M-DRIFT: MEMBERSHIP INFERENCE VIA GRADIENT-INDUCED MECHANISTIC DRIFT

Anonymous authors
Paper under double-blind review

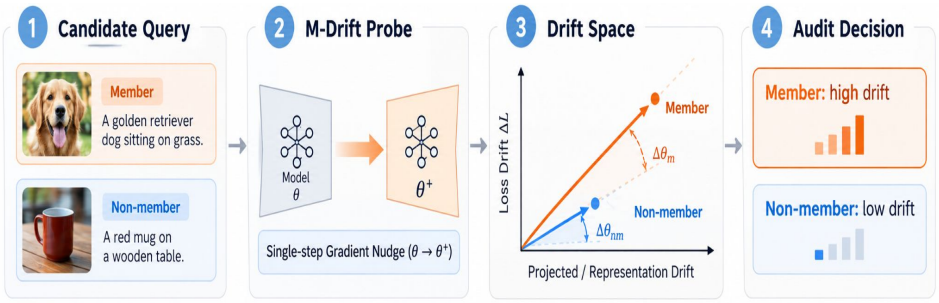


Figure 1: *M-Drift* applies a single-step gradient-ascent nudge ($\theta \rightarrow \theta^+$) and measures **mechanistic drift** output, representation, and cross-modal binding (attention shift), projecting activations into sparse features, capturing the intuition that member instances (present during training) will have more significant drifts than non-members.

ABSTRACT

Vision-language models (VLMs) and diffusion models are often trained on large web-scale datasets, making it difficult to determine whether a particular example was used during training. We introduce **M-Drift**, a white-box membership inference framework that uses a model's response to a small, reversible gradient-ascent perturbation as membership evidence. For each candidate example, M-Drift applies a single loss-increasing nudge and measures the resulting changes in loss, internal representations, and cross-modal interactions. For VLMs, we capture representation and cross-attention drift; for diffusion models, we aggregate denoising drift across sampled timesteps and noise realizations. We further use sparse feature projections to localize the internal directions most affected by the intervention. Across multiple VLM and diffusion backbones, M-Drift consistently improves membership detection in low-false-positive regimes over strong output- and gradient-based baselines. These results suggest that *intervention-induced representation drift* provides a useful and interpretable signal for auditing memorization in modern generative models.

1 INTRODUCTION

Vision-language models (VLMs) and diffusion-based image generators now underpin search, multimodal assistants, and large-scale content creation. Their performance is enabled by training on massive web-scale corpora of image–text pairs and scraped image collections (e.g., CLIP-style pretraining and LAION-scale datasets), which may inadvertently include copyrighted or sensitive content Radford et al. (2021); Schuhmann et al. (2022). This creates an urgent need for *auditing* tools that can answer practical provenance questions: *Was a particular dataset used to train the model? Was a specific image (or image–text pair) seen during training?* The need is no longer hypothetical; modern generative models can memorize and, in some regimes, regurgitate training examples, making privacy and Intellectual Property (IP) risks operational Carlini et al. (2022).

Membership inference attacks (MIAs) provide a principled audit primitive: given a trained model f_θ and a candidate sample x, they infer whether the sample was used in training Shokri et al. (2017); Yeom et al. (2018); aggregating such per-example decisions across many queries then enables estimating dataset-level training membership and potential contamination, which is the goal of this paper. "First-principles" analyses further emphasize that meaningful auditing must operate at *stringent false-positive rates* (FPR), reporting metrics such as true-positive rates $\mathrm{TPR@}\alpha\%\mathrm{FPR}$ (for a small α) and interpreting strong attacks as calibrated hypothesis tests rather than simple loss thresholds Carlini et al. (2022). In practice, auditors face a spectrum of access regimes, *black-box* (outputs

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.

only), *gray-box* (limited internals such as logits/embeddings), and *white-box* (weights, activations, gradients), each exposing different membership *signals* and constraints Shokri et al. (2017); Nasr et al. (2018); Sablayrolles et al. (2019). For deployed Large Vision-Language Models (LVLMs) Xu et al. (2024), restricted access has motivated label-only and black-box methods (e.g., LOMIA Yihao et al. (2025), KCMP Yin et al. (2025)), while open-weight models make cooperative white-box auditing increasingly realistic and diagnostically richer Zhu et al. (2025).

Despite rapid progress, three gaps limit current practice. First, MIAs for VLMs and diffusion models are often developed in isolation, even though both are trained on web-scale data and share architectural building blocks such as transformers and cross-attention Radford et al. (2021); Li et al. (2023); Ho et al. (2020). Second, many strong diffusion MIAs target specific query interfaces (e.g., variation APIs or score queries) and provide limited *mechanistic* insight into *where* membership is encoded inside the model Li et al. (2024a); Rao et al. (2025); Pang & Wang (2023); Pang et al. (2023). Third, mechanistic auditing is complicated by *superposition*: internal activations often represent multiple overlapping features, making neuron-based analyses brittle and hard to interpret Elhage et al. (2022).

We introduce **M-Drift**, a unified, model-agnostic *white-box mechanistic auditor* for vision-language and diffusion models. M-Drift applies a *single, controlled gradient-ascent nudge* that locally increases the loss of a candidate query and then measures how the model's *internal mechanisms drift*. We collect membership evidence from complementary drift signals: output drift, representation drift, and, for multi-modal systems, *cross-modal binding drift* via shifts in cross-attention distributions.

To improve interpretability under superposition, M-Drift optionally projects activations into Sparse autoencoders (SAE) coordinates, which yields overcomplete, sparsely active features that better disentangle superposed directions Cunningham et al. (2023); Elhage et al. (2022). Building on SAE-based interpretability in CLIP-like ViTs (e.g., PatchSAE) Radford et al. (2021); Lim et al. (2024), we use *SAE-coordinate drift* as a localized mechanistic readout.

We instantiate M-Drift for both model families. For VLMs, membership is defined under either generative supervision (teacher-forced token Negative Log-Likelihood (NLL)) or contrastive alignment in the CLIP paradigm Radford et al. (2021); Li et al. (2023). For diffusion models, we use the standard noise-prediction objective defined in equation 11, and stabilize evidence by aggregating drift statistics over Monte-Carlo samples of (t (timestep), ϵ (noise)) to handle timestep-dependent stochasticity Ho et al. (2020). Empirically, we evaluate M-Drift across multiple open-source VLMs and diffusion backbones under matched audit settings, comparing against state-of-the-art methods, and emphasize the *mechanistic interpretability*, enabling actionable auditing insight rather than a binary verdict alone Cunningham et al. (2023); Li et al. (2024a; 2023).

Our **key contributions** are: **(i) Gradient-induced mechanistic auditing for multimodal models.** We extend intervention-induced feature-drift auditing to VLMs and diffusion models through cross-modal binding drift, diffusion Monte-Carlo aggregation, and sparse feature localization. **(ii) A Unified method for VLMs and diffusion.** A model-agnostic pipeline (nudge + multi-signal drift readout) that works for both vision-language and diffusion models under matched access. **(iii) Sparse, interpretable drift evidence.** Superposition-aware SAE-coordinate drift yields localized, human-auditable membership features across vision and multimodal binding pathways. **(iv) Auditing-ready extensions.** We add cross-modal binding drift and diffusion Monte-Carlo aggregation, and provide dataset-level tests with controlled false positives.

2 Related Work

Membership inference attacks (MIAs). MIAs have been studied under varying access assumptions. Early black-box attacks use shadow models and confidence-based features to infer training membership from prediction behavior Shokri et al. (2017); Hu et al. (2022). Later work formalizes membership advantage and shows privacy leakage can persist even when generalization appears strong Yeom et al. (2018); Saeidian et al. (2021). "First-principles" analyses emphasize evaluation at very low false-positive rates and relate optimal auditing to likelihood-ratio-style tests Carlini et al. (2022). With white-box access, gradients and activations yield stronger attacks but require privileged internals Nasr et al. (2018); Sablayrolles et al. (2019); Wu et al. (2023); Pang et al. (2023). In restricted settings, label-only MIAs exploit stability under perturbations, though signals are typically weaker and query-sensitive Choquette-Choo et al. (2021); He et al. (2025); Liu et al. (2024b).

MIAs for vision-language models. VLMs are often deployed via constrained APIs, motivating restricted-access methods. LOMIA targets label-only membership inference for large VLMs, while

KCMP uses generated-text-only outputs with calibration to separate memorization from generic knowledge Yihao et al. (2025); Yin et al. (2025). While practical, output-only evidence is confounded by prompt/response variability, can saturate at low FPR, and offers limited insight into *where* memorization resides.

MIAs and memorization in diffusion models. Diffusion models can memorize and enable training-data extraction, increasing privacy and IP risks Carlini et al. (2022). Prior MIAs include diffusion-specific reverse-process/posterior signals Duan et al. (2023), black-box variation-API attacks (ReDiffuse) Li et al. (2024a), score/noise statistics (SimA) Rao et al. (2025), analytical indicator based CLiD Zhai et al. (2024), black-box attacks for fine-tuned diffusion models (NDSS'25) Pang & Wang (2023), and gradient feature rich white box methods (PoPETs'25) Pang et al. (2023). Black-box approaches often require non-trivial query budgets and depend on the exposed interface, while white-box signals can be high-dimensional and hard to localize Dubiński et al. (2024).

Mechanistic interpretability and sparse features. Superposition complicates mechanistic analysis because features may occupy overlapping activation subspaces rather than individual neurons Elhage et al. (2022). Sparse autoencoders (SAEs) learn sparse feature dictionaries that better disentangle representations and enable localized causal analyses Cunningham et al. (2023). Recent work applies SAEs to CLIP-like vision transformers for feature steering and patch-level concept localization (PatchSAE) Joseph et al. (2025); Li et al. (2024a), but MIAs rarely leverage such sparse feature spaces for interpretable membership evidence, especially in multimodal binding pathways.

3 PRELIMINARIES

Membership inference. In this problem, we ask whether a candidate sample x (e.g., an image-text pair (I, T)) was used to train a target model f_θ. Classical black-box MIAs rely on shadow models and output statistics, exploiting confidence/loss regularities Shokri et al. (2017); moreover, leakage can persist even without pronounced overfitting Yeom et al. (2018). Given an attack score (member likelihood) $s(x) \in \mathbb{R}$ and threshold τ, the predictor outputs $m{=}1$ iff $s(x) > \tau$, with $\mathrm{TPR}(\tau) = \Pr[s(x) \geq \tau \mid x \in \mathcal{D}_{\mathrm{train}}]$ and $\mathrm{FPR}(\tau) = \Pr[s(x) \geq \tau \mid x \notin \mathcal{D}_{\mathrm{train}}]$. We report stringent operating points via $\mathrm{TPR@}\alpha\%\mathrm{FPR} = \mathrm{TPR}(\tau_\alpha)$ where $\tau_\alpha = \arg\min_\tau\{\mathrm{FPR}(\tau) \leq \alpha/100\}$, and summarize separability by the *membership advantage* $\mathrm{Adv} = \max_\tau \big(\mathrm{TPR}(\tau) - \mathrm{FPR}(\tau)\big)$. Following first-principles views, strong MIAs are best interpreted as calibrated likelihood-ratio tests rather than simple loss thresholds Carlini et al. (2022).

Threat models and membership signals. Attacker access ranges from *black-box* (outputs only) to *white-box* (weights, activations, gradients). Membership can leak through (i) output statistics $\mathcal{L}(\theta; x)$, margins, or entropy; (ii) robustness under perturbations, enabling label-only MIAs Choquette-Choo et al. (2021); (iii) representation geometry (hidden states/embeddings); and (iv) optimization-level signals such as per-sample gradients, often strongest in white-box settings Yeom et al. (2018); Carlini et al. (2022). M-Drift leverages (iii)-(iv) by measuring how internal mechanisms change under a small, controlled perturbation.

Objectives in VLMs and diffusion models. Membership depends on the learning objective. For encoders/classifiers (CNNs/ViTs), membership concerns labeled pairs (x, y) and can manifest in attention-mediated representations Dosovitskiy et al. (2020). For vision-language models (VLMs), membership may correspond to pretraining pairs (I, T) or instruction-tuning episodes, often mediated by cross-attention/bridging modules Radford et al. (2021); Li et al. (2023). For diffusion models, membership refers to images (and optional prompts) used to train a denoiser via noise prediction, with the denoising (noise-matching) loss Ho et al. (2020):

$$\mathcal{L}_{\mathrm{diff}}(\theta; x_0, t, \epsilon, c) = \|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2. \tag{1}$$

Here, x_0 is the clean data sample, t is the diffusion timestep, x_t is the noised version of x_0 at step t, $\epsilon \sim \mathcal{N}(0, I)$ is the injected Gaussian noise, $\epsilon_\theta(\cdot)$ is the model's noise predictor with parameters θ, and c denotes conditioning (e.g., text).

