NOISE-FREE PROVABLE TRUSTWORTHY MACHINE LEARNING VIA DISTRIBUTION-SPECIFIC INDISTINGUISHABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Indistinguishability guarantees in trustworthy machine learning are usually obtained by injecting noise, most prominently through Differential Privacy (DP). We develop a noiseless alternative for data-influence problems such as poisoning, memorization, and copyright protection. Let A be the part of the training data whose influence we want to control (poisoned samples, a sensitive record, or a protected artwork), and let R be the rest. A model trained on R alone never sees A and is taken as a safe reference. Our goal is to make the model trained on $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$ statistically indistinguishable from this reference, so that the reference's trustworthiness carries over. Operationally, instead of injecting algorithmic randomness such as noise, we exploit input-wise randomness: the benign data R is modeled as a random sample from an unknown distribution, while A may be chosen adversarially after observing R. We formalize this as *Distribution-Specific Indistinguishability* (DiSI): a DiSI algorithm then transfers the reference's high-probability trustworthiness to the model trained on X. We instantiate DiSI through consistency-robust gradient operators that make the target and reference updates agree exactly, except on a failure event whose probability decays exponentially in the number of sources. Across image and text backdoor benchmarks, DiSI sharply reduces attack success while preserving far more utility than noise-based baselines; in generative-model fine-tuning, it also reduces outlier memorization, membership leakage, and resemblance to protected artworks.

1 INTRODUCTION

The rapid development of Artificial Intelligence (AI) in high-stakes domains (Kaur et al., 2022; Li et al., 2023a) has raised growing societal and security concerns. Traditionally, model accuracy and benchmark performance were treated as the primary evaluation criteria; over the last decade, substantial effort has been devoted to *simultaneously* improving additional trustworthiness properties, such as privacy (Abadi et al., 2016), memorization (Carlini et al., 2021; 2023), (adversarial) robustness (Barreno et al., 2006; Goodfellow et al., 2014) (including robustness to poisoning attacks (Rubinstein et al., 2009)), compliance (Li et al., 2025), and copyright (Vyas et al., 2023). However, the pursuit of Trustworthy Machine Learning is fundamentally constrained by a trifecta of hurdles.

(i) Semantic Definition. Many core trustworthiness concepts are semantically elusive. For example, what kinds of works deserve copyright protection remains controversial (Beckerman-Rodau, 2010; Matulionyte & Lee, 2022). Many trustworthy concepts shift with social context and domain specifics, and there is usually no universally accepted definition. Recent works have demonstrated under-specification or incompatibility issues regarding, for instance, memorization (Hartmann et al., 2023) and (knowledge) unlearning (Thudi et al., 2022).

(ii) Risk Quantification. Given intertwined, multi-dimensional trust requirements and the black-box nature of frontier machine learning models, we often lack sufficiently powerful mathematical tools to quantify the underlying trustworthiness risk. Many existing evaluation approaches are either attack-driven ("red-teaming") methods, such as membership inference (Shokri et al., 2017; Carlini et al., 2022) and differential-privacy auditing (Kong et al., 2024), or benchmark-driven evaluations (Liang et al., 2022). However, the optimality of adversarial strategies and the completeness of threat models are generally difficult to validate.

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.

(iii) Provable Guarantees. Another, arguably the most important, challenge is operationally how to convert an unsatisfactory data-processing mechanism into a trustworthy one. Failures of trustworthiness can lead to irrecoverable consequences. Yet, many existing defenses remain heuristic and lack provable guarantees, fueling a continuing cat-and-mouse dynamic (Hönig et al., 2024; Zhao et al., 2024; Shan et al., 2023; 2024; Huang et al., 2020).

1.1 Reference-Indistinguishability Methodology

A popular line of research addressing these challenges is rooted in Differential Privacy (DP) (Dwork et al., 2006b) and its variants (Xiao et al., 2025; Vyas et al., 2023). DP limits the influence of any *arbitrary* datapoint on the released outcome, and recent studies (Xiao et al., 2025; Vyas et al., 2023; Carlini et al., 2021; 2023) show that the indistinguishability it provides is a *sufficient* condition for transferring, from a safe reference, a broad range of data-influence trust properties such as robustness to poisoning (backdoors), non-memorization, and copyright protection. These works establish indistinguishability (statistical closeness) between a target model and a set of *safe* reference models. Below we interpret this indistinguishability from a probabilistic perspective (more discussion in Appendices A.1 and A.2).

Trustworthy algorithm design can be modeled as a constrained optimization: given an input domain $\mathcal{X}^*$ (e.g., training data), an output domain $\mathcal{O}^*$ (e.g., model weights), and a utility measure λ (e.g., test accuracy), a subset $\mathcal{T} \subseteq \mathcal{O}^*$ satisfies the desired trustworthiness requirement (e.g., robustness to backdoors), and the goal is an efficient algorithm that selects $o \in \mathcal{T}$ with the highest utility $\max_{o \in \mathcal{T}} \lambda(o)$.

