IMAGE MEMBERSHIP INFERENCE IN FINE-TUNED MEDICAL VISION LANGUAGE MODELS VIA DECODING SURFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Membership inference attacks (MIAs) are widely used to audit training-data leakage in fine-tuned medical vision–language models (VLMs). Existing VLM MIAs typically commit to a fixed decoding configuration or score family before evaluation, even though stochastic generation behavior can vary substantially with the inference policy. This restricted view may therefore miss stronger membership evidence elsewhere in the observable response space. We propose the *Decoding-Surface Attack* (DSA), a calibrated black-box MIA that models the multi-temperature response surface at each output budget by combining within-policy consistency, cross-temperature geometry, and response behavior. Beyond attack measurement, DSA uses the resulting leakage score to guide surface-access restriction and privacy-aware checkpoint selection, forming an audit-to-mitigation workflow without claiming formal privacy guarantees. Extensive evaluations across medical VQA benchmarks and model families compare DSA with fixed-policy, partial-surface, and recent VLM MIA baselines. DSA achieves 0.811 macro AUC in our primary evaluation, improving over fixed-policy MIA by 25.5 AUC points and the strongest reproduced VLM MIA baseline by 18.1 points. Comprehensive component ablations and same-pool controls further verify the contribution of complementary surface evidence and rule out the tested dataset shortcuts.

1 INTRODUCTION

Vision–language models (VLMs) have become a general interface for visual understanding: they answer questions, follow instructions, and generate free-form text conditioned on images (Chen et al., 2023; Li et al., 2023b; Liu et al., 2023; 2024). In medicine, this interface is particularly attractive because clinical evidence is inherently multimodal, spanning radiology images, pathology slides, reports, and visual questions (Moor et al., 2023; Tu et al., 2024; Li et al., 2023a; Zhang et al., 2024). Parameter-efficient adaptation methods such as LoRA (Hu et al., 2022) make it practical to specialize large VLMs on local clinical data, but they also create a concrete privacy risk: fine-tuning data may contain patient-derived images, rare findings, and institution-specific distributions. Even without verbatim memorization, a fine-tuned VLM may reveal whether a particular image was used for training through systematic changes in its generated answers. Membership inference attacks (MIAs) (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022) therefore provide a direct audit of privacy risk in medical VLM deployment.

Recent audits show that membership risk is not limited to closed-set classifiers. Generative models can leak membership in ways weak audits underestimate (Panda et al., 2025), and multimodal VQA systems remain vulnerable under realistic black-box access (Nguyen et al., 2025). Existing VLM MIAs expose leakage through confidence-derived scores (Li et al., 2024), temperature sensitivity (Hu et al., 2025), image-corruption robustness (Wu et al., 2025), or cross-modal semantic alignment (Li et al., 2026). Despite their methodological diversity, these attacks commit in advance to a prescribed observation protocol, extracting membership evidence through a predefined confidence score, perturbation family, alignment metric, or temperature schedule. We term this restricted perspective a fixed-policy audit. Even when it involves multiple queries, the audit observes only a predetermined slice of the model's generative behavior and may therefore miss stronger leakage exposed elsewhere in the decoding-policy space.

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

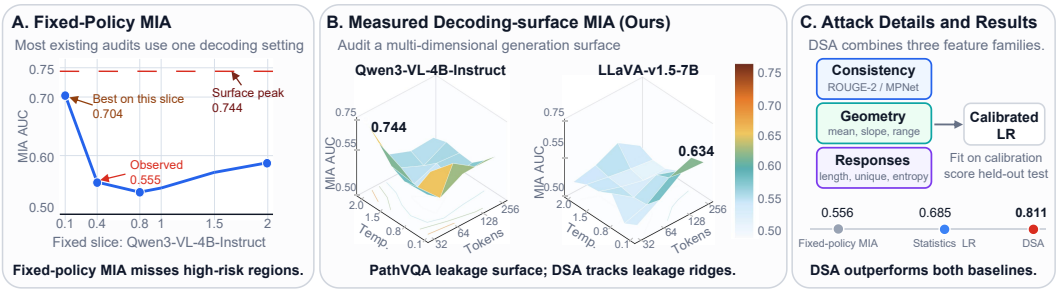


Figure 1: **Membership leakage varies across the decoding space.** (A) A fixed-policy MIA can miss higher-risk settings. (B) The empirical PATHVQA risk landscape varies across temperatures and separately evaluated output-length budgets. (C) At each fixed output budget, DSA fuses consistency, cross-temperature geometry, and response behavior into one membership score.

Our key observation is that membership evidence depends on the decoding policy used to query the model. Fig. 1 A illustrates a practical consequence of this dependence: a fixed policy may indicate only moderate leakage, while another policy reveals substantially stronger membership evidence. Fig. 1 B visualizes the broader risk landscape obtained by evaluating fixed policies across temperature and maximum generation length on PATHVQA (He et al., 2020). The landscape is not flat, and its high-risk regions differ across QWEN (Bai et al., 2025) and LLAVA (Liu et al., 2023). Motivated by this variation, we operationalize the audit at each fixed output budget by modeling the multi-temperature slice of the decoding surface; the four output budgets are evaluated separately rather than pooled into one attack.

We propose the *Decoding-Surface Attack* (DSA), a black-box MIA that turns this observation into a calibrated response-surface audit. As summarized in Fig. 1 C, DSA queries the target VLM repeatedly under multiple decoding policies and represents each candidate through three complementary aspects of generation behavior: within-policy answer consistency, cross-policy surface geometry, and response-form statistics such as length, diversity, empty responses, and entropy. A labeled calibration split is used to orient and normalize these measurements and to map them to a single membership score, which is then evaluated on held-out candidates. This surface view leads to stronger membership inference. Across QWEN and LLAVA and four output-length budgets on PATHVQA, DSA reaches 0.811 macro AUC, compared with 0.556 for fixed-policy MIA, 0.650 for an optimistic test-selected single-score oracle, and 0.630 for the strongest reproduced VLM MIA baseline (Li et al., 2024; Hu et al., 2025; Wu et al., 2025; Li et al., 2026). Ablations show that the gain comes from combining complementary surface evidence, and same-pool controls support that the signal reflects fine-tuning membership rather than image-pool or metadata artifacts. The same calibrated leakage score can also inform audit-guided deployment choices, such as restricting exposed decoding policies or selecting checkpoints with lower measured leakage.

Our contributions are:

- We identify fixed-policy auditing as a blind spot in medical VLM MIAs and formulate membership leakage as a decoding-surface property.
- We propose DSA, a calibrated black-box MIA that converts repeated generations into a label-free response-surface representation and improves over fixed-policy MIA by 25.5 AUC points on PATHVQA.
- We provide a complete empirical analysis: attack comparison against single-score and recent VLM MIA baselines, ablations quantifying the benefit of surface evidence, same-pool validity controls, and external PMC-VQA validation.
- We show that decoding-surface auditing can inform audit-guided risk control through surface-access restriction and pre-deployment checkpoint selection.

2 RELATED WORK

Membership inference and privacy auditing. Membership inference asks whether a target record was used to train a model. Early work showed that classifier confidence can reveal training member-

ship under black-box access (Shokri et al., 2017), and later analyses connected membership risk to overfitting and generalization behavior (Yeom et al., 2018). Modern attacks and audits refine this view by calibrating likelihood ratios, evaluating low-FPR regimes, and comparing examples to nearby alternatives (Carlini et al., 2022; Mattern et al., 2023). For large generative models, recent audits further show that weak attacks can give misleadingly small lower bounds on leakage (Panda et al., 2025). Our setting differs from likelihood- or loss-based audits: the adversary observes only generated text from a medical VLM and does not require parameters, gradients, logits, token probabilities, or training loss.