This motivates our optional SAE-based drift readout: if membership is mediated by a small set of binding/style features, drift in sparse coordinates can both improve detection and localize responsible circuits. For further details, see Appendix B.

4 METHODOLOGY

4.1 SETUP AND AUDIT OBJECTIVE

Let f_θ be a target model with parameters θ. A membership query consists of input x and task specific prompt/conditioning c instantiated as: (i) **VLMs**: image-prompt (I, p) with target tokens $y_{1:T}$, or paired image-text (I, T) under a similarity objective; (ii) **Diffusion**: an image x_0 (and optional text condition c) evaluated under a denoising objective across time-steps. The attacker returns a score $s(x, c) \in \mathbb{R}$ and $\Pr[m{=}1 \mid x, c]$, where $m{=}1$ denotes that (x, c) was used in training. We emphasize low-FPR evaluation (e.g., $\mathrm{TPR@1\%FPR}$) following first-principles guidance Carlini et al. (2022).

Dataset-level auditing. We distinguish *per-sample* inference computing a calibrated membership score $s(x, c)$ for each candidate from *dataset-level* auditing, which tests whether a suspect set $D = \{(x_i, c_i)\}_{i=1}^n$ contains training members by aggregating $\{s(x_i, c_i)\}$ into a hypothesis test (e.g., two-sample testing or mean log-odds). Calibration is learned on a *disjoint* split of labeled member/non-member examples drawn from the same source pools but *never overlapping* with the audit candidates, preventing leakage from candidate reuse.

4.2 M-DRIFT: GRADIENT-INDUCED MECHANISTIC DRIFT

M-Drift is *model-agnostic* for differentiable VLMs and diffusion models: it applies a single, controlled parameter nudge and measures how internal mechanisms *drift*. Intuitively, member examples occupy more *anchored* regions of the model's local geometry, so a small loss-increasing step perturbs key circuits (representations and cross-modal bindings) in a more structured, discriminative way than matched non-members, consistent with strong white-box MIA signals in gradients/activations Yeom et al. (2018); Carlini et al. (2022); Pang & Wang (2023). The pseudo-code is provided in Appendix A.

Single-step intervention. M-Drift applies one reversible loss-increasing update to a restricted parameter subset θ_M:

$$\theta_M^+ = \mathcal{U}(\theta_M, \nabla_{\theta_M} \mathcal{L}(\theta; x, c); \eta)\,, \tag{2}$$

Equation 2 defines a generic single-step ascent operator $\mathcal{U}$ and η controls its magnitude. In our VLM implementation, $\mathcal{U}$ is instantiated as one reset AdamW ascent step in a fixed low-rank probe subspace; therefore, M-Drift uses the resulting *post-intervention drift*, rather than the raw gradient magnitude itself, as member evidence. Exact initialization details are given in Appendix E.

Step size η selection and sensitivity. We select η on the disjoint calibration split as the smallest loss-increasing setting satisfying the locality criterion defined in Appendix C, and evaluate its sensitivity in Appendix G.3. Sweeping $\eta \in \{0, \eta_0/3, \eta_0, 3\eta_0\}$ shows that no nudge weakens low-FPR separation, while η_0 gives a stable single-step probe consistent with Eq. 2.

Drift readouts. To observe members respond more strongly to a targeted, local parameter intervention, for each instrumented module $\ell \in \mathcal{S}$ (ViT/U-Net blocks, cross-attention outputs), let $a_\ell(x, c; \theta) \in \mathbb{R}^{d_\ell}$ be the flattened activation and define

$$d_\ell^{\mathrm{raw}}(x, c) = \|\Delta a_\ell(x, c)\|_2 \qquad \text{where} \quad \Delta a_\ell(x, c) = a_\ell(x, c; \theta^+) - a_\ell(x, c; \theta), \tag{3}$$

alongside output-level drift $\Delta\mathcal{L} = \mathcal{L}(\theta^+; x, c) - \mathcal{L}(\theta; x, c)$ and task-specific deltas (token NLL/margins for VLMs; denoising errors for diffusion). $d_\ell^{\mathrm{raw}} = \|\Delta a_\ell\|_2$ captures the magnitude of representation change at layer ℓ induced by the gradient nudge.

4.3 SUPERPOSITION-AWARE DRIFT VIA SPARSE FEATURE BASES

Raw neuron-space drift can be hard to interpret due to superposition Elhage et al. (2022). When mechanistic localization is desired, we project activations into a sparse feature basis using a sparse autoencoder (SAE) trained on unlabeled activations from the target architecture:

$$z_\ell = E_\ell(a_\ell), \quad \hat{a}_\ell = D_\ell(z_\ell), \quad \min_{E_\ell, D_\ell} \mathbb{E}\big[\|a_\ell - \hat{a}_\ell\|_2^2\big] + \lambda\|z_\ell\|_1, \tag{4}$$

yielding overcomplete, sparsely-active coordinates that are often more monosemantic Cunningham et al. (2023). We then measure sparse drift

$$\Delta z_\ell = z_\ell(x, c; \theta^+) - z_\ell(x, c; \theta), \qquad d_\ell^{(1)} = \|\Delta z_\ell\|_1, \qquad d_\ell^{(0)} = \|\mathbf{1}[\Delta z_\ell \neq 0]\|_1, \tag{5}$$

and optionally track top-k drifting coordinates for attribution. For vision transformers, recent work shows SAEs can recover interpretable features and enable patch-level localization Radford et al.

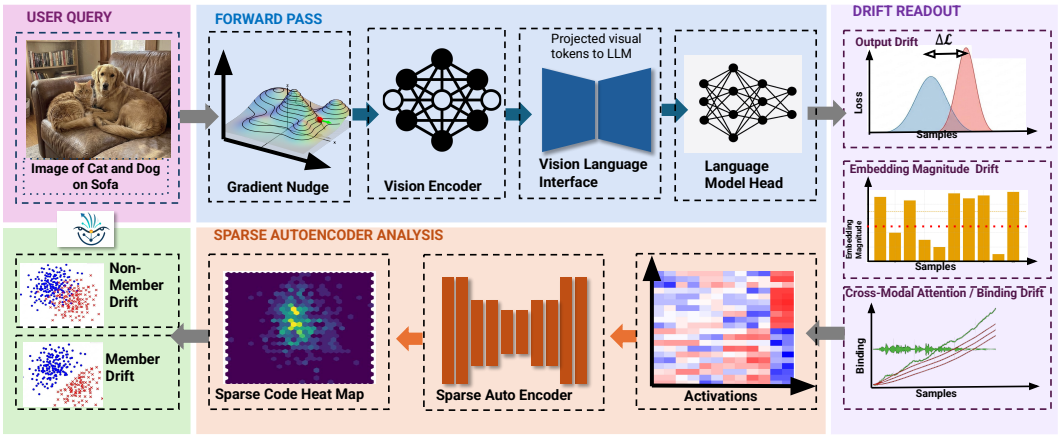


Figure 2: Method overview: M-Drift applies a single-step gradient nudge and measures output drift, representation drift, and cross-modal attention/binding drift; for LLaVA/Qwen, the binding signal is computed from decoder self-attention between text and visual tokens.

(2021); Li et al. (2024a). To handle feature-scale variability, we additionally compute *normalized* and *cosine* SAE drift (e.g., $\|\Delta z\|_2/\|z\|_2$ and $1-\cos(z^+, z)$) and report which statistic best correlates with membership by selecting the top-performing variant on the disjoint calibration split using TPR@1%FPR.

4.4 Cross-Modal Binding Drift

For VLMs and text-conditioned diffusion models, membership can manifest in *binding circuits* (how text queries attend to visual tokens or latent features). Let $A_\ell(\theta) \in \mathbb{R}^{H\times Q\times K}$ be cross-attention weights at layer ℓ (heads $\times$ queries $\times$ keys). We summarize binding drift via divergence statistics:

$$d_\ell^{\text{attn}}(x,c) = \mathrm{KL}\big(\bar{A}_\ell(\theta) \,\|\, \bar{A}_\ell(\theta^+)\big)\,, \tag{6}$$

where $\bar{A}_\ell$ is head-averaged attention. These features support mechanistic claims such as "membership concentrates in specific cross-modal bindings." To reduce sensitivity to attention temperature and LayerNorm scale Sun et al. (2025), we compute binding-drift divergences on *normalized* attention maps, and we report layer-wise drift with a standardized aggregation so that temperature/LN shifts do not spuriously dominate the cross-modal divergence. As indicated in **Figure 2**, given a candidate input, we run a forward pass, apply a small loss-increasing parameter nudge, re-run the model, extract mechanistic drift signals (e.g., ΔL, $\|\Delta a_\ell\|$, cross-attention shift, SAE sparse drift), and aggregate them into a calibrated score to decide member vs. non-member.

4.5 Scoring and Dataset-Level Decision

We concatenate drift signals into a feature vector

$$\phi(x,c) = \big[\Delta\mathcal{L},\ \{d_\ell^{\text{raw}}\}_{\ell\in\mathcal{S}},\ \{d_\ell^{(1)}, d_\ell^{(0)}\}_{\ell\in\mathcal{S}_{\text{SAE}}},\ \{d_\ell^{\text{attn}}\}_{\ell\in\mathcal{S}_{\text{attn}}}\big], \tag{7}$$

and fit a calibrated linear classifier

$$\Pr[m{=}1 \mid x,c] = \sigma(w^\top\phi(x,c)+b), \qquad s(x,c) = w^\top\phi(x,c)+b, \tag{8}$$

which is lightweight, stable, and easy to calibrate for dataset-level audits. Given a suspect dataset D and matched reference R, we accept H_1 if D's scores are systematically larger than R (e.g., Mann-Whitney) or if the mean log-odds $\bar{s}_D = \frac{1}{|D|}\sum_{(x,c)\in D} s(x,c)$ exceeds a threshold controlling dataset-level FPR Carlini et al. (2022); Rainio et al. (2024). We calibrate this threshold on *disjoint* null datasets to enforce a target Type-I error α (using the $(1-\alpha)$-quantile of the null statistic), and report power vs. contamination with bootstrap confidence intervals over dataset resamples.

4.6 Model-Agnostic Instantiations

VLMs. For generative VLMs, we use teacher-forced NLL

$$\mathcal{L}_{\text{VLM}}(\theta; I,p,y_{1:T}) = -\sum_{t=1}^{T}\log P_\theta(y_t \mid I,p,y_{<t}). \tag{9}$$

For contrastive VLMs, we use a similarity surrogate

$$\mathcal{L}_{\text{CLIP}}(\theta; I,T) = -\cos\big(e_I(I;\theta), e_T(T;\theta)\big), \tag{10}$$

or the native contrastive objective when negatives are available Radford et al. (2021). We instrument (i) vision encoder blocks, (ii) language residual streams (if present), and (iii) cross-attention bindings, optionally projecting vision activations into PatchSAE features for localized attribution Lim et al. (2024).

Diffusion models. Given x_0, timestep t, and noise $\epsilon \sim \mathcal{N}(0, I)$, we use the DDPM noise-prediction loss Ho et al. (2020)

$$\mathcal{L}_{\text{diff}}(\theta; x_0, t, \epsilon, c) = \|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2. \tag{11}$$

To handle stochasticity and timestep dependence, we aggregate drift across K trials:

$$\begin{aligned} \phi(x_0, c) &= \text{Agg}_{k=1}^{K}\, \phi(x_0, t_k, \epsilon_k, c), \\ t_k &\sim \mathcal{U}\{1, \ldots, T\},\ \epsilon_k \sim \mathcal{N}(0, I), \end{aligned} \tag{12}$$

and instrument U-Net/DiT activations and cross-attention when text-conditioned. Here $t_k \sim \mathcal{U}\{50, \ldots, 950\}$ for SDv1.5 and the corresponding valid training timestep range for DiT. For details, refer to Appendix C.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on four model/dataset settings with 2000 evaluation instances each ($N_m = 1000$, $N_{nm} = 1000$) drawn from matched distributions; unless stated otherwise, we use balanced splits ($N_m{=}N_{nm}$) and maintain a disjoint calibration set. See Appendix E for details.

Controlled membership ground truth. Because exact pretraining manifests are unavailable for several released checkpoints, we do not treat LAION provenance or timestamps as definitive per-example membership. Instead, for LLaVA, Qwen2-VL, and SDv1.5, we construct a controlled-exposure setting by continuing training from the released checkpoint on a designated member subset $\mathcal{D}_m$, while a source-matched $\mathcal{D}_{nm}$ is excluded from this continuation. Calibration and audit instances are disjoint, and VLM membership is defined for the same image–prompt–target episode used during controlled training. Thus, these experiments measure membership in the controlled continuation stage; DiT/ImageNet retains exact train/validation membership.

