm .0130$	$\mathbf{7.42 \pm 3.20}$	4.69 ± 0.39
M10	$\mathbf{.5528 \pm .0039}$	$.5316 \pm .0173$	$-.0211 \pm .0184$	$\mathbf{.5514 \pm .0030}$	$.5501 \pm .0147$	9.24 ± 2.93	$\mathbf{11.33 \pm 1.79}$
M11	$\mathbf{.5142 \pm .0071}$	$.4978 \pm .0091$	$-.0164 \pm .0120$	$\mathbf{.5345 \pm .0030}$	$.5299 \pm .0049$	4.56 ± 1.97	$\mathbf{7.94 \pm 3.16}$
M12	$\mathbf{.5849 \pm .0291}$	$.5614 \pm .0360$	$-.0235 \pm .0651$	$\mathbf{.5866 \pm .0249}$	$.5853 \pm .0126$	5.08 ± 1.79	$\mathbf{7.42 \pm 3.85}$
M13	$\mathbf{.5827 \pm .0104}$	$.5651 \pm .0365$	$-.0176 \pm .0469$	$.5781 \pm .0153$	$\mathbf{.5801 \pm .0147}$	$\mathbf{18.49 \pm 4.88}$	14.19 ± 2.66
Average	$\mathbf{.5415 \pm .0072}$	$.5262 \pm .0054$	$-.0153 \pm .0090$	$\mathbf{.5494 \pm .0036}$	$.5456 \pm .0024$	$\mathbf{8.19 \pm 0.08}$	7.75 ± 0.33