Membership inference for multimodal and vision–language models. Membership inference has also been extended to multimodal systems. DocMIA studies document-level leakage in DocVQA models (Nguyen et al., 2025); VL-MIA proposes confidence-based MaxRényi-$K\%$ scores for training-image detection (Li et al., 2024); Hu et al. examine practical black-box VLM leakage (Hu et al., 2025); ICIMIA exploits sensitivity to image corruption (Wu et al., 2025); and CSA-MIA uses cross-modal semantic alignment between images and generated text (Li et al., 2026). These attacks show that multimodal generations contain membership signal, but each audits a prescribed view of the model, such as confidence, perturbation sensitivity, alignment, or a selected decoding condition. We instead ask whether decoding policy itself forms a structured attack surface whose joint behavior reveals leakage missed by fixed-policy audits.

Medical VLM adaptation and privacy. Medical foundation models aim to integrate images, reports, and clinical reasoning into generalist systems (Moor et al., 2023; Tu et al., 2024; Zhang et al., 2024). Instruction-tuned biomedical VLMs such as LLaVA-Med (Li et al., 2023a) and medical VQA datasets such as PATHVQA (He et al., 2020) make this setting operational: a model is fine-tuned on image–question–answer data and later queried through natural language. Parameter-efficient adaptation such as LoRA (Hu et al., 2022) lowers the cost of local specialization, but also increases the need for deployment-time privacy audits. Our work focuses on this regime: black-box membership auditing for fine-tuned medical VLMs, where privacy evidence is expressed through generated answers rather than internal model states.

3 DECODING-SURFACE ATTACK

3.1 OVERVIEW

Fig. 2 summarizes DSA at a fixed output budget ℓ. Given a candidate image-question pair, DSA queries the target VLM across multiple temperatures and records only generated text. The repeated generations form a temperature-indexed response tensor. A label-free operator converts this tensor into within-temperature consistency, cross-temperature geometry, and response-behavior evidence. A labeled calibration split then fixes the preprocessing and scoring map used for held-out candidates. DSA is instantiated separately for each output budget; different budgets are compared experimentally but are not pooled into one attack.

3.2 BLACK-BOX RESPONSE SURFACE

Let (x, q) denote a candidate image–question pair. The broader decoding space is $\Pi = \mathcal{T} \times \mathcal{L}$, where $\mathcal{T}$ contains sampling temperatures and $\mathcal{L}$ contains output-length budgets. Fig. 1 B visualizes risk across this evaluation space. Operationally, DSA is instantiated separately for each fixed output budget $\ell \in \mathcal{L}$ and models the temperature slice

$$\Pi_\ell = \{(t, \ell) : t \in \mathcal{T}\}. \tag{1}$$

For each $t \in \mathcal{T}$, DSA draws K stochastic generations:

$$y_{t,k}^{(\ell)}(x, q) \sim M_\theta(\cdot \mid x, q, t, \ell), \qquad k = 1, \ldots, K. \tag{2}$$

The observed response tensor for budget ℓ is

$$\mathcal{Y}_\ell(x, q) = \{y_{t,k}^{(\ell)}(x, q) : t \in \mathcal{T},\ k = 1, \ldots, K\}. \tag{3}$$

The four output budgets are evaluated separately and are not concatenated into a single attack representation. DSA uses only generated text and does not access model parameters, gradients, logits, token probabilities, hidden states, or training losses.

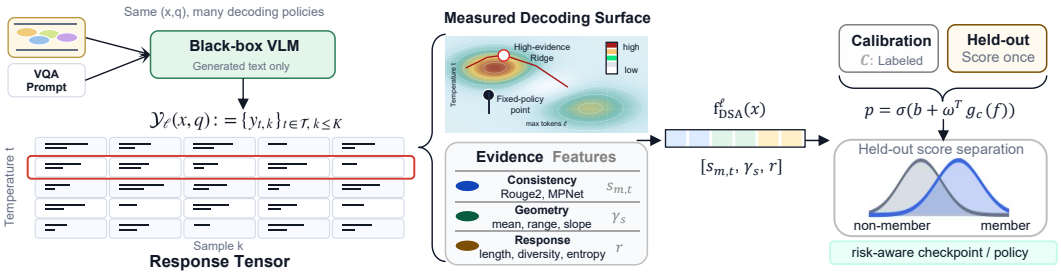


Figure 2: **Decoding-Surface Attack at a fixed output budget.** DSA summarizes repeated generations across temperatures into consistency, cross-temperature geometry, and response-behavior evidence, then calibrates the resulting surface-slice representation into held-out membership scores.

3.3 Surface Evidence Operator

DSA first converts the generated response tensor into a label-free feature vector:

$$\mathbf{f}^{(\ell)}_{\mathrm{DSA}}(x) = \Phi_\ell(\mathcal{Y}_\ell(x,q)) \,. \tag{4}$$

Here Φ_ℓ is a deterministic feature extractor: it uses only the generated strings in $\mathcal{Y}_\ell(x,q)$ and does not access membership labels. It computes three evidence families: within-policy consistency, cross-policy geometry, and response behavior.

Within-policy consistency. For a fixed output budget ℓ, and for each similarity metric $m \in \mathcal{M}$ and temperature $t \in \mathcal{T}$, DSA measures the agreement among the K responses generated under the same policy (t,ℓ):

$$s^{(\ell)}_{m,t}(x) = \frac{2}{K(K-1)} \sum_{1\le i<j\le K} \mathrm{sim}_m\Big(y^{(\ell)}_{t,i}(x,q), y^{(\ell)}_{t,j}(x,q)\Big) \,. \tag{5}$$

Thus $s^{(\ell)}_{m,t}(x)$ is the average pairwise similarity among the K sampled answers at temperature t. We use lexical ROUGE-2 F-score (Lin, 2004) and semantic MPNET embedding cosine similarity (Song et al., 2020).

Cross-policy geometry. A fixed-policy score observes only one location on the decoding surface. DSA additionally summarizes how the consistency profile changes across temperatures. For each metric m, define

$$\mathbf{s}^{(\ell)}_m(x) = \Big(s^{(\ell)}_{m,t}(x)\Big)_{t\in\mathcal{T}} \,. \tag{6}$$

Let Γ_s denote the temperature-summary operator for scalar consistency curves. It returns the mean, standard deviation, minimum, maximum, range, low-to-high temperature difference, and least-squares slope over temperature. We collect the surface-geometry block as

$$\boldsymbol{\gamma}^{(\ell)}_s(x) = \Big\{\Gamma_s\Big(\mathbf{s}^{(\ell)}_m(x)\Big) : m \in \mathcal{M}\Big\} \,. \tag{7}$$

This is the $\gamma_s(\cdot)$ block in Fig. 2; it captures whether membership evidence is localized, monotone, dispersed, or unstable across decoding policies.

Response behavior. Pairwise similarity captures whether repeated answers agree in content, but membership evidence can also appear in the form of the answers themselves. For example, members and non-members may produce responses with similar semantic agreement but different verbosity, degeneracy, or diversity. For each temperature t, DSA therefore computes response-level statistics from $\{y^{(\ell)}_{t,k}(x,q)\}_{k=1}^K$, including answer length, unique-response ratio, empty-response rate, and character-level entropy. We denote this per-temperature statistic vector by

$$\boldsymbol{\rho}^{(\ell)}_t(x) = \rho\Big(\{y^{(\ell)}_{t,k}(x,q)\}_{k=1}^K\Big) \,, \tag{8}$$

where $\rho(\cdot)$ computes the response-level statistics listed above. Let $\Gamma_r(\cdot)$ denote the response-summary operator that applies temperature summaries component-wise to these statistic vectors. The response-behavior block is then

$$\mathbf{r}^{(\ell)}(x) = \Gamma_r\left(\left(\boldsymbol{\rho}_t^{(\ell)}(x)\right)_{t\in\mathcal{T}}\right). \tag{9}$$

This is the r block in Fig. 2; it summarizes how response form and diversity vary across the decoding surface.

Combining the three evidence families gives

$$\mathbf{f}_{\mathrm{DSA}}^{(\ell)}(x) = \left[\{s_{m,t}^{(\ell)}(x)\}_{m,t}, \boldsymbol{\gamma}_s^{(\ell)}(x), \mathbf{r}^{(\ell)}(x)\right]. \tag{10}$$

The first term is the local evidence available to fixed-policy attacks; the second and third terms encode the geometry and response behavior of the observed surface.