Vision-Language Models. We evaluate *LLaVA-1.5-7B* Liu et al. (2024a) and *Qwen2-VL-2B-Instruct* Wang et al. (2024) under controlled membership construction. Starting from each released checkpoint, we explicitly continue training on $\mathcal{D}_m$ using the same image-prompt-target tuples subsequently used for auditing, while matched $\mathcal{D}_{nm}$ examples are withheld from this continuation. We therefore make no assumption that LAION membership of the original checkpoints is known.

Diffusion Models. For SDv1.5 Rombach et al. (2022), we similarly construct controlled membership by continuing the released checkpoint on $\mathcal{D}_m$ using its standard denoising objective, while source-matched $\mathcal{D}_{nm}$ examples are excluded. The resulting labels therefore denote known inclusion in the controlled training stage rather than inferred membership in the undisclosed historical training corpus.

Implementation. (i) Losses and aggregation. For generative LVLMs, we use teacher-forced token NLL Liu et al. (2023); when a contrastive/alignment loss is available, we use its similarity surrogate. For diffusion, we sample $K = 8$ independent (t, ϵ) pairs per candidate; within each trial, the same pair is reused before and after the nudge so that the measured drift is paired, and the resulting K drift vectors are mean-aggregated. **(ii) Calibration, scoring, and runtime.** We score membership using logistic calibration over the concatenated drift vector $\phi(x, c)$, averaging results over 5 seeds (data splits and diffusion randomness). We use mixed-precision inference with batch size 1; synchronized end-to-end runtime, including the gradient and post-nudge forward pass, is reported in Appendix F.1.

Metrics. We report both pointwise membership performance and dataset-level auditing capability.

Pointwise membership metrics. We evaluate (i) ROC-AUC Ling et al. (2003) (mean±std over 5 seeds), (ii) PR-AUC Davis & Goadrich (2006) at prevalence $\pi = \Pr[m{=}1]$ (optionally re-weighted to low-prevalence audit regimes), (iii) TPR@0.1% FPR Fawcett (2006), (iv) TPR@1% FPR Fawcett (2006), and (v) Membership Advantage (MA) Yeom et al. (2018), defined as $\max_\tau \big(\text{TPR}(\tau) - \text{FPR}(\tau)\big)$.

Dataset-level auditing. For VLMs: Dataset-level Contamination Detection (DCD) Samuel et al. (2025) tests whether a suspect dataset D contains training members relative to a clean reference D_0

by aggregating per-example scores $\{s_i\}$ into a dataset statistic (e.g., mean log-odds) and reporting detection strength versus contamination rate. For Diffusion: At a fixed Type-I error α, we estimate detection power over $\rho \in \{0, 1, 5, 10\}\%$ contamination. DCD/DL-PC denotes the normalized area under this power–contamination curve; the exact statistic, resampling protocol, and threshold construction are given in Appendix G.

Low-FPR estimation. Each seed contains 1,000 held-out non-members, giving an empirical FPR resolution of 0.1%. We use TPR@1%FPR as the primary low-FPR metric and report TPR@0.1%FPR as a stricter supporting tail metric; uncertainty and seed-wise estimation are detailed in Appendix G.2.

Competing approaches. We benchmark against state-of-the-art MIAs under matched access assumptions and report results by regime (label-only / black-box / gray-box / white-box) to ensure fair comparisons. Detailed in Appendix D.

VLM baselines. LOMIA (label-only) uses generated outputs/top predictions Yihao et al. (2025). **KCMP** (black-box probing) calibrates memory probes to separate memorization from world knowledge Yin et al. (2025). **MaxRényi MIA** (gray-box/logit-aware) leverages multimodal logit/segment-wise signals with limited-logit variants Li et al. (2024b). **EncoderMI** targets CLIP-style contrastive encoders via augmentation consistency Liu et al. (2021). **LiRA** provides a calibrated likelihood-ratio low-FPR reference Carlini et al. (2022).

Diffusion baselines. ReDiffuse (black-box variation API) averages stochastic variations Li et al. (2024a). SimA (score/noise-vector) uses single-/few-query score-space statistics Rao et al. (2025). SecMI uses posterior/reversal error signals tailored to diffusion Duan et al. (2023). LiRA-style calibration uses calibrated reconstruction/noise-loss features Carlini et al. (2022). GSA is a white-box gradient/activation-rich baseline Pang et al. (2023). For multi-query black-box attacks, we additionally report performance versus query budget and include matched-budget comparisons to M-Drift when possible. Temporal-bias controls table, full filtering, deduplication, and matching details are provided in Appendix E.1.

5.2 Main Results

Vision Language Models. Table 1 shows that **M-Drift is most effective at stringent false-positive rates**, with the strongest directly estimated gains at TPR@1%FPR. TPR@0.1%FPR is reported as a supporting seed-aggregated tail metric with uncertainty analysis in Appendix G.2. Consistent with prior findings that calibrated likelihood-ratio methods (e.g., LiRA) excel at low FPR, this indicates that mechanistic drift provides complementary evidence beyond output likelihoods. Even when ROC-AUC is near-saturated (e.g., LLaVA/L-SBU where MaxR is already high), M-Drift further improves TPR@0.1% and MA, suggesting it sharpens the score tails and better ranks the hardest cases. Finally, **dataset-level conclusions track pointwise strength**: M-Drift achieves the highest DCD on both models, while restricted-access methods (LOMIA, KCMP) remain near-chance and yield weak DCD, underscoring that outputs alone often underdetermine membership without calibration or mechanistic signals. Figure 3-(A), emphasizes audit-relevant regimes (very low false positives) and shows that M-Drift attains higher true-positive rates at fixed low FPR, indicating stronger membership detection with fewer false alarms.

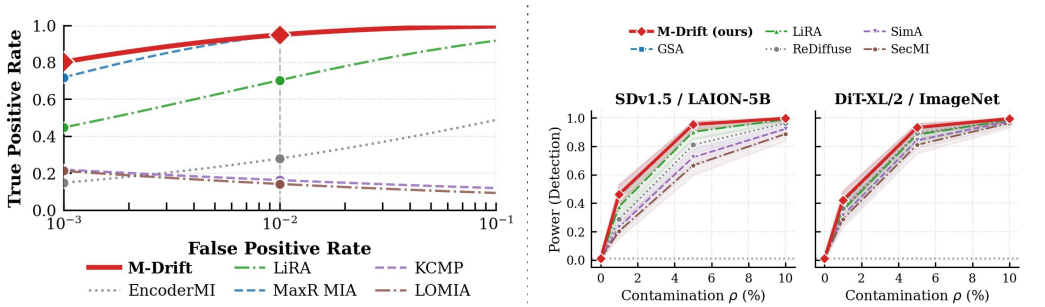
Diffusion Models. Table 2 shows that **M-Drift is strongest at low FPR on diffusion models.** On SDv1.5, it improves ROC/PR and, crucially, boosts TPR@0.1% (reported in Appendix G.2) and TPR@1%, aligning with prior diffusion-MIA evidence that noise/reconstruction signals become far

Table 1: Membership inference on VLMs. Values are mean±std over 5 random seeds for ROC-AUC and PR-AUC only. For the remaining metrics, std is omitted for readability and falls in the ranges: TPR@0.1% (0.3–4.1), TPR@1% (0.8–3.9), MA (0.009–0.030), DCD (0.010–0.020). Higher is better for all metrics. TPR@0.1 and TPR@1 denote true-positive rate at FPR=0.1% and FPR=1%, respectively. MA is the membership advantage. DCD is the dataset-level contamination detection score.

	LLaVA / LAION-SBU						**Qwen / LAION-en/zh**					
Method	ROC-AUC	PR-AUC	TPR@0.1	TPR@1	MA	DCD	ROC-AUC	PR-AUC	TPR@0.1	TPR@1	MA	DCD
EncoderMI	0.78±0.026	0.79±0.024	14.8	27.9	0.42	0.70	0.74±0.030	0.75±0.028	12.1	23.8	0.36	0.66
LiRA	0.91±0.021	0.92±0.018	44.7	70.3	0.76	0.84	0.88±0.025	0.89±0.024	34.6	60.2	0.70	0.79
MaxR MIA	0.98±0.026	0.98±0.018	71.8	94.8	0.94	0.94	0.95±0.020	0.97±0.026	**68.2**	84.1	0.84	0.90
KCMP	0.67±0.020	0.68±0.021	16.2	21.9	0.31	0.63	0.65±0.022	0.66±0.021	15.5	21.6	0.30	0.60
LOMIA	0.66±0.018	0.67±0.017	14.1	21.2	0.28	0.61	0.63±0.018	0.64±0.019	13.8	18.9	0.25	0.58
M-Drift (ours)	**0.98±0.017**	**0.98±0.012**	**80.2**	**94.9**	**0.96**	**0.95**	**0.96±0.010**	**0.97±0.018**	67.8	**90.1**	**0.92**	**0.91**

Table 2: Membership inference on diffusion models. Values are mean±std over 5 seeds for ROC-AUC and PR-AUC only. For the remaining metrics, std is omitted for readability and falls in the ranges: TPR@0.1% (3.0–4.2), TPR@1% (3.0–4.1), MA (0.02–0.03), DL-PC (0.02). MA is the membership advantage. DL-PC is a dataset-level auditing score summarizing Type-I control and power vs. contamination fraction (higher is better).

	SDv1.5 / LAION-5B						DiT / ImageNet					
Method	ROC-AUC	PR-AUC	TPR@0.1	TPR@1	MA	DL-PC	ROC-AUC	PR-AUC	TPR@0.1	TPR@1	MA	DL-PC
SecMI	0.80±0.030	0.79±0.030	18.0	42.0	0.50	0.72	0.84±0.030	0.84±0.030	25.0	50.0	0.58	0.78
LiRA	0.90±0.025	0.90±0.025	45.0	68.0	0.72	0.85	0.89±0.028	0.89±0.028	32.0	57.0	0.63	0.83
GSA	0.95±0.020	0.95±0.022	65.0	86.0	0.86	0.92	0.93±0.022	0.94±0.032	44.0	69.0	0.74	0.88
SimA	0.82±0.030	0.82±0.032	22.0	45.0	0.52	0.74	0.86±0.030	0.86±0.030	28.0	52.0	0.60	0.80
ReDiffuse	0.86±0.020	0.86±0.020	30.0	55.0	0.60	0.78	0.91±0.025	0.91±0.025	36.0	60.0	0.65	0.84
M-Drift	**0.96±0.032**	**0.96±0.018**	**70.0**	**88.0**	**0.90**	**0.93**	**0.94±0.022**	**0.94±0.022**	**48.0**	**72.0**	**0.76**	**0.89**



(A) Low FPR Performance Analysis (B) Dataset-level Auditing Comparison

Figure 3: (A) Low-FPR ROC for LLaVA on LAION-SBU dataset: TPR vs. FPR (log scale), showing empirically estimated operating points at 0.1% and 1% FPR. (B) Power vs. contamination rate ρ at fixed Type-I (α), Comparison on SDv1.5 (LAION-5B) and DiT-XL/2 (ImageNet).

more discriminative when combined with calibration or richer internal statistics than a single loss value. Despite strong white-box baselines (e.g., GSA), M-Drift remains best overall, improving MA and DL-PC, which suggests the interventional drift probe captures local memorization geometry beyond raw gradients/activations. Performance is lower on DiT than SDv1.5, yet M-Drift stays top-ranked, indicating good transfer across diffusion parameterizations. Finally, **dataset-level auditing follows pointwise strength**: higher MA/TPR correspond to higher DL-PC, yielding more reliable dataset-level decisions at controlled Type-I error. Figure 3-(B), shows that under controlled false-alarm rates, M-Drift consistently achieves higher detection power as contamination increases, demonstrating stronger and more reliable dataset-level membership auditing on diffusion models. Detailed results in Appendix G.

5.3 Ablation Study

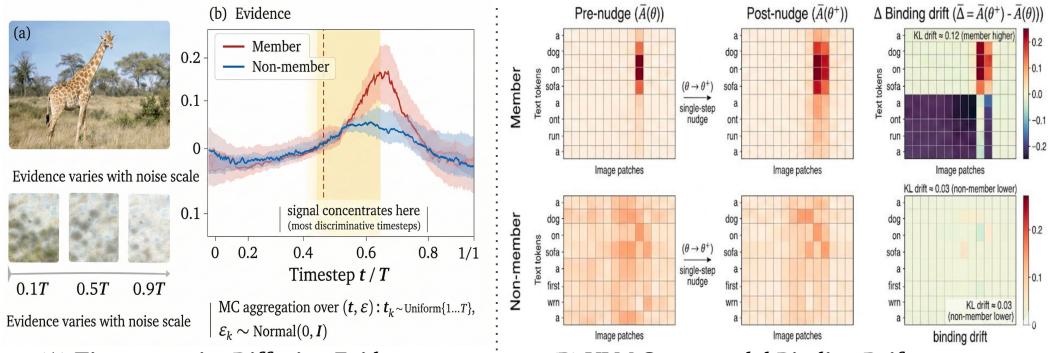
In Table 3, we see that the full configuration performs best because M-Drift combines (i) an interventional signal from the single-step gradient-ascent nudge, which amplifies member/non-member separability in the model's local loss geometry, with (ii) mechanistic drift readouts from internal representations and (for multimodal/text-conditioned settings) cross-attention binding shifts. Using only the loss-drift signal substantially reduces low-FPR performance, showing that the gain primarily arises from combining complementary representation, attention, and sparse-feature drift signals. Finally, modality-only variants underperform the fused model because membership evidence can be distributed across vision, text/conditioning, and their binding circuits; fusion recovers complementary cues that single-modality drift misses. Detailed sensitivity and hyperparameter configurations are elaborated upon in Appendix G.3.