3.4 Calibration and Held-Out Scoring

DSA uses membership labels only through a calibration split $\mathcal{C}$ containing known members and non-members. Let $\mathcal{X}_{\mathcal{C}}$ denote the calibration candidates and $\boldsymbol{\eta}_{\mathcal{C}}$ their image-level membership labels. The response-surface feature extractor Φ_ℓ is label-free; membership labels are used only to fit the final preprocessing and scoring map. All fitted quantities are frozen before evaluation on the held-out split $\mathcal{H}$. Let $g_c(\cdot)$ denote the preprocessing map fitted exclusively on $\mathcal{C}$. It removes zero-variance features, replaces missing values with calibration-set medians, and standardizes every remaining feature using calibration-set means and standard deviations. A class-balanced logistic regression then maps the processed DSA representation to a membership score:

$$p_{\mathrm{DSA}}^{(\ell)}(x) = \sigma\left(b + \boldsymbol{\omega}^\top g_c\left(\mathbf{f}_{\mathrm{DSA}}^{(\ell)}(x)\right)\right), \tag{11}$$

where b and $\boldsymbol{\omega}$ are estimated using $\mathcal{C}$ only. Because logistic regression learns signed feature weights, Full DSA does not require a separate direction-selection step for each individual feature. The sigmoid output is used as a membership ranking score for AUC and operating-point evaluation, rather than being interpreted as a formally calibrated posterior probability.

The logistic head is intentionally low-capacity: its role is to combine measured response-surface evidence rather than to obtain attack power from classifier expressivity. Scalar reference attacks require a separate direction-orientation procedure because each of them exposes only one predefined score; their exact calibration rules are provided in Appendix B. No held-out membership label is used for imputation, normalization, model fitting, feature selection, or hyperparameter selection.

3.5 Query Budget and Complexity

For each candidate image–question pair and output budget ℓ, DSA makes $K|\mathcal{T}|$ black-box generation queries. Pairwise consistency requires

$$|\mathcal{M}|\,|\mathcal{T}|\,\frac{K(K-1)}{2} \tag{12}$$

text-similarity evaluations per candidate, computed offline from generated strings. The dominant cost is therefore model querying and text-similarity computation; the calibration model is inexpensive. This separation is important for auditing: the expensive part measures the response surface, while the downstream attack head remains simple and reproducible.

4 Experiments

4.1 Protocol

We evaluate DSA as a calibrated black-box membership audit for fine-tuned medical VLMs. The main benchmark is PathVQA (He et al., 2020); we audit Qwen (Bai et al., 2025) and LLaVA (Liu et al., 2023) at maximum generation lengths $\ell \in \{32, 64, 128, 256\}$. Each model–length setting contains 300 member images from the fine-tuning set and 300 held-out non-member images. The attacker

Algorithm 1 Decoding-Surface Attack for output budget ℓ

Require: Target VLM M_θ, temperature set $\mathcal{T}$, output budget ℓ, sample count K, similarity metrics $\mathcal{M}$, calibration set $\mathcal{C}$, held-out set $\mathcal{H}$.
Ensure: Held-out membership scores $p^{(\ell)}_{\mathrm{DSA}}(x)$.

function SURFACEFEATURES(x, q)
  for each temperature $t \in \mathcal{T}$ **do**
    for $k = 1$ to K **do**
      Query M_θ to obtain $y^{(\ell)}_{t,k}(x, q)$.
    end for
    Compute response-behavior statistics from $\{y^{(\ell)}_{t,k}(x, q)\}_{k=1}^{K}$.
    for each metric $m \in \mathcal{M}$ **do**
      Compute $s^{(\ell)}_{m,t}(x)$ using Eq. 5.
    end for
  end for
  Summarize consistency with Γ_s and response behavior with Γ_r.
  return $\mathbf{f}^{(\ell)}_{\mathrm{DSA}}(x)$ using Eq. 10.
end function
Construct calibration features $\{\mathbf{f}^{(\ell)}_{\mathrm{DSA}}(x) : (x, \eta) \in \mathcal{C}\}$, where $\eta \in \{0, 1\}$ is membership label.
Fit median imputation, zero-variance filtering, feature standardization, and class-balanced logistic regression on $\mathcal{C}$.
for each held-out candidate $x \in \mathcal{H}$ **do**
  Compute $\mathbf{f}^{(\ell)}_{\mathrm{DSA}}(x)$ using SURFACEFEATURES(x, q).
  Apply the frozen calibrator to obtain $p^{(\ell)}_{\mathrm{DSA}}(x)$.
end for
Report AUC and low-FPR operating points using held-out labels only.

Table 1: Stability of calibration-selected single-score attacks over repeated splits. Oracle gap quantifies residual test-time selection advantage.

Model	Tokens	Most selected	Selection rate	Validated AUC	Oracle gap
Qwen	32	ROUGE-2, $T = 2$	56.7%	0.710	+0.038
Qwen	64	ROUGE-2, $T = 0.1$	100.0%	0.686	+0.000
Qwen	128	ROUGE-2, $T = 0.1$	96.7%	0.643	+0.004
Qwen	256	ROUGE-2, $T = 0.1$	53.3%	0.613	+0.023
LLaVA	32	ROUGE-2, $T = 0.8$	93.3%	0.619	+0.013
LLaVA	64	ROUGE-2, $T = 0.1$	80.0%	0.559	+0.042
LLaVA	128	ROUGE-2, $T = 0.1$	100.0%	0.616	+0.007
LLaVA	256	ROUGE-2, $T = 0.4$	73.3%	0.601	+0.025

observes generated text only, sampled across temperatures $\mathcal{T} = \{0.1, 0.2, 0.4, 0.8, 1.0, 1.5, 2.0\}$ for Qwen and $\mathcal{T} = \{0.1, 0.4, 0.8, 1.0, 1.5, 2.0\}$ for LLAVA.

All AUCs are means over 30 identical stratified calibration/test splits, with 65% of records used for calibration and 35% held out. Calibration fixes score direction, imputation, normalization, hyperparameters, and logistic weights; held-out labels are used only after the scoring rule is frozen. This matters because raw similarity scores do not have a universal direction: depending on the decoding policy, either higher or lower self-consistency can indicate membership. Our experiments ask whether decoding policy is a privacy variable, whether surface modeling improves attack strength over fixed-policy and partial-surface references, which DSA components carry signal, whether the signal reflects fine-tuning membership rather than dataset artifacts, and whether the audit can guide lower-risk deployment choices.

4.2 DECODING POLICY IS A PRIVACY VARIABLE

A fixed-policy MIA is reliable only if the chosen policy is representative of the target model's broader generation behavior. Tab. 1 shows that this assumption fails in our setting. The calibration-selected single score changes across model families and output lengths: QWEN at 32 tokens most often selects ROUGE-2 at $T = 2.0$, whereas QWEN at 64 and 128 tokens selects ROUGE-2 at $T = 0.1$ almost always; LLAVA selects $T = 0.8$ at 32 tokens but $T = 0.4$ at 256 tokens. These are not small

Table 2: Comparison with recent VLM membership inference attacks on PathVQA. Values are held-out AUC (mean ± standard deviation) over 30 identical calibration/test splits. Hyperparameters and score directions are selected on calibration data only.