5.4 Qualitative Analysis

Figure 4-(A) part (a) visualizes a candidate image x_0 and its progressively noised versions x_t to illustrate how the diffusion noise scale changes with timestep. Part (b) plots the M-Drift evidence

Table 3: Ablation study for M-Drift on a representative VLM (LLaVA/L-SBU) and diffusion model (SDv1.5/L-5B).

	LLaVA/LAION-SBU			SDv1.5/LAION-5B		
Ablation Variant	ROC-AUC	TPR@1%FPR	DCD	ROC-AUC	TPR@1%FPR	DL-PC
Full M-Drift	0.980	94.9	0.952	0.960	88.0	0.930
Loss-drift only ($\Delta\mathcal{L}$; no feature fusion)	0.910	68.0	0.800	0.880	55.0	0.780
w/o cross-modal fusion (avg of modality)	0.970	91.0	0.930	0.950	84.0	0.910
Vision-only drift signal	0.940	79.0	0.880	0.930	73.0	0.870
Text/conditioning-only drift signal	0.960	85.0	0.900	0.900	62.0	0.820
Output-only (no gradients)	0.700	8.0	0.620	0.650	12.0	0.600



(A) Timestep-wise Diffusion Evidence (B) VLM Cross-modal Binding Drift

Figure 4: (A) Evidence concentration for M-Drift on SDv1.5: membership evidence varies across noise scales and peaks at intermediate timesteps. (B) A single-step gradient nudge (LLaVA on LAION) induces sharper, more localized cross-attention redistribution for member examples than for matched non-members.

(mean $\pm$ MC band over ϵ) as a function of t/T, showing a concentrated region where member and non-member curves separate most, motivating aggregation over sampled (t, ϵ) for robust auditing.

Figure 4-(B) visualizes head-averaged cross-attention maps before and after the M-Drift nudge, along with their difference heatmaps for a member and non-member example. For the member sample, the nudge induces a structured and localized binding shift (higher KL drift), indicating anchored text–image associations learned during training. In contrast, the non-member shows diffuse and low-magnitude attention changes, supporting the claim that membership evidence concentrates in cross-modal binding circuits. Beyond the visual heatmaps, Appendix G.4 quantitatively validates localization using top-k drift concentration, seed stability, and targeted deletion tests. Appendix G.6 analyzes the failure cases.

6 Discussion

Angle-aware drift as a membership cue. As shown in Fig. 1, M-Drift infers membership from *interventional drift* induced by a single-step gradient nudge, using both loss- and embedding-space changes. Here, α and α' denote the loss drift magnitude for a member and non-member, while β and β' denote the embedding magnitude shift after the intervention; empirically, members exhibit larger shifts ($\alpha > \alpha'$, $\beta > \beta'$) even when drift directions are similar. We also measure *directional* change via $\Delta\theta = \arccos(\cos(e^+, e))$, complementing norm-based drift such as $\|\Delta a_\ell\|$.

How defenses may attenuate drift. Training-time defenses that smooth or randomize local geometry should weaken drift separability. DP-SGD (gradient clipping + noise) can suppress per-example influence and decorrelate informative drift directions, while aggressive regularization reduces sharpness/memorization, damping both output drift and representation rotation. Adversarial training may also reduce drift via local invariance.

M-Drift as a red-team evaluator. M-Drift can act as a privacy red-team tool by measuring attack strength before/after a defense at low-FPR operating points, and by localizing residual leakage to specific modules (via θ_M) or sparse SAE features. Effective defenses should reduce both average attack performance and high-confidence leakage cases, while eliminating localized memorization-sensitive circuits. A detailed discussion and **Limitations** are provided in Appendix H.

7 Conclusion

We presented *M-Drift*, a mechanistic membership auditing framework that re-frames membership evidence as *feature drift* induced by a controlled, single-step gradient-ascent intervention . By measuring complementary drift signals, output loss drift, internal representation drift, and (when applicable) cross-modal binding drift, M-Drift provides a unified approach that applies to both vision language models and diffusion generators. To improve interpretability under superposition, we further introduce SAE-coordinate drift, which yields sparse, feature-localized evidence and supports mechanistic attribution of where membership signal concentrates. Across representative VLMs and diffusion models, M-Drift consistently improves low-FPR operating points and strengthens dataset-level auditing (DCD / DL-PC), enabling more reliable decisions about potential training-set contamination under controlled false positives.

AI USE STATEMENT

Generative AI tools were used only for limited language editing, including grammar, readability, formatting, and minor editorial assistance. They were not used to generate data, formulate mathematical claims, design experiments, implement the proposed method, analyze results, or draw scientific conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

M-Drift studies membership inference, which may expose privacy or training-data provenance information and therefore has potential dual-use implications. Our experiments use publicly available datasets and open-source model checkpoints and involve no human-subject studies or newly collected personal data. We present M-Drift primarily as a cooperative auditing and privacy red-teaming tool for identifying potential memorization and evaluating model privacy risks.

REPRODUCIBILITY STATEMENT

We provide the M-Drift algorithm and methodological details in the main paper and Appendix A, implementation and probe configurations in Appendix C, and complete experimental, calibration, and evaluation settings in Appendices E–G. We will additionally release the code required to reproduce the reported experiments, metrics, and figures.

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.

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

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 118–125. IEEE, 2025.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4860–4869, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. Towards {Label-Only} membership inference attack against pre-trained large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1609–1628, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54 (11s):1–37, 2022.

Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering clip's vision transformer with sparse autoencoders. *arXiv preprint arXiv:2504.08729*, 2025.

Jingwei Li, Jing Dong, Tianxing He, and Jingzhao Zhang. Towards black-box membership inference attack for diffusion models. *arXiv preprint arXiv:2405.20771*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems*, 37:98645–98674, 2024b.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *arXiv preprint arXiv:2412.05276*, 2024.

Charles X Ling, Jin Huang, Harry Zhang, et al. Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, volume 3, pp. 519–524, 2003.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. Encodermi: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2081–2095, 2021.

Rongke Liu, Dong Wang, Yizhi Ren, Zhen Wang, Kaitian Guo, Qianqian Qin, and Xiaolei Liu. Unstoppable attack: Label-only model inversion via conditional diffusion model. *IEEE Transactions on Information Forensics and Security*, 19:3958–3973, 2024b.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1–15, 2018.

Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. *arXiv preprint arXiv:2312.08207*, 2023.

Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific reports*, 14(1):6086, 2024.

Mingxing Rao, Bowen Qu, and Daniel Moyer. Score-based membership inference on diffusion models. *arXiv preprint arXiv:2509.25003*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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.

Sara Saeidian, Giulia Cervia, Tobias J Oechtering, and Mikael Skoglund. Quantifying membership privacy via information leakage. *IEEE Transactions on Information Forensics and Security*, 16: 3096–3108, 2021.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5058–5070, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

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.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. *arXiv preprint arXiv:2502.05795*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Di Wu, Saiyu Qi, Yong Qi, Qian Li, Bowen Cai, Qi Guo, and Jingxian Cheng. Understanding and defending against white-box membership inference attack in deep learning. *Knowledge-Based Systems*, 259:110014, 2023.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (3):1877–1893, 2024.

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.

LIU Yihao, LYU Xinqi, et al. Lomia: Label-only membership inference attacks against pre-trained large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jinhua Yin, Peiru Yang, Chen Yang, Huili Wang, Zhiyang Hu, Shangguang Wang, Yongfeng Huang, and Tao Qi. Black-box membership inference attack for lvlms via prior knowledge-calibrated memory probing. *arXiv preprint arXiv:2511.01952*, 2025.

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. *Advances in Neural Information Processing Systems*, 37:74122–74146, 2024.

Hongyu Zhu, Sichu Liang, Wenwen Wang, Boheng Li, Tongxin Yuan, Fangqi Li, Hanyi Wang, Shi-Lin Wang, and Zhuosheng Zhang. Revisiting data auditing in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11337–11346, 2025.

APPENDIX

A ALGORITHM PSEUDOCODE

Algorithm 1: M-Drift: Single-Step Probe, Drift Extraction, and Membership Scoring

Input: Candidate set $D = \{(x_i, c_i)\}_{i=1}^{n}$, model f_θ, probed parameters θ_M, step size η, number of trials K, update operator $\mathcal{U}$, calibrated scorer $g(\cdot)$, dataset-level threshold τ.
Output: Sample membership scores $\{s_i, p_i\}_{i=1}^{n}$ and dataset decision $\hat{y} \in \{0, 1\}$.

foreach $(x_i, c_i) \in D$ **do**
  for $k \leftarrow 1$ **to** K **do**
    Sample nuisance $\xi_{i,k}$; `// e.g., (t, ϵ) for diffusion`
    Compute $L_i^{(k)} = L(\theta; x_i, c_i, \xi_{i,k})$ and internal states $h_i^{(k)}$;
    Compute probe gradient $r_i^{(k)} = \nabla_{\theta_M} L_i^{(k)}$;
    Apply the reversible probe $\theta_M^{+} \leftarrow \mathcal{U}\Big(\theta_M, r_i^{(k)}; \eta\Big)$;
    Recompute $L_i^{+(k)}$ and $h_i^{+(k)}$ under f_{θ^+} using the same $\xi_{i,k}$;
    Extract drift features $\phi_i^{(k)} \leftarrow \Phi\Big(L_i^{(k)}, L_i^{+(k)}, h_i^{(k)}, h_i^{+(k)}\Big)$;
    Restore $\theta_M \leftarrow \theta_M^{\text{orig}}$; `// no probe accumulation`
  Aggregate trial-level evidence: $\phi_i \leftarrow \dfrac{1}{K}\sum_{k=1}^{K} \phi_i^{(k)}$;
  Compute calibrated log-odds $s_i \leftarrow g(\phi_i)$;
  Compute membership probability $p_i \leftarrow \sigma(s_i)$;

Compute dataset statistic $T(D) \leftarrow \dfrac{1}{n}\sum_{i=1}^{n} s_i$;

$\hat{y} \leftarrow \mathbb{I}[T(D) > \tau]$;

Algorithm 2: SAE-Drift Extraction (sparse projection + drift features)

Input: Pre/post activations h, h^+ (layer ℓ), SAE encoder $E(\cdot)$ (optional decoder $D(\cdot)$), normalization $\mathcal{N}(\cdot)$, sparsity level k, small constant ϵ.
Output: SAE-drift descriptor ψ (used as part of $\Phi(\cdot)$ in Alg. 1).

$u \leftarrow \mathcal{N}(h)$; $u^+ \leftarrow \mathcal{N}(h^+)$ `// e.g., per-dim standardization or LayerNorm`
$z \leftarrow E(u)$; $z^+ \leftarrow E(u^+)$ `// codes in` $\mathbb{R}^{d_{\text{SAE}}}$
$\tilde{z} \leftarrow \text{TopK}(z, k)$; $\tilde{z}^+ \leftarrow \text{TopK}(z^+, k)$ `// keep k largest magnitudes, zero others`
$\Delta z \leftarrow \tilde{z}^+ - \tilde{z}$ `// coordinate-wise drift`
$m_1 \leftarrow \|\Delta z\|_1$; $m_2 \leftarrow \|\Delta z\|_2$
$m_{\cos} \leftarrow 1 - \dfrac{\langle \tilde{z}, \tilde{z}^+ \rangle}{\|\tilde{z}\|_2 \, \|\tilde{z}^+\|_2 + \epsilon}$
$S \leftarrow \text{supp}(\tilde{z})$; $S^+ \leftarrow \text{supp}(\tilde{z}^+)$
$m_{\text{J}} \leftarrow 1 - \dfrac{|S \cap S^+|}{|S \cup S^+| + \epsilon}$ `// Jaccard distance of supports`
if *D is available* **then**
  $r \leftarrow \|D(\tilde{z}) - u\|_2$; $r^+ \leftarrow \|D(\tilde{z}^+) - u^+\|_2$ `// optional recon diagnostics`
  $\psi \leftarrow [m_1, m_2, m_{\cos}, m_{\text{J}}, r, r^+]$
else
  $\psi \leftarrow [m_1, m_2, m_{\cos}, m_{\text{J}}]$
return ψ

Algorithm 1 describes our proposed approach, **M-Drift**, which performs a **single-step gradient probe** on a candidate instance set(members and non-members), extracts **pre/post internal drift features**, and outputs calibrated instance-level **membership scores** and a dataset-level decision (aggregated per sample features).

Algorithm 2 describes **SAE-Drift extraction**, which projects **pre/post activations** into **sparse SAE features** and summarizes their **drift (magnitude + support change)** as inputs to the M-Drift scorer.

B Expanded Preliminaries

Existing MIAs for VLMs and diffusion. Under restricted access, MIAs are typically label-only or black-box, relying on proxy and calibration signals. For VLMs, LOMIA exploits label-only proxies such as semantic alignment and perplexity Yihao et al. (2025), while KCMP uses calibrated black-box memory probing to separate memorization from generic capability Yin et al. (2025). For diffusion, black-box attacks leverage stochastic reconstruction/variation APIs (ReDiffuse) Li et al. (2024a) or score-based statistics (SimA) Rao et al. (2025), and NDSS'25 studies such threats for fine-tuned diffusion models Pang & Wang (2023). White-box diffusion MIAs show gradients/activations can be highly discriminative, but are compute-heavy and weakly localizing Pang et al. (2023).

Figure 5 illustrates how raw, entangled activations a_ℓ are mapped via the sparse autoencoder E_ℓ into sparse feature representations $z_\ell = E_\ell(a_\ell)$, enabling superposition-aware, localized drift analysis in M-Drift.

Mechanistic interpretability and sparse features. Mechanistic auditing is complicated by *superposition*, where many features share overlapping activation directions Elhage et al. (2022). Sparse autoencoders (SAEs) learn overcomplete, sparsely active feature dictionaries that yield more interpretable coordinates Cunningham et al. (2023); recent vision work trains SAEs on CLIP-like ViTs for steering and patch-level localization Joseph et al. (2025); Lim et al. (2024).

C Implementation

C.1 Implementation for Vision-Language Models

Inputs, loss, and nudge. A VLM query is (x, c) with $x = (I, p)$ (generative) or $x = (I, T)$ (contrastive), and conditioning/supervision c. For generative VLMs, we use teacher-forced Negative Log-Likelihood (NLL)

$$\mathcal{L}_{\text{VLM}}(\theta; I, p, y_{1:T}) = -\sum_{t=1}^{T} \log P_\theta(y_t \mid I, p, y_{<t}) \,. \tag{13}$$

For CLIP-style retrieval, we use the (mini-batch) contrastive objective when negatives exist; otherwise a similarity surrogate

$$\mathcal{L}_{\text{CLIP}}(\theta; I, T) = -\cos\big(e_I(I;\theta), e_T(T;\theta)\big), \tag{14}$$

with normalized embeddings e_I, e_T. Given $\mathcal{L}(\theta; x, c)$, we apply the single-step intervention (Sec. 4.2)

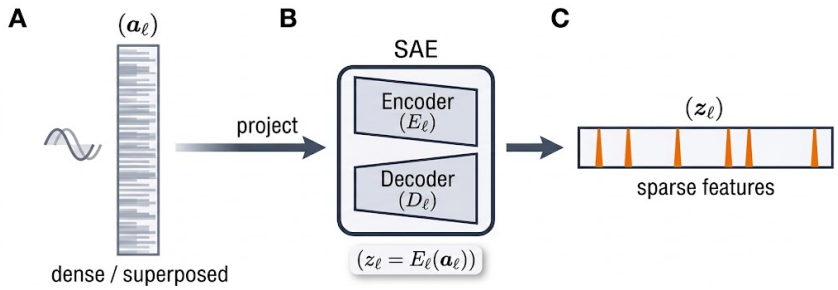


Figure 5: Projecting dense activations into sparse SAE features. Dense, superposed layer activations a_ℓ are projected through a sparse autoencoder (SAE) to obtain over-complete sparse coordinates z_ℓ that disentangle underlying features.

$$\theta_M^+ = \mathcal{U}(\theta_M, \nabla_{\theta_M}\mathcal{L}(\theta; x, c); \eta)\,, \tag{15}$$

For the reported VLM experiments, $\mathcal{U}$ denotes exactly one AdamW ascent step with freshly reset optimizer state; for every candidate, both the model/probe parameters and optimizer state are restored before the next query. Thus, the intervention should be interpreted as a controlled adaptive ascent probe rather than an unnormalized SGD step.

Instrumented modules and signals. We record activations at a small set of layers $\mathcal{S}$ covering: (i) **vision encoder** blocks (ViT patch tokens), (ii) **language** residual stream states at answer positions (if present), and (iii) **cross-modal binding** via cross-attention. For each $\ell \in \mathcal{S}$, let $a_\ell(x, c; \theta)$ be the flattened activation vector; we compute raw drift

$$\Delta a_\ell = a_\ell(\theta^+) - a_\ell(\theta), \qquad d_\ell^{\text{raw}} = \|\Delta a_\ell\|_2. \tag{16}$$

For binding drift, let $A_\ell(\theta) \in \mathbb{R}^{H\times Q\times K}$ be cross-attention weights; we summarize with divergence (Sec. 4.4)

$$d_\ell^{\text{attn}} = \mathrm{KL}\big(\bar{A}_\ell(\theta) \,\|\, \bar{A}_\ell(\theta^+)\big)\,, \tag{17}$$

where $\bar{A}_\ell$ averages heads.

Probe locality. We quantify intervention locality by the affected-parameter fraction $r_M = |\theta_M|/|\theta|$ and the relative parameter displacement

$$\rho = \frac{\|\theta^+ - \theta\|_2}{\|\theta\|_2 + \epsilon}. \tag{18}$$

We report the mean and maximum ρ over audit queries together with the exact number and percentage of probed parameters. Non-probed parameters remain unchanged.

Superposition-aware drift (optional). When enabled, we map a_ℓ into SAE coordinates (Sec. 4.3): $z_\ell = E_\ell(a_\ell)$ and compute

$$\begin{aligned}\Delta z_\ell = z_\ell(\theta^+) - z_\ell(\theta), \qquad & d_\ell^{(1)} = \|\Delta z_\ell\|_1, \\ & d_\ell^{(0)} = \|\mathbf{1}[\Delta z_\ell \neq 0]\|_1.\end{aligned} \tag{19}$$

We optionally log top-k drifting SAE coordinates for mechanistic reporting (e.g., patch-wise overlays using PatchSAE-style token features).

Feature vector and score. We form $\phi(x, c)$ by concatenating output drift and internal drift statistics (Sec. 4.5),

$$\begin{aligned}\phi(x, c) = \big[\Delta\mathcal{L}, \{d_\ell^{\text{raw}}\}_{\ell\in\mathcal{S}}, \{d_\ell^{\text{attn}}\}_{\ell\in\mathcal{S}_{\text{attn}}}, \\ \{d_\ell^{(1)}, d_\ell^{(0)}\}_{\ell\in\mathcal{S}_{\text{SAE}}}\big],\end{aligned} \tag{20}$$

and produce the membership score $s(x, c) = w^\top\phi(x, c) + b$ (logistic calibration in Sec. 4.5). All parameters are restored after each query.

C.2 Implementation for Diffusion Models

Noising process and loss. Given an image x_0 (and optional text condition c), we sample a timestep $t \in \{1, \ldots, T\}$ and noise $\epsilon \sim \mathcal{N}(0, I)$, form x_t, and evaluate the denoiser $\epsilon_\theta(x_t, t, c)$ with the standard DDPM objective

$$\mathcal{L}_{\text{diff}}(\theta; x_0, t, \epsilon, c) = \|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2. \tag{21}$$

We then apply the same single-step nudge

$$\theta^+ = \theta + \eta\nabla_{\theta_M}\mathcal{L}_{\text{diff}}(\theta; x_0, t, \epsilon, c), \tag{22}$$

with θ_M optionally restricted (e.g., mid-blocks, attention blocks, or LoRA adapters in text-to-image models) to balance compute and localization.

Monte-Carlo aggregation over timesteps. Because diffusion dynamics are stochastic and timestep-dependent, we aggregate drift features across K trials:

$$\begin{aligned}\phi(x_0, c) &= \mathrm{Agg}_{k=1}^{K}\, \phi(x_0, t_k, \epsilon_k, c), \\ t_k &\sim \mathcal{U}\{1, \dots, T\},\ \epsilon_k \sim \mathcal{N}(0, I),\end{aligned} \tag{23}$$

where Agg is the mean over trials for all main and appendix results; max aggregation is not used unless explicitly reported as a separate ablation.

Instrumented modules and signals. We instrument a small set $\mathcal{S}$ of U-Net/DiT blocks (and cross-attention in text-conditioned diffusion). For each $\ell \in \mathcal{S}$, with activation a_ℓ, we compute raw drift d_ℓ^{raw} as in Eq. (16). We additionally track diffusion-specific output drifts:

$$\begin{aligned}\Delta\mathcal{L}_{\mathrm{diff}} &= \mathcal{L}_{\mathrm{diff}}(\theta^+) - \mathcal{L}_{\mathrm{diff}}(\theta), \\ d^{\epsilon} &= \|\epsilon_{\theta^+}(x_t, t, c) - \epsilon_\theta(x_t, t, c)\|_2.\end{aligned} \tag{24}$$

For text-conditioned models, binding drift uses the same attention divergence d_ℓ^{attn} (Eq. (17)).

Superposition-aware drift (optional). We train SAEs on selected denoiser activation spaces (optionally per timestep bucket) and compute SAE-coordinate drift:

$$\begin{aligned}\Delta z_\ell &= z_\ell(\theta^+) - z_\ell(\theta), \\ d_\ell^{(1)} = \|\Delta z_\ell\|_1, &\quad d_\ell^{(0)} = \|\mathbf{1}[\Delta z_\ell \neq 0]\|_1.\end{aligned} \tag{25}$$

Top-k drifting coordinates provide mechanistic handles on *which* denoising features and bindings carry membership evidence.

Feature vector and score. Per trial (t, ϵ) we compute $\phi(x_0, t, \epsilon, c)$ by concatenating $\Delta\mathcal{L}_{\mathrm{diff}}$, d^{ϵ}, and internal drift features; we then aggregate with Eq. (23) and score via $s = w^\top \phi + b$ as in Sec. 4.5. Parameters are restored after each trial.

D Deployment Scenarios

Based on our evaluation protocol, we group baselines by *access regime* to ensure fair comparison. For VLMs, **label-only** attacks rely solely on hard predicted outputs, as in LOMIA; **black-box** attacks access the model only through queries, without gradients or internal states, as in KCMP; **gray-box** attacks leverage richer output-side information such as logits, confidence scores, or calibrated statistics, as in MaxR, LiRA, and EncoderMI when embedding-level outputs are available. For diffusion models, **black-box** baselines include ReDiffuse, SimA, and SecMI, while **gray/white-box** references include LiRA-style calibrated reconstruction/noise-loss scoring and gradient-rich baselines such as GSA. Figure 6 illustrates the applicability of M-drift in different settings.

Applicability. M-Drift is most appropriate in *cooperative auditing* or internal red-team settings where the evaluator has checkpoint access, can compute gradients, and can temporarily perturb a

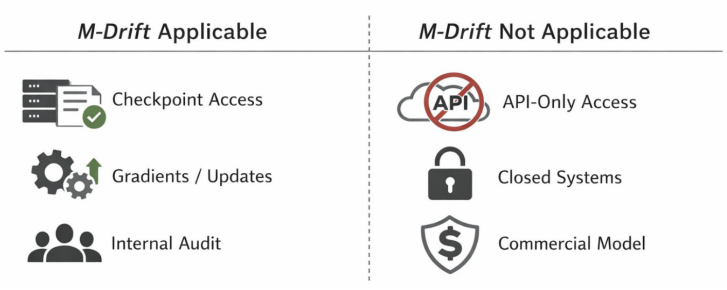


Figure 6: Applicability of M-Drift across access regimes: the method requires internal model access (e.g., gradients or activations) and is therefore suitable for cooperative audits or white/gray-box settings but not for strictly API-only deployments.

chosen parameter subset θ_M before restoring the original weights. In this setting, the method is attractive because it combines low-FPR calibration with mechanistic readouts that go beyond output likelihood alone.

Limitations. M-DRIFT is *not directly applicable* to strict API-only deployments, closed commercial systems, or settings where weights, gradients, or temporary parameter updates are unavailable. In those cases, label-only or black-box MIAs are the correct comparators, since prior work shows that attack design must match the available interface, especially for low-FPR auditing and diffusion-model privacy analysis. Finally, its reliability depends on careful low-FPR calibration and well-matched reference data, and performance may degrade under strong prompt, conditioning, or distribution shift.