Method	Qwen				LLaVA				Macro
	32	64	128	256	32	64	128	256	Avg.
Hu Image-only	0.668 ± 0.022	0.598 ± 0.028	0.526 ± 0.029	0.549 ± 0.030	0.491 ± 0.029	0.525 ± 0.041	0.573 ± 0.029	0.621 ± 0.029	0.569 ± 0.030
MaxRényi-K%	0.854 ± 0.020	0.681 ± 0.035	0.655 ± 0.027	0.646 ± 0.029	0.514 ± 0.036	0.519 ± 0.022	0.563 ± 0.038	0.577 ± 0.049	0.626 ± 0.032
ICIMIA (text)	0.710 ± 0.027	0.666 ± 0.030	0.660 ± 0.031	0.647 ± 0.029	0.590 ± 0.035	0.605 ± 0.039	0.586 ± 0.039	0.573 ± 0.031	0.630 ± 0.033
CSA-MIA	0.536 ± 0.035	0.620 ± 0.024	0.616 ± 0.025	0.625 ± 0.028	0.557 ± 0.032	0.543 ± 0.025	0.626 ± 0.025	0.611 ± 0.026	0.592 ± 0.028
DSA (ours)	$\mathbf{0.915 \pm 0.015}$	$\mathbf{0.883 \pm 0.013}$	$\mathbf{0.850 \pm 0.022}$	$\mathbf{0.850 \pm 0.018}$	$\mathbf{0.706 \pm 0.029}$	$\mathbf{0.742 \pm 0.031}$	$\mathbf{0.754 \pm 0.032}$	$\mathbf{0.785 \pm 0.017}$	$\mathbf{0.811 \pm 0.022}$

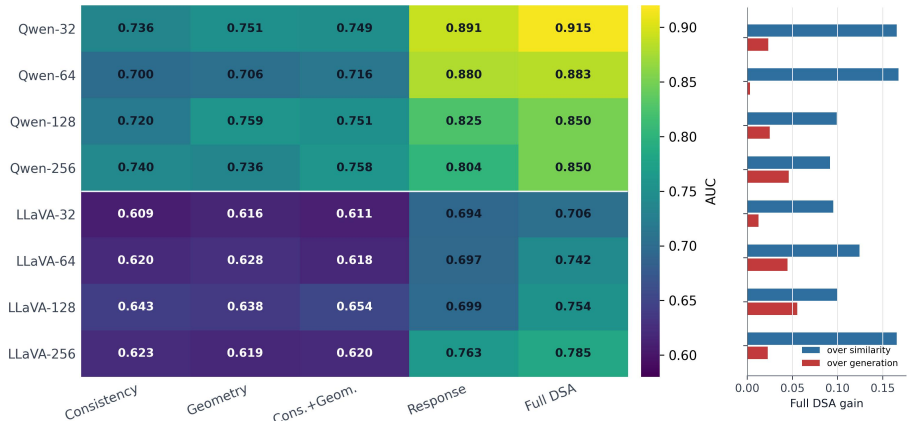


Figure 3: **DSA ablation.** Full DSA consistently improves over consistency, geometry, and response-behavior components across models and output lengths.

formatting choices. They change which part of the model's generation behavior is audited. The same table also shows why calibration cannot be replaced by a conventional "higher similarity means member" rule. Among 104 raw single-score conditions, 31 reverse direction before calibration; for LLaVA, this occurs in 25 of 48 conditions. Thus, fixed-policy attacks face two coupled risks: selecting an uninformative decoding point and interpreting the membership direction incorrectly. This motivates the response-surface view introduced in Section 3: decoding policy should be treated as part of the attack surface, not as a nuisance hyperparameter fixed before evaluation.

4.3 Main Attack Strength on PathVQA

Tab. 2 compares DSA with recent VLM membership attacks on the same PathVQA records and model–length settings: Hu Image-only (Hu et al., 2025), MaxRényi-$K\%$ from VL-MIA (Li et al., 2024), ICIMIA (Wu et al., 2025), and CSA-MIA (Li et al., 2026). DSA is strongest in every setting, with 0.811 macro AUC versus 0.630 for the best reproduced baseline, ICIMIA. Access assumptions differ: Hu Image-only, ICIMIA, and CSA-MIA use fewer black-box generations, whereas MaxRényi-$K\%$ requires full next-token distributions. We therefore keep native access patterns and interpret the comparison as evidence that fixed-view VLM MIAs leave membership signal unused when a richer response surface is observable. A query-controlled analysis further shows that DSA retains its advantage at the evaluated budgets of $Q \in \{1, 2, 5\}$, improving macro AUC by 0.033–0.184 over CSA-MIA, ICIMIA, and Hu Image-only (Appendix C.1).

Tab. 3 gives the main same-budget comparison. Mean single reflects typical fixed-policy risk by averaging individual metric–temperature AUCs. Avg-all and Validated-best are deployable calibrated references: Avg-all averages oriented single scores, while Validated-best selects one score on calibration data. Oracle-single is non-deployable because it selects the best individual score on the held-out split; it serves only as an optimistic single-score reference. DSA is strongest in all eight model–length settings, reaching 0.811 macro AUC. It improves over Multi-temperature LR by 0.126 AUC and over Oracle-single by 0.161 AUC, showing that the gain is not merely better selection among scalar scores. On Qwen, DSA reaches 0.915 AUC at 32 tokens and remains between 0.850 and 0.883 at longer budgets. On LLaVA, it rises from 0.706 at 32 tokens to 0.785 at 256 tokens, indicating that longer responses expose more usable behavioral variation for this model.

Table 3: Strongest-attack comparison over 30 matched splits. Entries are held-out AUC mean ± standard deviation; Oracle single is test-selected and macro averages span eight settings.

Model	Tokens	Mean single	Avg-all	Validated-best	Oracle-single	Multi-temp. LR	DSA (Ours)
Qwen	32	0.659 ± 0.017	0.712 ± 0.025	0.710 ± 0.022	0.747 ± 0.026	0.749 ± 0.028	**0.915** ± 0.015
Qwen	64	0.564 ± 0.014	0.624 ± 0.028	0.686 ± 0.029	0.686 ± 0.029	0.716 ± 0.029	**0.883** ± 0.013
Qwen	128	0.548 ± 0.015	0.640 ± 0.038	0.643 ± 0.030	0.648 ± 0.027	0.751 ± 0.024	**0.850** ± 0.022
Qwen	256	0.561 ± 0.014	0.690 ± 0.040	0.613 ± 0.030	0.636 ± 0.026	0.758 ± 0.024	**0.850** ± 0.018
LLaVA	32	0.531 ± 0.011	0.590 ± 0.033	0.619 ± 0.034	0.632 ± 0.024	0.611 ± 0.029	**0.706** ± 0.029
LLaVA	64	0.519 ± 0.012	0.574 ± 0.045	0.559 ± 0.045	0.601 ± 0.029	0.618 ± 0.030	**0.742** ± 0.031
LLaVA	128	0.531 ± 0.007	0.586 ± 0.026	0.616 ± 0.037	0.623 ± 0.032	0.654 ± 0.032	**0.754** ± 0.032
LLaVA	256	0.535 ± 0.010	0.584 ± 0.030	0.601 ± 0.030	0.625 ± 0.023	0.620 ± 0.027	**0.785** ± 0.017
Macro Avg.	–	0.556 ± 0.012	0.625 ± 0.033	0.631 ± 0.032	0.650 ± 0.027	0.685 ± 0.028	**0.811** ± 0.022

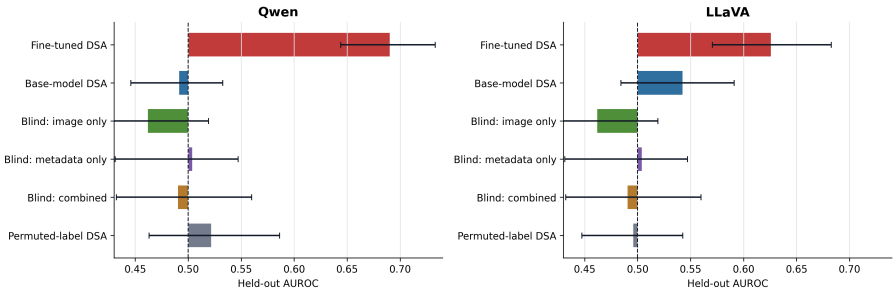


Figure 4: **Same-pool validity controls on PATHVQA.** At 32 tokens and $Q = 30$, all controls share candidates and data splits but vary the source of membership evidence.