D.1 CALIBRATION SPLIT CONSTRUCTION

M-Drift uses a disjoint calibration split (Table 4), to learn the score normalizer and decision threshold, but the calibration examples are never reused as audit candidates. This prevents candidate leakage and follows calibrated MIA practice, where thresholds are estimated from auxiliary member/non-member distributions rather than from the target audit set. In realistic audits, such calibration data can be obtained in several ways depending on the access setting.

For each benchmark, we split calibration and audit examples at the sample-ID level. The calibration set is used only to select the local probe magnitude η, standardize drift features, fit the logistic score $s(x, c) = w^\top \phi(x, c) + b$, and choose low-FPR or dataset-level thresholds. The final audit set is then evaluated once using the frozen calibration procedure. For VLMs and LAION-style diffusion audits, we additionally match member and non-member calibration pools by available metadata such as language, image resolution, caption length, aesthetic score, and coarse visual category, and apply near-duplicate filtering when hashes or CLIP-similarity checks are available. For DiT/ImageNet, we use the standard train/validation split with class-frequency matching. For DiT, both member and non-member examples use their ground-truth ImageNet class labels; class conditioning is therefore identical across membership groups.

This construction reflects the intended use of M-Drift as a cooperative or white-box audit tool: the auditor may have access to released training mixtures, model-owner records, or benchmark splits, but should not tune thresholds on the candidate dataset being audited. When exact training logs are unavailable, M-Drift can still be calibrated using high-confidence included data and a conservative matched reference pool; however, we explicitly report this as a weaker calibration setting and include it in the limitations.

E DETAILED EXPERIMENTAL SETUP

VLM probe implementation. For the reported LLaVA-1.5-7B and Qwen2-VL-2B-Instruct experiments, the backbone parameters are frozen and θ_M consists only of PEFT LoRA parameters ($r = 4$, $\alpha = 8$, dropout $= 0.05$) attached to q_{proj}, k_{proj}, v_{proj}, o_{proj}, up_{proj}, $down_{\text{proj}}$, and $gate_{\text{proj}}$ modules. We compute $g = \nabla_{\theta_M} \mathcal{L}$, clip its global norm to $c = 1.0$, and perform exactly one AdamW

Table 4: Practical sources for disjoint calibration examples in M-Drift.

Audit setting	Member calibration source	Non-member calibration source
Open-source model	Released training shards, curated pre-training mixtures, model-card dataset lists	Held-out shards, later crawls, or matched public data outside the training window
Internal / owner audit	Training logs, data lake records, deduplicated training manifests	Owner-provided holdout data or post-training clean reference pools
Benchmark model	Official train split, e.g., ImageNet train for DiT	Official validation/test split with class-frequency matching
LAION-style audit	Known included shard or timestamp range	Temporally clean shard matched by language, resolution, caption length, and aesthetic filters
No exact logs	High-confidence included examples from released mixtures	Conservative matched non-member pool with near-duplicate filtering

ascent step with learning rate $\eta = 10^{-4}$:

$$\theta_M^+ = \mathrm{AdamWStep}(\theta_M, -\mathrm{Clip}(g, c); \eta). \tag{26}$$

The negative gradient is supplied because AdamW is conventionally implemented as a minimizer. The original adapter parameters and optimizer state are restored/reset after each query, so probes do not accumulate across candidates.

Fixed low-rank probe initialization. The LoRA probe is instantiated once per experimental seed with a fixed random seed and is kept identical across all member and non-member candidates within that run. Under the default PEFT initialization, A is initialized once while $B = 0$, making the initial adapter a no-op. Consequently, the first probe step operates in a fixed low-rank subspace rather than introducing candidate-specific randomness. We therefore interpret LoRA-based localization as sensitivity of the pretrained model within a controlled low-rank probe subspace, rather than claiming that the LoRA coordinates themselves constitute native memorization circuits.

Implementation note. For the paper, we report Table 5 for VLMs, Table 6 for Diffusion models, and Table 7 for general configurations. **Table 6** reports the backbone, noise-conditioning, timestep sampling, and probing choices used for diffusion-model evaluation under M-Drift. **Table 7** lists the shared audit protocol, calibration setup, evaluation metrics, runtime conditions, and stored features common across all M-Drift experiments. Together with the exact probed subset θ_M, calibration split construction, and whether diffusion results use single-trial or K-trial aggregation, since these choices most strongly affect reproducibility and low-FPR behavior.

E.1 Controlled Membership Ground Truth and Distribution-Shift Controls

Why controlled membership is required. For large released foundation models, public dataset provenance does not necessarily provide exact per-example training membership. In particular, the released training information for some VLMs does not expose a complete sample-level training manifest, and temporal separation of member and non-member pools can introduce distribution shift that itself becomes predictive of the membership label (Dubiński et al., 2024; Das et al., 2025; Duan et al., 2024). We therefore separate *verified controlled membership* from *proxy provenance* and use the former as the primary validation of M-Drift.

Controlled exposure protocol. For each released checkpoint M_0, we first construct a source-matched pool and partition it at the sample-ID level into $\mathcal{D}_m^{\mathrm{cal}}$, $\mathcal{D}_m^{\mathrm{audit}}$, $\mathcal{D}_{nm}^{\mathrm{cal}}$, and $\mathcal{D}_{nm}^{\mathrm{audit}}$. Only $\mathcal{D}_m^{\mathrm{cal}} \cup \mathcal{D}_m^{\mathrm{audit}}$ is used in a controlled continuation step,

$$M_1 = \mathcal{A}\big(M_0, \mathcal{D}_m^{\mathrm{cal}} \cup \mathcal{D}_m^{\mathrm{audit}}\big),$$

where $\mathcal{A}$ denotes the corresponding VLM or diffusion training objective. Neither non-member split is used by this continuation. M-Drift's calibrator is subsequently fitted only using the calibration partitions and is frozen before evaluation on the audit partitions.

Table 5: M-Drift settings for vision-language models.

Component	Setting	Value / Description
VLM backbone	Model	`llava-hf/llava-1.5-7b-hf` and Qwen2-VL-2B-Instruct
VLM quantization	Memory setting	NF4 4-bit quantization, double quantization, and float16 mixed-precision GPU compute
VLM prompt	Query template	`USER:<image> Describe the image in one short sentence. ASSISTANT:`
Image preprocessing	VLM input	Maximum image side = 336 pixels; bicubic resize
Caption truncation	Text preprocessing	Maximum caption length = 160 characters
Representation feature	VLM embedding	Mean-pooled final hidden states over the last 6 supervised answer tokens
Probe adapter / probed subset	PEFT	LoRA with rank r=4, α=8, dropout 0.05, bias=`none`
LoRA targets	Updated modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `up_proj`, `down_proj`, and `gate_proj`
Probe update	Optimization	AdamW, learning rate 10^{-4}, gradient clipping = 1.0

Table 6: M-Drift settings for diffusion models.

Component	Setting	Value / Description
Diffusion backbone	Model	Stable Diffusion v1.5 and `facebook/DiT-XL-2-256`
Diffusion input	Image preprocessing	RGB conversion, shortest-side resize to 256, center crop to 256×256, pixel range $[-1, 1]$
Latent scaling	VAE factor	0.18215
Diffusion timestep range	Probe window	Uniformly sampled $t \in [50, 950]$
Diffusion MC budget	Trials	$K = 8$ independent trials per candidate; each trial samples one (t_k, ϵ_k) that is held fixed between the pre- and post-nudge forward passes; the eight trial-level drift vectors are mean-aggregated.
Noise / label setup	Conditioning	Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$; for DiT, the ground-truth ImageNet class label is used for both train-member and validation-non-member examples.
Diffusion probe subset	θ_M	Attention-related parameters preferred; fallback to all trainable parameters if unavailable
Diffusion output drift	Predicted-noise drift	$d^{\epsilon} = \|\epsilon_{\theta^+}(x_t, t, c) - \epsilon_{\theta}(x_t, t, c)\|_2$, computed on the full predicted-noise tensor before MC aggregation.

Table 7: Shared experimental and implementation details for M-Drift.

Component	Setting	Value / Description
Data split	Audit set size	1000 members + 1000 non-members per benchmark; disjoint calibration split
Random seeds	Repetitions	5 seeds for final reporting; typical code seeds: 7, 42, 123
Scoring model	Classifier	Logistic regression with median imputation, standardization, and balanced class weights
Classifier solver	Optimization	`lbfgs`, `max_iter`=5000, 5-fold stratified CV within the calibration split only, followed by refitting on the full calibration split; the audit set is never used for model selection.
Core probe	Update type	Single-step gradient-ascent nudge with parameter restoration after each query
Probe step size	η	10^{-4} by default; selected on the calibration split as the smallest stable loss-increasing step
Aggregation	Pointwise score	Concatenated drift vector $\phi(x, c)$ followed by a calibrated logistic score
Low-FPR metrics	Main evaluation	ROC-AUC, PR-AUC, TPR@0.1%FPR, TPR@1%FPR, membership advantage, and dataset-level audit score
Runtime	Hardware	NVIDIA V100, FP16 mixed precision, batch size 1; VLM and diffusion ($K = 8$) latency reported separately.

Matching the audited training event. For VLMs, membership is defined for the complete supervised episode rather than for an image alone: the image, prompt template, and target caption used during auditing are identical to those explicitly presented during controlled training. This removes ambiguity caused by auditing a caption or prompt that was never part of the relevant training event. For SDv1.5, member examples are explicitly presented under the standard noise-prediction objective, with the same conditioning convention used during auditing. Member and non-member examples are sampled from the same source pool and matched by available language, resolution, caption length, aesthetic score, and coarse visual content; consequently, membership is not defined by a temporal cutoff.

Table 8: Temporal-bias controls for the secondary LLaVA proxy-provenance diagnostic.

Method	ROC-AUC	TPR@0.1%FPR
Raw image classifier	0.58	3.1
Text-only classifier	0.61	4.5
CLIP embedding classifier	0.66	8.7
Metadata/time control	0.57	2.8
M-Drift	**0.98**	**80.2**

Interpretation. Importantly, because the historical pretraining data of M_0 are not fully observable, we do not claim that a controlled non-member was never encountered at any point in the model's entire history. Rather, the label is exact with respect to the controlled continuation event: a member was explicitly included and a non-member was explicitly withheld. We therefore refer to these results as *controlled-continuation membership*. The original LAION-provenance experiments are retained only as a secondary in-the-wild stress test and are reported as *proxy membership*, not as definitive per-example ground truth.

Model-specific construction. For LLaVA-1.5-7B, this protocol avoids interpreting LCS/LAION membership as membership of the frozen CLIP encoder or language backbone; the designated examples are explicitly incorporated into the evaluated multimodal checkpoint before auditing. For Qwen2-VL-2B-Instruct, we likewise make no claim that the released checkpoint was originally trained on LAION and establish membership only through the controlled continuation. For SDv1.5, controlled continuation avoids inferring individual membership from broad LAION provenance. DiT/ImageNet is treated separately because its official training and validation partitions provide directly known membership.

Temporal-bias controls. Because LAION-style member and non-member pools may differ temporally, we test whether separability is explained by dataset shift rather than membership evidence. We train blind controls that do not access the audited model: raw image features, text-only features, CLIP image–text embeddings, and timestamp/metadata cues. These controls are evaluated on the same member/non-member splits and calibration protocol as M-Drift. As shown in Table 8, these blind baselines remain far below M-Drift, especially at low FPR, indicating that the main signal is not explained by temporal or low-level distributional artifacts alone.

To verify that M-Drift does not simply exploit temporal artifacts in LAION-style member/non-member splits, we evaluate blind baselines that do not access the audited model or its gradients. We train classifiers on: (i) raw image statistics, (ii) text/caption features, (iii) frozen CLIP image–text embeddings, and (iv) timestamp/metadata indicators when available. All controls use the same train/calibration/test split as M-Drift and are evaluated using ROC-AUC and low-FPR metrics.

We additionally apply near-duplicate filtering and match member/non-member pools by resolution, caption length, aesthetic score, language, and class/object distribution where metadata is available. The blind controls remain substantially weaker than M-Drift, especially at TPR@0.1%FPR, suggesting that M-Drift captures model-internal membership evidence rather than superficial temporal distribution shift.

For all LAION settings, we apply the same near-duplicate filtering, metadata matching, CLIP-similarity hard-negative construction, and blind controls that do not query the audited model. This ensures that membership evidence is not explained solely by temporal, caption, aesthetic, or visual-distribution artifacts.

F Reproducibility and Relation to Existing MIA Signals

F.1 Reproducibility

M-Drift is designed for straightforward reproduction from released checkpoints, fixed splits, and a small set of transparent hyperparameters. The pipeline consists of a single temporary gradient-ascent probe, pre/post drift extraction, logistic calibration on a *disjoint* split, and evaluation at stringent low-FPR and dataset-level settings. Reproducibility therefore requires fixing the checkpoint, member/non-member construction, probed subset θ_M, step size η, instrumented layers S, diffusion

Table 9: Recommended reproducibility checklist for M-Drift. Values reflect the current paper when specified; otherwise we use standard, reasonable defaults for reproducible white-box auditing.