4.4 COMPONENT ABLATION

The method defines three evidence families: within-policy consistency, cross-policy geometry, and response behavior. Fig. 3 evaluates their complementarity. Consistency-only, geometry-only, and their combination achieve macro AUCs of 0.674, 0.682, and 0.685, respectively. These results are close to Multi-temperature LR in Tab. 3, since both use similarity evidence over the temperature grid. Response behavior is stronger at 0.782, capturing answer length, diversity, empty-response rate, and entropy, but omitting semantic agreement among repeated outputs. Full DSA combines these complementary signals, reaches 0.811 macro AUC, and outperforms every non-full variant across all eight settings. The largest gains occur for LLAVA at 128 tokens (+0.055) and QWEN at 256 tokens (+0.046). Full results appear in Appendix D.

4.5 EXTERNAL MEDICAL VQA COMPARISON

The main PATHVQA benchmark shows that DSA improves over fixed-policy and partial-surface attacks across the full model–length grid. Tab. 4 tests whether this pattern transfers to PMC-VQA (Zhang et al., 2023), using the per-sample VQA question, a 32-token cap, $Q = 30$ sampled responses, and identical stratified calibration/test splits. Full DSA reaches 0.950 AUC, improving over Multi-temp LR by 0.034 and over Oracle-single by 0.054; Appendix G shows that these gains hold on all 30 paired splits. We treat PMC-VQA as external validation because PATHVQA remains the only benchmark evaluated across both model families and all output lengths.

4.6 VALIDITY CONTROLS

A high AUC may reflect dataset artifacts rather than fine-tuning membership. We therefore conduct a same-pool PATHVQA audit at 32 tokens and $Q = 30$, assigning members and non-members from a common source pool before training separate targets. All controls share candidates and splits but vary the signal source: the adapted target, base model, blind image/metadata covariates, or permuted labels. Fine-tuned DSA achieves 0.690 AUC for QWEN and 0.626 for LLAVA, whereas all controls remain near chance. These values are below the main 32-token results (0.915 and 0.706), consistent with reduced source-pool shortcuts, although separate target training and fewer QWEN queries also contribute to the gap. Thus, detectable membership signal persists under the stricter same-source protocol. Full results appear in Appendix F.

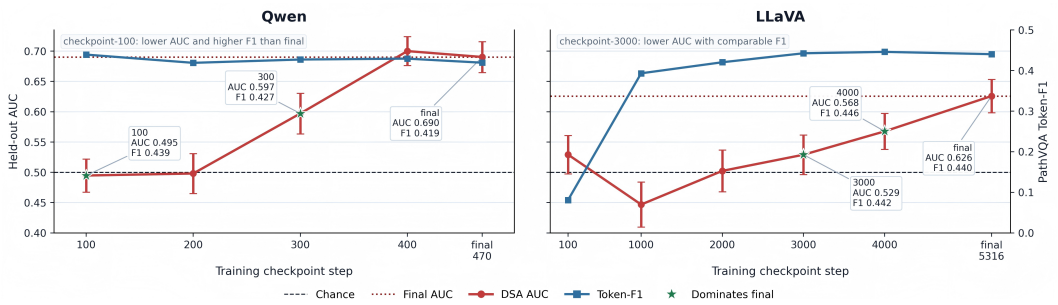


Figure 5: DSA-guided checkpoint selection.

Table 4: **PMC-VQA validation.** All attacks use per-sample questions, a 32-token cap, 30 sampled responses, and identical calibration/test splits; entries report held-out AUC mean ± standard deviation.

Dataset	Mean single	Avg-all	Validated-best	Oracle-single	Multi-temp LR	Full DSA
PMC-VQA	0.761 ± 0.020	0.873 ± 0.022	0.882 ± 0.021	0.896 ± 0.019	0.917 ± 0.020	**0.950 ± 0.012**

Table 5: Surface-access restriction. The audit limits the number of exposed temperature probes and reports risk reduction relative to full surface access within each model.

Model	Temp. probes	Surface exposed	AUC	AUC Drop
Qwen	1	14%	0.795 ± 0.036	0.060
Qwen	2	29%	0.824 ± 0.032	0.031
Qwen	3	43%	0.836 ± 0.026	0.019
Qwen	4	57%	0.843 ± 0.027	0.012
Qwen	Full	100%	**0.855** ± 0.034	0.000
LLaVA	1	17%	0.706 ± 0.031	0.035
LLaVA	2	33%	0.726 ± 0.048	0.014
LLaVA	3	50%	0.730 ± 0.045	0.011
LLaVA	4	67%	0.737 ± 0.041	0.004
LLaVA	Full	100%	**0.741** ± 0.040	0.000

4.7 DSA-Guided Risk Control

All attack-strength results above use the final fine-tuned checkpoint; the following analyses use DSA only to guide lower-risk deployment choices. Because the two controls use different audit settings, their absolute AUCs are not directly comparable. First, surface-access restriction limits the decoding temperatures exposed by the API. Tab. 5 shows that one-temperature access reduces AUC from 0.855 to 0.795 for Qwen and from 0.741 to 0.706 for LLaVA, although leakage increases as more temperatures become available. Thus, restricting access narrows but does not eliminate the observable privacy channel. Second, DSA-guided checkpoint selection jointly audits leakage and Token-F1 during training. As shown in Fig. 5, checkpoint-100 reduces Qwen AUC from 0.690 to 0.495 while improving Token-F1 from 0.419 to 0.439; checkpoint-3000 reduces LLaVA AUC from 0.626 to 0.529 while preserving Token-F1 (0.442 versus 0.440). This provides a practical audit-driven selection policy, not a formal privacy guarantee. Exact results appear in Appendix H.

5 Conclusion

We study membership inference for fine-tuned medical VLMs through the lens of decoding behavior. Our results show that leakage is not fully characterized by a single temperature, output length, similarity metric, or fixed score direction; instead, it varies across the decoding surface exposed by black-box generation. We propose DSA, a calibrated black-box attack that fuses within-policy consistency, cross-policy geometry, and response-behavior statistics into a per-image membership score. Across PathVQA and an external PMC-VQA validation, DSA improves over fixed-policy, partial-surface, and recent VLM MIA baselines, while same-pool controls support that the measured signal reflects fine-tuning membership rather than dataset artifacts. Finally, we show that the same audit can guide lightweight risk controls such as surface-access restriction and checkpoint selection. These findings suggest that privacy evaluation for generative medical VLMs should audit the decoding surface, not only a single inference policy.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, 2022.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11441.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. Membership inference attacks against vision-language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1589–1608, Seattle, WA, aug 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/hu-yuke.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://papers.nips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Jiaqing Li, Yajuan Lu, Xiaochuan Shi, Gang Wu, ZhongYuan Wang, and Chao Liang. Single-sample black-box membership inference attack against vision-language models via cross-modal semantic alignment. *arXiv preprint arXiv:2605.17341*, 2026. URL https://arxiv.org/abs/2605.17341.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3131. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b2c892312af07f8a77afbeed188391f4-Abstract-Conference.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fed9263d7f1086e735904ff690527a53-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023. doi: 10.1038/s41586-023-05881-4. URL https://www.nature.com/articles/s41586-023-05881-4.

Khanh Nguyen, Raouf Kerkouche, Mario Fritz, and Dimosthenis Karatzas. DocMIA: Document-level membership inference attacks against DocVQA models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/231a8daf4f64935d0c7c6586f290b24f-Abstract-Conference.html.

Ashwinee Panda, Xinyu Tang, Christopher Choquette-Choo, Milad Nasr, and Prateek Mittal. Privacy auditing of large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e76814a0012a46ade4f997591e86f972-Abstract-Conference.html.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 2024. doi: 10.1056/AIoa2300138. URL https://research.google/pubs/towards-generalist-biomedical-ai/.

Zongyu Wu, Minhua Lin, Zhiwei Zhang, Fali Wang, Xianren Zhang, Xiang Zhang, and Suhang Wang. Image corruption-inspired membership inference attacks against large vision-language models. *arXiv preprint arXiv:2506.12340*, 2025. URL https://arxiv.org/abs/2506.12340. Accepted to EACL 2026.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, 2018.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30:3129–3141, 2024. doi: 10.1038/s41591-024-03185-2. URL https://www.nature.com/articles/s41591-024-03185-2.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. URL https://arxiv.org/abs/2305.10415.

A Additional Experimental Details

Split construction. For each model–length setting, we form a balanced candidate pool with 300 member and 300 non-member images. Each of the 30 repetitions uses a stratified 65/35 calibration/test split, so calibration contains 195 members and 195 non-members, and the held-out test split contains 105 members and 105 non-members. The same split indices are reused for all attacks within a setting. This ensures that performance differences are attributable to the scoring rule rather than to easier or harder test partitions.

Target models and fine-tuning. We instantiate QWEN with Qwen3-VL-4B-Instruct and LLaVA with LLaVA-v1.6-Mistral-7B. Both targets are adapted using LoRA with rank 8 over all linear modules. QWEN is trained for 10 epochs with learning rate 10^{-4}, per-device batch size 12, gradient accumulation 8, FP16 precision, and a context length of 2048. LLaVA is trained for 6 epochs with learning rate 10^{-5}, per-device batch size 1, gradient accumulation 5, BF16 precision, and a context length of 4096.

What calibration is allowed to do. Calibration may use membership labels only for the calibration records. It fixes score orientation, missing-value imputation, feature normalization, hyperparameter selection, and the weights of any logistic calibrator. Once these choices are fit, they are frozen. Held-out labels are used only to compute final AUC and operating-point statistics. This design mirrors a realistic audit setting in which an evaluator has a labeled reference set but must score new candidate records without using their labels.

Macro averaging and uncertainty. Per-setting results report the mean and standard deviation over the 30 held-out splits. Macro averages are unweighted over the eight PATHVQA model–length settings. When a table reports "Macro avg." with a $\pm$ term, both the mean AUC and the standard-deviation term are separately averaged across settings; the latter summarizes typical split-to-split variability rather than the standard error of the macro mean. The repeated splits quantify sensitivity to calibration/test partitioning but do not capture variability across independently fine-tuned target-model seeds.

Query accounting. In the main attack-strength benchmark, DSA uses 10 stochastic generations at each of 7 temperatures for QWEN and 5 stochastic generations at each of 6 temperatures for LLaVA. The output-length budgets $\ell \in \{32, 64, 128, 256\}$ are evaluated separately. Surface-access restriction is a separate risk-control study that varies the number of exposed temperature probes; its results are interpreted within each model rather than compared as a query-matched benchmark.

B Exact Baseline Definitions

This section defines the calibration procedures used by the scalar and partial-surface reference attacks. These procedures are separate from Full DSA: scalar references require explicit score-direction selection, whereas Full DSA learns feature directions through the signed weights of its logistic-regression head. Let $j = (m, t) \in \mathcal{J} = \mathcal{M} \times \mathcal{T}$ index one similarity metric m and one temperature t. Therefore, $|\mathcal{J}| = 14$ for QWEN and $|\mathcal{J}| = 12$ for LLaVA. Let $\mathcal{X}_\mathcal{C}$ and $\boldsymbol{\eta}_\mathcal{C}$ denote the calibration candidates and their membership labels, and let $\mathcal{X}_\mathcal{H}$ and $\boldsymbol{\eta}_\mathcal{H}$ denote the corresponding held-out candidates and labels.

Scalar-score orientation and standardization. For each scalar score s_j, the calibration split determines whether larger or smaller values indicate membership:

$$a_j^\mathcal{C} = \mathrm{AUC}\left(\boldsymbol{\eta}_\mathcal{C}, s_j(\mathcal{X}_\mathcal{C})\right), \qquad d_j = \begin{cases} +1, & a_j^\mathcal{C} \geq 0.5, \\ -1, & a_j^\mathcal{C} < 0.5. \end{cases} \tag{13}$$

After calibration-median imputation, the oriented and standardized score is

$$z_j(x) = d_j \frac{s_j(x) - \mu_j}{\sigma_j}, \tag{14}$$

where μ_j and σ_j are computed on $\mathcal{X}_\mathcal{C}$ only. The imputation value, direction, mean, and standard deviation are frozen before held-out evaluation.

Mean single. Mean single reports the average held-out AUC of all individual scalar scores:

$$A_{\text{mean}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \text{AUC}\left(\boldsymbol{\eta}_{\mathcal{H}}, z_j(\mathcal{X}_{\mathcal{H}})\right). \tag{15}$$

Because it averages multiple test AUC values rather than producing one membership score per image, Mean single is a descriptive summary of typical fixed-policy leakage and is not itself a deployable attack.

Validated-best. Validated-best selects one scalar score using calibration labels only:

$$j^* = \arg\max_{j \in \mathcal{J}} \text{AUC}\left(\boldsymbol{\eta}_{\mathcal{C}}, z_j(\mathcal{X}_{\mathcal{C}})\right). \tag{16}$$

The selected and frozen score $z_{j^*}(x)$ is then applied to the held-out candidates. This is a deployable calibration-selected fixed-policy attack.

Oracle-single. Oracle-single uses held-out labels to select both the direction and the best individual scalar score:

$$A_{\text{oracle}} = \max_{j \in \mathcal{J}} \max\left\{A_j^{\mathcal{H}}, 1 - A_j^{\mathcal{H}}\right\}, \qquad A_j^{\mathcal{H}} = \text{AUC}\left(\boldsymbol{\eta}_{\mathcal{H}}, s_j(\mathcal{X}_{\mathcal{H}})\right). \tag{17}$$

Oracle-single is deliberately non-deployable and is used only as an optimistic upper comparator for individual scalar scores. It is not an upper bound on attacks that fuse multiple measurements.

Avg-all. Avg-all produces one membership score for each candidate by averaging all oriented and standardized scalar scores:

$$u_{\text{all}}(x) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} z_j(x). \tag{18}$$

A single held-out AUC is then computed from $u_{\text{all}}(\mathcal{X}_{\mathcal{H}})$. Thus, Mean single averages several held-out AUCs, whereas Avg-all first averages per-image scores and then computes one AUC.

Learned partial-surface attacks. All learned attacks use the same calibration protocol as Full DSA: zero-variance filtering, median imputation, feature standardization, and class-balanced logistic regression are fitted using $\mathcal{C}$ only. For a baseline feature vector $\mathbf{f}_B(x)$, the resulting score is

$$p_B(x) = \sigma\left(b_B + \mathbf{w}_B^\top g_{B,\mathcal{C}}\left(\mathbf{f}_B(x)\right)\right). \tag{19}$$

Multi-temperature LR uses

$$\mathbf{f}_{\text{sim}}(x) = \left[\{s_{m,t}(x)\}_{m,t}; \mathbf{h}(x)\right], \tag{20}$$

where $\mathbf{h}(x)$ contains, for each similarity metric, the cross-temperature mean, standard deviation, minimum, maximum, range, least-squares slope, and high-minus-low temperature difference. This produces 28 features for QWEN and 26 features for LLAVA.

Full DSA additionally includes response-count, empty-response ratio, unique-response ratio, word-length moments, character length, and character-level entropy at every temperature, together with the cross-temperature mean, standard deviation, and range of each response statistic. Unlike the scalar baselines in Eqs. 13–14, Full DSA does not explicitly orient individual features; their directions are learned through the signed logistic-regression weights in Eq. 11.

C EXTERNAL VLM MIA ADAPTATIONS

All external attacks use the same 600 records per model–length setting and the same 30 stratified 65/35 calibration/test splits as DSA. For every scalar score, calibration data alone determine whether higher or lower values indicate membership.

Hu Image-only. For K stochastic descriptions $\{y_a(x)\}_{a=1}^K$, the score is the mean pairwise ROUGE-2 F-measure. We use $T = 0.4$, the closest available decoding point to the source setting $T = 0.5$. The native budget is $K = 10$ for QWEN and $K = 5$ for LLAVA.