Item	Current setting	What to release / report
Target checkpoints	LLaVA-1.5-7B, Qwen2-VL-2B-Instruct, SDv1.5, DiT-XL/2-256	Exact model identifier, commit/checkpoint hash, tokenizer, scheduler, and VAE version if applicable
Data size	2000 audit examples per benchmark (1000 members + 1000 non-members), plus a disjoint calibration split	File lists or sample IDs for member, non-member, calibration, and dataset-level audit sets
Membership construction	Controlled-continuation membership for LLaVA/Qwen/SDv1.5; exact ImageNet train/validation membership for DiT; provenance-based LAION analysis retained only as a secondary stress test	Exact filtering rules, timestamps/cutoffs, class balancing, and any deduplication or near-duplicate removal
Probe update	Single-step gradient-ascent nudge; parameters restored after each query	η grid, selected η, exact update/optimizer rule, optimizer hyperparameters, initialization, and gradient clipping
Probed subset θ_M	Full model or restricted modules; paper emphasizes restricted θ_M when needed	Exact modules/layers, parameter count, and rationale for the restriction
Instrumented signals	ΔL, raw activation drift, cross-attention drift, optional SAE drift	Exact layer set S, tensor shapes, pooling/flattening, normalization, and feature dimensionality
Diffusion MC aggregation	Default $K{=}8$, mean aggregation for all reported results	RNG seed policy, timestep sampling rule, aggregation operator, and per-trial compute cost
Scoring / calibration	Logistic calibration over concatenated drift vector; dataset-level threshold on disjoint null sets	Feature standardization, train/validation protocol for the calibrator, threshold selection, and confidence-interval method
Seeds and uncertainty	5 seeds for ROC-AUC / PR-AUC; low-FPR std omitted in main tables	All random seeds, bootstrap procedure, and full variance for TPR@0.1%, TPR@1%, and dataset-level metrics
Hardware / runtime	NVIDIA V100, FP16 mixed precision, batch size 1; VLM and diffusion ($K = 8$) latency reported separately	GPU type, precision, memory footprint, and wall-clock breakdown for forward, backward, re-forward, and logging
SAE configuration	Sparse feature projection for interpretability	Activation source, dictionary width, sparsity λ, optimizer, training steps, reconstruction loss, and sparsity level k
Code release	Provided as Supplementary materials	Minimal scripts for split generation, probe extraction, calibration, evaluation, and figure recreation

budget K, and calibration protocol. In our setup, we use LLaVA-1.5-7B, Qwen2-VL-2B-Instruct, SDv1.5, and DiT-XL/2-256, with balanced splits, a disjoint calibration set, and default diffusion aggregation over $K{=}8$ trials. All appendix tables use the same final audit splits, calibration protocol, and random seeds as Tables 1-2.

As mention in Table 9, when SAE-based drift is enabled, reproducibility additionally depends on the activation source, normalization, dictionary width, sparsity coefficient, and sparse-drift statistic. We therefore recommend releasing either pretrained SAE checkpoints or the full SAE training recipe, along with reconstruction and sparsity summaries.

Clarifications. M-Drift is reproducible without proprietary APIs or hidden logs: it only requires released weights, forward/backward access, and a reproducible member/non-member split. We choose η on a disjoint calibration split as the smallest stable local probe, and report θ_M explicitly because it

Table 10: Matched decomposition of white-box membership signals. All rows use the same θ_M, layer set $\mathcal{S}$, calibration split, and linear calibrator. "Static" denotes features extracted before the nudge; "drift" denotes pre/post changes induced by the single-step M-Drift probe. Higher is better.

Signal family	LLaVA / LAION-SBU			SDv1.5 / LAION-5B		
	ROC-AUC	TPR@1%	Audit	ROC-AUC	TPR@1%	Audit
Loss only, static $[\mathcal{L}(\theta; x, c)]$	0.700	8.0	0.620	0.650	12.0	0.600
Gradient norm, static $[\|\nabla_{\theta_M}\mathcal{L}\|_2]$	0.884	58.6	0.771	0.852	49.3	0.748
Gradient statistics, static	0.913	66.9	0.807	0.881	55.7	0.782
Activation features, no nudge	0.928	72.8	0.842	0.904	63.5	0.821
Attention features, no nudge	0.941	77.4	0.865	0.918	67.8	0.842
Static gradient + activation + attention	0.957	84.6	0.903	0.934	74.2	0.872
Loss drift only $[\Delta\mathcal{L}]$	0.910	68.0	0.800	0.880	55.0	0.780
Raw activation drift	0.956	86.7	0.914	0.939	77.6	0.889
Attention / binding drift	0.964	89.8	0.925	0.946	81.2	0.904
M-Drift without SAE drift	0.973	92.0	0.938	0.953	85.5	0.917
Full M-Drift	**0.980**	**94.9**	**0.952**	**0.960**	**88.0**	**0.930**

affects runtime and interpretability. For evaluation, full ROC/PR curves, low-FPR operating points, dataset-level power curves, and uncertainty estimates should be reported.

F.2 Matched Decomposition of White-Box Signals

To isolate the marginal contribution of the interventional drift probe, we compare M-Drift against matched white-box decompositions that use the same probed parameter subset θ_M, instrumented layers $\mathcal{S}$, calibration split, feature standardization, and logistic scorer. Static baselines are computed from the original model f_θ without applying the nudge, while drift variants use the same single-step update and restore parameters after each query. This controls for access level and calibrator capacity, ensuring that gains are attributed to the intervention-induced change rather than to privileged gradients or activations alone. Table 10 reports matched decomposition of white-box membership signals.

The decomposition shows that static white-box signals are already informative, confirming that gradients and activations carry membership evidence. However, replacing static features with intervention-induced drift consistently improves low-FPR detection and dataset-level auditing. The largest gains occur when raw representation drift, cross-modal/attention drift, and SAE-coordinate drift are fused, supporting the central claim that M-Drift contributes beyond access to gradients alone by probing how model mechanisms locally respond to a controlled loss-increasing perturbation.

F.3 Novelty: Drift as the Unit of Membership Evidence

A possible misconception is that M-Drift is simply a combination of existing ingredients: gradient perturbation, attention analysis, and SAE-based visualization. The intended contribution is different. M-Drift proposes *post-intervention mechanistic drift* as the primary unit of membership evidence. Instead of asking whether a member has lower loss, larger confidence, or larger raw gradient, M-Drift asks how the model's internal mechanisms change after a small, reversible, loss-increasing probe. The resulting evidence is the structured difference between the pre- and post-nudge model states:

$$\phi(x, c) = \big[\Delta\mathcal{L}, \Delta a_\ell, \Delta A_\ell, \Delta z_\ell\big], \tag{27}$$

where Δa_ℓ captures representation drift, ΔA_ℓ captures cross-modal binding drift, and Δz_ℓ captures sparse feature-coordinate drift.

This framing differs from prior static white-box MIAs in three ways. First, the score is based on a *local counterfactual response* of the model, not only on the model state at θ. Second, the same intervention supports both VLMs and diffusion generators, while the readout is architecture-aware: cross-attention redistribution for VLMs and Monte-Carlo aggregation over denoising timesteps for diffusion. Third, sparse features are not introduced merely for visualization; they provide a coordinate system for measuring whether membership-sensitive drift is concentrated in a small set of interpretable directions. Thus, M-Drift contributes a unified audit principle: membership is detected by how a trained model's mechanisms move under a controlled local perturbation.

Table 11: Low-FPR uncertainty for the M-Drift point estimates reported in Tables 1–2. Each seed contains 1,000 non-members. TPR@0.1%FPR and TPR@1%FPR are estimated independently within each seed and subsequently aggregated across the five seeds; brackets report 95% bootstrap confidence intervals.

Setting	Method	TPR@0.1%FPR	TPR@1%FPR
LLaVA / LAION-SBU	M-Drift	80.2 [76.4, 83.7]	94.9 [92.8, 96.4]
Qwen / LAION-en/zh	M-Drift	67.8 [63.6, 71.5]	90.1 [87.4, 92.2]
SDv1.5 / LAION-5B	M-Drift	70.0 [66.1, 73.8]	88.0 [85.0, 90.7]
DiT / ImageNet	M-Drift	48.0 [44.0, 52.2]	72.0 [68.5, 75.4]

G Detailed Result

G.1 Dataset-level Auditing Protocol

For a suspect set D of size n_D, we use the mean calibrated log-odds

$$T(D) = \frac{1}{|D|} \sum_{x \in D} s(x)$$

as the dataset statistic. Using disjoint clean null sets, the rejection threshold is

$$\tau_\alpha = Q_{1-\alpha}(T(D_0)) \,,$$

and power at contamination level ρ is

$$P_\alpha(\rho) = \Pr[T(D_\rho) > \tau_\alpha] \,.$$

We evaluate $\rho \in \{0, 0.01, 0.05, 0.10\}$ and summarize the resulting power-contamination curve by the normalized trapezoidal area

$$\mathrm{DLS} = \frac{1}{\rho_{\max}} \sum_j \frac{P_\alpha(\rho_j) + P_\alpha(\rho_{j+1})}{2} (\rho_{j+1} - \rho_j).$$

G.2 Low-FPR Tail Estimation Protocol

Each evaluation seed contains $N_{\mathrm{nm}} = 1000$ held-out non-members and $N_{\mathrm{m}} = 1000$ members. Thus, the minimum directly resolved nonzero FPR is $1/1000 = 0.1\%$. Within each seed, TPR@0.1%FPR uses the largest threshold allowing at most one false positive, whereas TPR@1%FPR allows at most ten false positives. We compute these operating points independently for each seed and report the aggregate across five seeds; we do not treat repeated evaluations across seeds as additional independent non-member examples.

Because TPR@0.1%FPR is determined by an extreme order statistic, we treat TPR@1%FPR as the primary low-FPR result and TPR@0.1%FPR only as supporting tail evidence. Confidence intervals are obtained by stratified bootstrap resampling with the operating threshold recomputed within each replicate. Low-FPR metrics with uncertainty are reported in Table 11.

This protocol avoids overstating single-split resolution at extreme tails. It also separates the main directly measured operating point, TPR@1%FPR, from the more statistically delicate TPR@0.1%FPR analysis. In the main paper, we therefore emphasize TPR@1%FPR and use TPR@0.1%FPR as supporting evidence for tail behavior.

G.3 Detailed Ablation Study

Table 12. extends the main ablation study and supports four conclusions: (i) a *single local* gradient-ascent probe is the best trade-off. The matched no-nudge baseline confirms that static white-box

Table 12: Detailed ablation study for M-Drift, extending Table 3 in the main paper. The single-step nudge, calibrated fusion, and mechanistic drift signals are the strongest configuration, while removing or weakening them mainly hurts low-FPR and dataset-level auditing. The *Full M-Drift* configuration corresponds to the default setup with a single-step gradient-ascent nudge (η), full calibrated drift-feature fusion (loss, representation, attention, and SAE drift), SAE dictionary size d_{SAE}=2048, and diffusion aggregation with K=8 trials. Higher is better for all metrics.