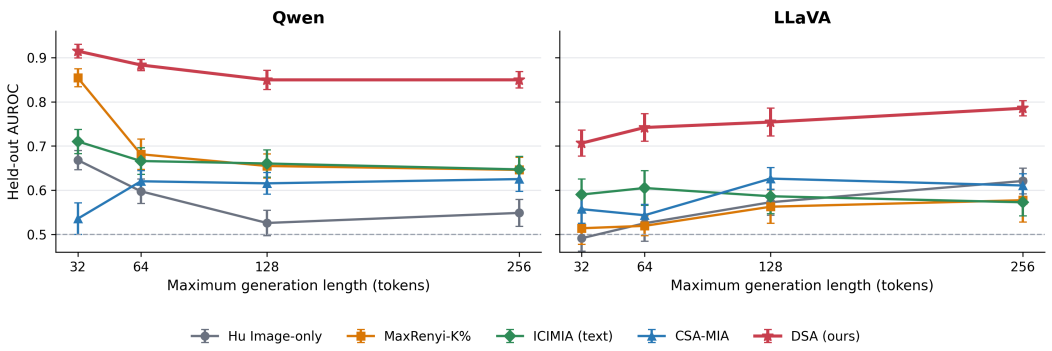


Figure 6: **Comparison with recent VLM membership attacks.** Points report held-out AUC over the same 30 calibration/test splits and error bars show one standard deviation.

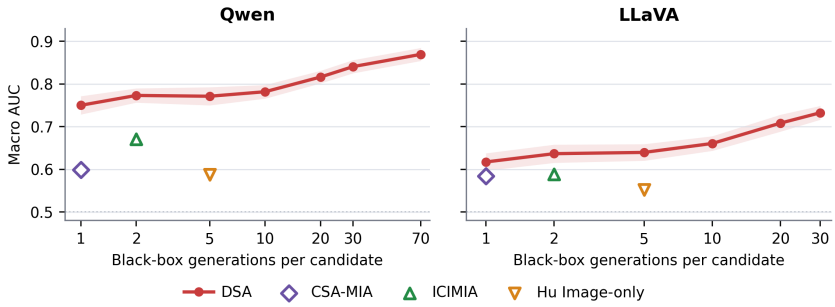


Figure 7: **Query-matched attack efficiency.** DSA remains stronger than external black-box VLM MIAs when restricted to their generation budgets. Hu Image-only uses $K = 5$ for both model families; shading denotes one standard deviation over 30 matched splits.

MaxRényi-$K\%$. We teacher-force the image, instruction, and its first $T = 0.1$ description through the target. For description position i with full-vocabulary distribution p_i, the method aggregates high-entropy token positions under candidate Rényi orders. Calibration selects one entropy order, one token percentage, and the score direction. This is the only comparator requiring full next-token distributions.

ICIMIA and CSA-MIA. ICIMIA compares descriptions of the original image and a Gaussian-blurred version using a text encoder. CSA-MIA measures CLIP alignment between the image and a generated description. Both are evaluated with calibration-only score orientation under the same splits as DSA.

C.1 Query-Matched Attack Efficiency

Because the main comparison retains each baseline's native access pattern, query count may confound attack strength. We therefore restrict DSA to the same black-box generation budget as each text-output baseline. For every model–length setting, a deterministic nested schedule selects exactly Q responses across the available temperatures, and DSA is calibrated using only the evidence available at that budget. We compare against CSA-MIA (Li et al., 2026), ICIMIA (Wu et al., 2025), and Hu Image-only (Hu et al., 2025) at $Q = 1$, $Q = 2$, and $Q = 5$, respectively; Hu Image-only uses $K = 5$ for both model families. All methods use the same PathVQA records and 30 matched stratified calibration/test splits. Model-level macro AUC is computed by averaging the four output-length AUCs within each split.

Fig. 7 shows that DSA's advantage persists after controlling for query count. On Qwen, DSA improves over CSA-MIA, ICIMIA, and Hu Image-only by 0.151, 0.102, and 0.184 AUC, respectively, and wins all 30 matched splits. On LLaVA, the corresponding gains are 0.033, 0.048, and 0.087, with

Table 6: **Component ablation of DSA on PathVQA.** Consistency uses within-policy ROUGE-2/MPNet agreement; Geometry uses cross-policy summaries of the consistency surface; Response behavior uses generation statistics such as length, diversity, empty-response rate, and entropy. Full gain is measured over the strongest non-full component or partial-surface reference in the same row.

Model	Tokens	Consistency	Geometry	Cons.+Geom.	Response	Full DSA	Full gain
Qwen	32	0.736	0.751	0.749	0.891	**0.915**	+0.023
Qwen	64	0.700	0.706	0.716	0.880	**0.883**	+0.003
Qwen	128	0.720	0.759	0.751	0.825	**0.850**	+0.025
Qwen	256	0.740	0.736	0.758	0.804	**0.850**	+0.046
LLaVA	32	0.609	0.616	0.611	0.694	**0.706**	+0.013
LLaVA	64	0.620	0.628	0.618	0.697	**0.742**	+0.045
LLaVA	128	0.643	0.638	0.654	0.699	**0.754**	+0.055
LLaVA	256	0.623	0.619	0.620	0.763	**0.785**	+0.023
Macro avg.	–	0.674	0.682	0.685	0.782	**0.811**	+0.029

DSA winning 27/30, 28/30, and 30/30 splits. At $Q = 1$, only single-response behavior is observable; pairwise consistency and cross-policy structure become available as the budget grows. The results therefore show that DSA's improvement cannot be explained solely by additional black-box queries.

D Component Ablation Details

Tab. 6 decomposes DSA into the three evidence families introduced in Section 3. *Consistency* uses only within-policy agreement among repeated generations, measured with lexical and semantic similarity. *Geometry* summarizes how those consistency scores vary across decoding policies, such as their mean, dispersion, range, and slope over temperature. *Response behavior* uses statistics of the generated answers themselves, including length, diversity, empty-response rate, and entropy. The full method concatenates all three families before calibration.

The ablation answers whether DSA is merely a larger logistic-regression feature set or whether the response-surface view contributes complementary evidence. The result supports the latter. On PathVQA, consistency and geometry alone give macro AUCs of 0.674 and 0.682, and their combination reaches 0.685. Response behavior is stronger, with 0.782 macro AUC, showing that fine-tuning membership affects not only semantic agreement but also the form and variability of generated text. Full DSA reaches 0.811 macro AUC and is best in every model–length setting. The average full gain over the strongest non-full component is +0.029 AUC, with particularly large gains for LLaVA at 64 and 128 tokens (+0.045 and +0.055). Thus the complete attack is not driven by a single evidence family; it benefits from fusing consistency, surface geometry, and response behavior.

E Decoding-Policy Selection Heatmaps

Fig. 8 expands the main-text selection summary by showing the full calibration-selection heatmap for Validated-best MIA. Each row is one model–length setting, and each column is one candidate single score defined by a similarity metric and decoding temperature. A cell reports the percentage of the 30 calibration splits in which that candidate is selected as the best single-score attack. Gray cells denote temperatures that were not available for a model.

The heatmap illustrates why fixed-policy auditing is brittle. The selected policy is highly configuration-dependent: Qwen at 64 tokens selects ROUGE-2 at $T = 0.1$ in 100% of splits, whereas Qwen at 32 tokens splits between ROUGE-2 at $T = 2.0$ and MPNet at $T = 0.1$. LLaVA follows a different pattern: the 32-token setting selects ROUGE-2 at $T = 0.8$ in most splits, while the 256-token setting shifts toward ROUGE-2 at $T = 0.4$. These changes occur under the same calibration protocol and candidate pool. Therefore, a single preselected temperature or metric would audit different parts of the response surface for different models, and may miss stronger leakage elsewhere. This motivates modeling the surface rather than committing to one decoding point.

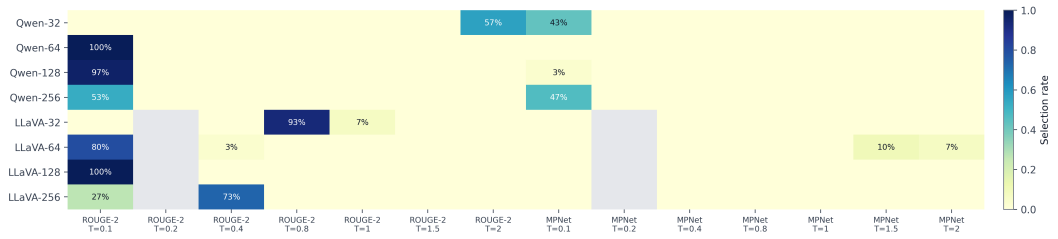


Figure 8: **Calibration selection is configuration-dependent.** Each cell reports the percentage of 30 calibration splits in which Validated-best MIA selects a given metric–temperature score; gray cells indicate unavailable decoding temperatures.