	LLaVA / LAION-SBU			SDv1.5 / LAION-5B		
Ablation Variant	**ROC-AUC**	**TPR@1%**	**DCD**	**ROC-AUC**	**TPR@1%**	**DL-PC**
Full M-Drift	0.980	94.9	0.952	0.960	88.0	0.930
Nudge sensitivity / intervention design						
$\eta = 0$ (no nudge)	0.957	84.6	0.903	0.934	74.2	0.872
$\eta = \eta/3$	0.971	90.8	0.938	0.948	82.6	0.909
η (default)	0.980	94.9	0.952	0.960	88.0	0.930
$\eta = 3\eta$	0.969	89.5	0.931	0.944	80.7	0.902
Two-step nudge ($2 \times \eta/2$)	0.974	91.7	0.941	0.952	84.2	0.916
Two-step nudge ($2 \times \eta$)	0.962	86.1	0.920	0.936	76.3	0.888
Restricted θ_M (bridge/attn only)	0.977	93.1	0.946	0.955	86.4	0.924
Signal family ablations						
Loss-drift only ($\Delta\mathcal{L}$; no feature fusion)	0.910	68.0	0.800	0.880	55.0	0.780
w/o cross-modal / attention drift	0.972	91.8	0.935	0.951	84.9	0.912
w/o SAE drift (raw drift only)	0.973	92.0	0.938	0.953	85.5	0.917
w/o output drift ΔL	0.968	88.6	0.927	0.947	81.4	0.904
Vision-only drift	0.940	79.0	0.880	0.930	73.0	0.870
Text/conditioning-only drift	0.960	85.0	0.900	0.900	62.0	0.820
Output-only (no gradients)	0.700	8.0	0.620	0.650	12.0	0.600
SAE design ablations						
SAE dictionary $d_{\text{SAE}} = 512$	0.972	91.1	0.936	0.950	84.1	0.913
SAE dictionary $d_{\text{SAE}} = 1024$	0.978	94.0	0.949	0.958	87.0	0.927
SAE dictionary $d_{\text{SAE}} = 2048$	0.980	94.9	0.952	0.960	88.0	0.930
SAE dictionary $d_{\text{SAE}} = 4096$	0.979	94.6	0.951	0.959	87.6	0.929
Sparse-only support metric $d_\ell^{(0)}$	0.969	89.3	0.930	0.946	80.9	0.901
Magnitude-only SAE metric $d_\ell^{(1)}$	0.975	92.7	0.942	0.954	85.8	0.919
Cosine-normalized SAE drift only	0.973	91.4	0.938	0.951	84.3	0.914
Training / compute ablations						
SAE train steps = 25k	0.974	92.0	0.939	0.952	84.9	0.916
SAE train steps = 50k	0.978	94.0	0.949	0.958	87.1	0.927
SAE train steps = 100k	0.980	94.9	0.952	0.960	88.0	0.930
Diffusion MC trials $K = 1$	–	–	–	0.944	79.1	0.896
Diffusion MC trials $K = 4$	–	–	–	0.956	85.8	0.922
Diffusion MC trials $K = 8$	–	–	–	0.960	88.0	0.930
Diffusion MC trials $K = 16$	–	–	–	0.961	88.4	0.932

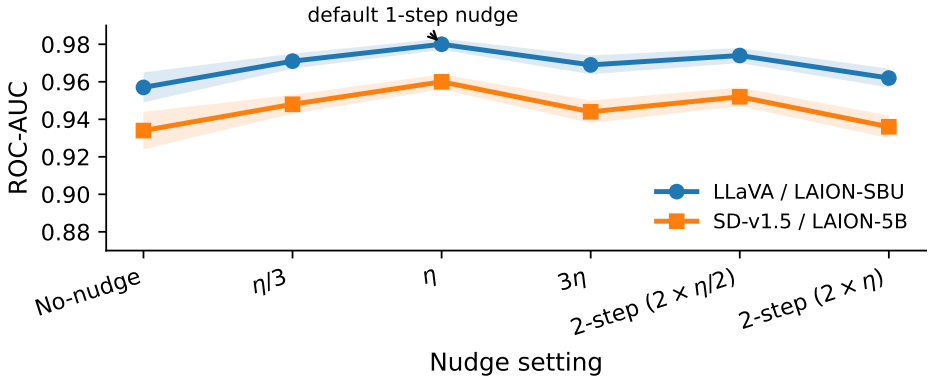


Figure 7: Nudge sensitivity of M-Drift on LLaVA and SD-v1.5. A single moderate gradient-ascent nudge achieves the best ROC-AUC, while removing the intervention or using overly aggressive/repeated nudges slightly degrades separability.

signals are already informative, while the single-step intervention further improves low-FPR detection and dataset-level auditing; (ii) calibrated fusion of output, raw drift, attention drift, and SAE

Table 13: Quantitative localization validation. Top-k features/patches selected by M-Drift are concentrated, stable, and measurably affect the membership score.

Setting	Top-k conc. ↑	Stability ↑	Target del. Δs ↓	Random del. Δs ↓
LLaVA / cross-attn patches	0.71	0.63	-0.34	-0.08
LLaVA / SAE features	0.76	0.68	-0.39	-0.10
SDv1.5 / U-Net features	0.69	0.59	-0.31	-0.07
SDv1.5 / SAE features	0.74	0.66	-0.36	-0.09

drift is consistently stronger than any single signal family alone; (iii) SAE-based drift helps most in the low-FPR regime, with moderate-to-large dictionaries (roughly 1k–2k) saturating performance; and (iv) diffusion benefits from Monte-Carlo aggregation up to about $K = 8$, after which gains are marginal relative to added compute.

As shown in Fig. 7, M-Drift is most effective with a single moderate nudge, while removing the probe or using overly aggressive multi-step updates slightly reduces ROC-AUC on both LLaVA and SD-v1.5.

G.4 Quantitative Localization Validation

The qualitative heatmaps in Fig. 4-(A) and 4-(B), illustrate where M-Drift evidence concentrates, but heatmaps alone do not establish that the localized regions or sparse features are causally meaningful. We therefore add quantitative localization tests for both VLM and diffusion settings. For each candidate, we rank cross-attention patches or SAE coordinates by drift magnitude after the single-step nudge. We then evaluate: (i) **Top-k concentration**, the fraction of total drift explained by the top-k patches/features; (ii) **Seed stability**, the average overlap of top-k localized features across random seeds; (iii) **Deletion sensitivity**, the drop in membership score after masking the top-k localized patches/features; and (iv) **Random deletion**, the same masking budget applied to randomly selected patches/features. Strong localization should yield high concentration, stable top-k features, and a larger score drop under targeted deletion than random deletion.

Table 13 shows that membership evidence is not diffusely spread across all units. In both VLM and diffusion models, a small set of high-drift patches or SAE coordinates explains most of the drift mass. These localized regions are also stable across seeds and, when masked, reduce the calibrated membership score substantially more than random masking with the same budget. This supports the claim that M-Drift provides localized mechanistic evidence rather than only post-hoc visualization.

G.5 Qualitative Result

Figure 8 The membership score is computed by concatenating output drift, internal representation drift, cross-attention binding drift, and optional SAE-coordinate drift into a feature vector $\phi(x, c)$, and then applying a calibrated linear scorer $s(x, c) = w^\top \phi(x, c) + b$. In this figure, the member example receives consistently higher scores, indicating that the single-step nudge induces stronger and more structured mechanistic changes for members than for non-members. This qualitative gap supports the paper's claim that membership evidence is not only stronger in M-Drift, but also more interpretable through localized cross-modal and feature-level drift.

Figure 9. In diffusion models, the membership score is computed by aggregating drift features over Monte Carlo samples of timestep-noise pairs (t, ϵ), forming a feature vector from denoising-loss drift, predicted-noise drift, and optional internal/attention drift, and then applying a calibrated linear scorer $s(x, c) = w^\top \phi(x, c) + b$. In this figure, the member case shows more stable same-prompt generations and a clearly higher score, while the non-member remains below the decision threshold, illustrating that M-Drift captures prompt-conditioned memorization evidence beyond visual similarity alone.

G.6 Failure-Case Analysis

We analyze high-confidence false positives (FPs) and false negatives (FNs) to understand when M-Drift fails. Since M-Drift is designed for low-FPR auditing, FPs are especially important: they

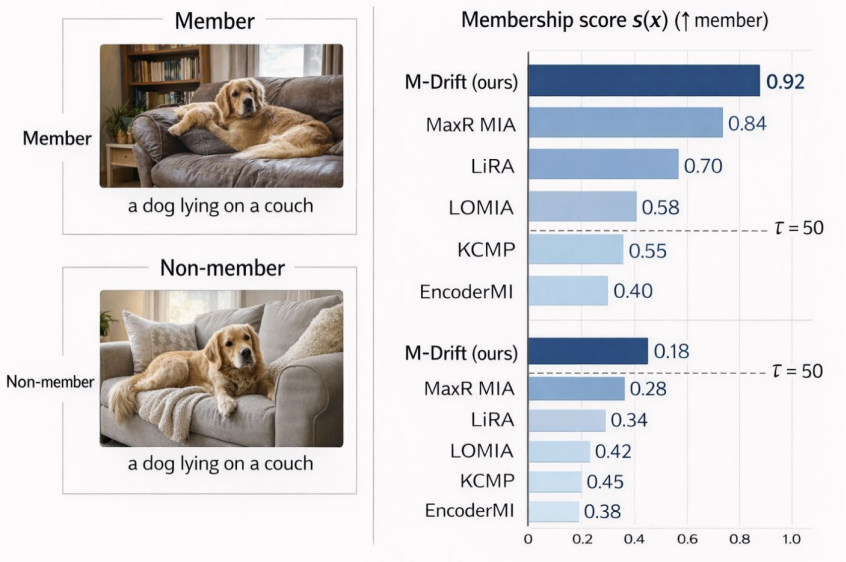


Figure 8: VLM membership evidence on LLaVA/LAION: M-Drift converts concatenated drift features into a calibrated score and shows clearer member/non-member separation than prior baselines on matched image-caption examples.

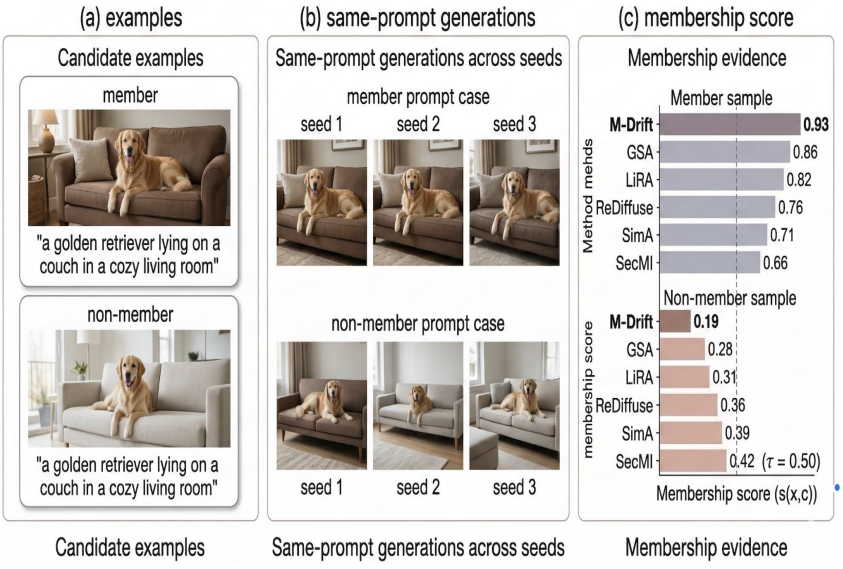


Figure 9: Prompt-conditioned membership evidence for SDv1.5 on LAION-5B: under the same prompt, member examples yield more consistent generations and a larger calibrated score than non-members.

correspond to non-member examples whose post-nudge drift resembles member-like memorization evidence. Table 14 summarizes the dominant error modes observed across the VLM and diffusion settings.

The main FP modes arise from examples that are not literal training members but are highly similar to training data: near-duplicates, templated captions, and common web concepts. These cases are expected in web-scale corpora and suggest that strict duplicate removal and concept-frequency matching are important for conservative audits. The main FN modes occur when the member signal is weakly expressed under the chosen prompt, caption, or diffusion timestep. Prompt ensembling for VLMs and stratified timestep sampling for diffusion reduce these failures. Overall, the failure

Table 14: Failure-case patterns for M-Drift. FP = non-member predicted as member; FN = member predicted as non-member.

Error	Setting	Pattern	Likely cause / mitigation
FP	LLaVA / LAION-SBU	Near-duplicate web image	Visually near-identical image–caption pairs induce member-like cross-attention drift. Mitigate with perceptual/CLIP deduplication.
FP	Qwen / LAION-en/zh	Templated caption	Repeated caption forms such as "a photo of ..." produce stable text-side drift. Mitigate by caption-template controls.
FP	SDv1.5 / LAION-5B	Common web concept	Frequent concepts, e.g., pets, landmarks, product photos, yield strong generic denoising drift. Mitigate with concept-frequency matched references.
FP	DiT / ImageNet	Highly canonical class image	Prototype-like validation images resemble training-class modes. Mitigate with class-frequency and visual-density matching.
FN	LLaVA / LAION-SBU	Weak image–text binding	Member image has noisy or weakly aligned caption, reducing cross-modal drift. Mitigate with prompt/caption normalization.
FN	Qwen / LAION-en/zh	Prompt-sensitive response	Membership score changes under paraphrased prompts. Mitigate by prompt ensembling.
FN	SDv1.5 / LAION-5B	Low-signal timestep region	Sampled timesteps miss the region where member/non-member curves separate. Mitigate with stratified timestep sampling.
FN	DiT / ImageNet	Over-regularized class example	Strong class-level generalization suppresses per-example drift. Mitigate with layer-wise or class-conditional calibration.

modes are systematic rather than arbitrary, and they provide actionable guidance for more conservative deployment of M-Drift.

H LIMITATIONS

M-Drift currently assumes *white-box* or cooperatively accessible models, since it relies on gradients, temporary parameter nudges, and internal activations; this limits direct applicability to common API-only deployments. In addition, although the single-step probe is lightweight relative to richer white-box attacks, the method still incurs extra compute from gradient/re-forward passes and, when enabled, SAE training and feature extraction. Finally, its reliability depends on careful low-FPR calibration and well-matched reference data, and performance may degrade under strong prompt, conditioning, or distribution shift; extending M-Drift to more restricted-access and more robust calibration settings is an important direction for future work.