Model	Method	AUC	95% interval	TPR@1% FPR	TPR@5% FPR
Qwen	Fine-tuned DSA	**0.690 ± 0.025**	[0.644, 0.733]	0.097	0.209
Qwen	Base-model DSA	0.491 ± 0.025	[0.446, 0.533]	0.023	0.068
Qwen	Blind: image only	0.462 ± 0.035	[0.401, 0.519]	0.010	0.034
Qwen	Blind: metadata only	0.504 ± 0.031	[0.431, 0.547]	0.012	0.043
Qwen	Blind: combined	0.491 ± 0.033	[0.432, 0.560]	0.013	0.041
Qwen	Permuted-label DSA	0.522 ± 0.034	[0.463, 0.586]	0.022	0.062
LLaVA	Fine-tuned DSA	**0.626 ± 0.028**	[0.571, 0.683]	0.124	0.202
LLaVA	Base-model DSA	0.542 ± 0.031	[0.484, 0.591]	0.016	0.061
LLaVA	Blind: image only	0.462 ± 0.035	[0.401, 0.519]	0.010	0.034
LLaVA	Blind: metadata only	0.504 ± 0.031	[0.431, 0.547]	0.012	0.043
LLaVA	Blind: combined	0.491 ± 0.033	[0.432, 0.560]	0.013	0.041
LLaVA	Permuted-label DSA	0.496 ± 0.026	[0.447, 0.543]	0.020	0.054

Table 7: **Same-pool validity controls at 32 generated tokens and $Q = 30$ black-box queries.** All methods use identical calibration/test image splits. AUC is reported as mean ± standard deviation over 30 splits; 95% intervals are empirical percentile intervals; TPR@1%/5% FPR reports held-out member detection at low false-positive rates.

F Same-Pool Validity Controls

Tab. 7 reports the numerical values behind Fig. 4. The goal is to test whether high attack performance reflects fine-tuning membership rather than easier artifacts such as image-pool separability, metadata leakage, or a calibration pipeline that overfits arbitrary labels. All methods use the same balanced PathVQA same-pool candidate set, the same 32-token, $Q = 30$ query budget, and the same calibration/test image splits.

The control rows change only the source of evidence. Base-model DSA uses the corresponding pretrained/base VLM on the same candidates, so a high score there would indicate that the image pool is already separable before task adaptation. Blind image-only and metadata-only controls use no VLM generations and test whether low-level image statistics, question/answer length, group counts, or coarse metadata explain membership labels. The permuted-label control keeps the fine-tuned DSA features but randomizes membership labels before calibration, testing whether the pipeline can manufacture apparent leakage when the training relation is removed. The fine-tuned DSA row separates from all controls: for Qwen, AUC is 0.690 versus 0.491 for the base model and 0.504 for the strongest blind control; for LLaVA, AUC is 0.626 versus 0.542 and 0.504. The controls remain near chance, supporting the interpretation that DSA measures fine-tuning-induced membership signal rather than a trivial shortcut.

The table also reports operating-point statistics. The 95% interval is the empirical 2.5-97.5 percentile range of AUC across the 30 repeated calibration/test splits, and is included to show split-to-split stability. TPR@1% FPR and TPR@5% FPR measure how many true members are detected when the decision threshold is set so that only 1% or 5% of non-members are falsely flagged on the held-out split. These low-FPR metrics are stricter than AUC and are relevant for privacy auditing, where a useful attack should identify members while producing few false accusations. The fine-tuned DSA rows have both higher AUC and higher low-FPR TPR than the controls.

Baseline	ΔAUC: Full DSA − baseline	DSA wins / folds
Mean single	+0.189 ± 0.019	30/30
Avg-all	+0.077 ± 0.019	30/30
Validated-best	+0.068 ± 0.016	30/30
Oracle-single	+0.054 ± 0.015	30/30
Multi-temp LR	+0.034 ± 0.013	30/30

Table 8: **Paired PMC-VQA comparison.** Differences are computed fold-by-fold under the same strong-attack protocol as Table 4. Positive values mean Full DSA has higher held-out AUC than the baseline.

Model	Checkpoint	Step	DSA AUC	Token-F1	AUC drop	Token-F1 Δ
Qwen	**checkpoint-100**	100	**0.495 ± 0.027**	0.439	+0.196	+0.020
Qwen	checkpoint-200	200	0.498 ± 0.033	0.419	+0.192	-0.000
Qwen	**checkpoint-300**	300	**0.597 ± 0.033**	0.427	+0.093	+0.008
Qwen	checkpoint-400	400	0.700 ± 0.024	0.429	-0.010	+0.010
Qwen	final	470	0.690 ± 0.025	0.419	+0.000	+0.000
LLaVA	checkpoint-100	100	0.529 ± 0.032	0.081	+0.097	-0.359
LLaVA	checkpoint-1000	1000	0.447 ± 0.037	0.393	+0.179	-0.047
LLaVA	checkpoint-2000	2000	0.502 ± 0.035	0.420	+0.123	-0.020
LLaVA	**checkpoint-3000**	3000	**0.529 ± 0.033**	0.442	+0.097	+0.002
LLaVA	**checkpoint-4000**	4000	**0.568 ± 0.030**	0.446	+0.058	+0.006
LLaVA	final	5316	0.626 ± 0.028	0.440	+0.000	+0.000

Table 9: **DSA-guided checkpoint selection.** AUC drop and Token-F1 Δ are computed against the final checkpoint of the same model. Bold checkpoints Pareto-dominate the final checkpoint by lowering DSA AUC while preserving or improving Token-F1.

G External Dataset Details

Tab. 8 gives paired split differences for the PMC-VQA validation summarized in Tab. 4. The main text reports aggregate AUCs; this table checks whether the improvement is stable on the exact same calibration/test splits. For each split, we subtract the baseline AUC from Full DSA AUC. Positive values therefore mean that DSA improves on that matched split rather than only after averaging across different partitions.

The paired results strengthen the external-validation claim. Full DSA beats every listed reference in all 30 matched splits. The mean paired gains are +0.189 over Mean single, +0.077 over Avg-all, +0.068 over Validated-best, +0.054 over Oracle-single, and +0.034 over Multi-temp LR. The last comparison is the most stringent among deployable surface baselines because Multi-temp LR already uses the similarity surface; the remaining gain shows that the full response-surface representation contributes information beyond similarity scores alone. We treat PMC-VQA as external validation rather than the primary benchmark because only PathVQA is evaluated across both model families and all four output lengths.

H Risk-Control Details

Tab. 9 gives exact checkpoint-selection values for Fig. 5. This analysis is separate from the attack-strength comparisons in the main text, which use the final fine-tuned checkpoint unless otherwise stated. Here DSA is used as an audit signal along the training trajectory: for each checkpoint, we measure held-out membership AUC and downstream PathVQA Token-F1, then compare both quantities against the final checkpoint of the same model.

The table reports two derived quantities. AUC drop is positive when a checkpoint has lower measured membership leakage than the final checkpoint. Token-F1 Δ is positive when the checkpoint preserves or improves utility. A checkpoint Pareto-dominates the final model when both conditions hold. For Qwen, checkpoint-100 reduces AUC from 0.690 to 0.495 while improving Token-F1 from 0.419 to 0.439; checkpoint-300 also reduces leakage while improving Token-F1. For LLaVA, checkpoints

3000 and 4000 both lower AUC relative to the final checkpoint while maintaining or improving Token-F1. This is not a formal privacy guarantee, but it shows that DSA can be used operationally to choose a lower-leakage deployment checkpoint when multiple checkpoints have comparable utility.