In practice, approximating $\mathcal{T}$ is extremely challenging, let alone optimizing over it. An alternative is to rely implicitly on statistical closeness to known, safe *references*. Suppose an algorithm Alg and a reference input R satisfy the trustworthiness condition with high probability, $\Pr(\mathsf{Alg}(\mathsf{R}) \in \mathcal{T}) \geq 1 - \delta_0$. If the output distribution $\mathbb{P}_{\mathsf{Alg}(\mathsf{X})}$ of a target input X is close to $\mathbb{P}_{\mathsf{Alg}(\mathsf{R})}$, *indistinguishability* transfers this guarantee to $\mathsf{Alg}(\mathsf{X})$: in total variation (TV) distance $\mathsf{TV}(\mathbb{P}\|\mathbb{Q}) = \frac{1}{2}\int |\mathbb{P} - \mathbb{Q}|$ (Sason & Verdú, 2016), $\Pr\big(\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}\big) \geq 1 - \delta_0 - \mathsf{TV}\big(\mathbb{P}_{\mathsf{Alg}(\mathsf{R})} \| \mathbb{P}_{\mathsf{Alg}(\mathsf{X})}\big)$: a target model that is statistically close to the reference inherits its properties with high probability.

Many data-influence trust notions can be framed via such *references*. In memorization mitigation (Tirumala et al., 2022), for $\mathsf{X} = \{x_1, \ldots, x_n\}$ let $\mathcal{T}_i$ be the set of generative models that do not reproduce x_i when prompted with a segment of it. The model trained on the leave-one-out reference $\mathsf{R}_i = \mathsf{X} \setminus x_i$ never sees x_i and belongs to $\mathcal{T}_i$ with high probability once x_i is sufficiently unique; if $\mathsf{Alg}(\mathsf{X})$ is *simultaneously* close to *each* $\mathsf{Alg}(\mathsf{R}_i)$, memorization mitigation holds for every x_i.

Backdoor and poisoning defense (Wu et al., 2025) can be framed similarly. Let each x_i be the data of a distinct source, one unknown source i_0 being malicious and free to craft its data adaptively, given the benign data of the other sources, to implant a trigger. Among the leave-one-source-out references $\mathsf{R}_i = \mathsf{X} \setminus x_i$, the unknown R_{i_0} is entirely clean, so $\mathsf{Alg}(\mathsf{R}_{i_0})$ is robust to the *unseen* trigger with high probability; if $\mathsf{Alg}(\mathsf{X})$ is *simultaneously* indistinguishable from *each* $\mathsf{Alg}(\mathsf{R}_i)$, this yields backdoor robustness of $\mathsf{Alg}(\mathsf{X})$.

The above examples illustrate the general workflow of "reference indistinguishability": rather than defining $\mathcal{T}$ explicitly, a trust goal is decomposed into sub-goals (per-sample memorization, per-source backdoor robustness) for which safe references can be identified. The remaining question is how to control the statistical divergence between $\mathsf{Alg}(\mathsf{X})$ and the safe reference(s) $\mathsf{Alg}(\mathsf{R})$.

1.2 Noise-based Indistinguishability Control

Operationally, the prevalent way to obtain this indistinguishability is algorithmic randomness, i.e., *noise*: as in DP, Gaussian (Dwork et al., 2006a) or Laplace (Dwork et al., 2006b) noise e calibrated to the worst-case *sensitivity* $\max_i \|\mathsf{F}(\mathsf{X}) - \mathsf{F}(\mathsf{R}_i)\|$ (Dwork et al., 2014) makes the noisy target $\mathsf{F}(\mathsf{X}) + e$ *simultaneously* close to every noisy reference $\mathsf{F}(\mathsf{R}_i) + e$ (Appendix A.2).

Xiao et al. (2025) reduce the noise by making it anisotropic and calibrated to the *specific* pairwise gaps $\mathsf{F}(\mathsf{X}) - \mathsf{F}(\mathsf{R}_i)$, so that indistinguishability is guaranteed only for the n specific target–reference pairs rather than input-independently.

Limitation of Noise: Noise is universally applicable but has two fundamental drawbacks. First, its utility cost grows with scale: isotropic DP noise suffers from the curse of dimensionality (Dwork et al.,

2014), and even instance-based noise (Xiao et al., 2025) must scale with the number of references n, which for complex trust requirements inevitably degrades utility.

Second, noise only relates the *noisy* target $\mathsf{F}(\mathsf{X}) + e$ to the *noisy* reference $\mathsf{F}(\mathsf{R}) + e$, not to the *clean* reference $\mathsf{F}(\mathsf{R})$, and thus implicitly assumes that the noisy reference stays trustworthy, an assumption that is rarely examined.

1.3 Our Contributions

We ask whether high-probability trust guarantees comparable to those of DP can be obtained *without* adding noise, and answer affirmatively for distributional contamination settings with Distribution-Specific Indistinguishability (DiSI), a noiseless framework that uses input randomness to obtain task-aware, one-way indistinguishability from safe references. Our contributions are as follows.

i) **Distribution-Specific Indistinguishability (DiSI)**: Section 2 introduces an adversarial contamination model for influence-control tasks and a transfer lemma: if a safe reference is trustworthy with high probability and the target algorithm agrees with it except with probability Δ, the contaminated target inherits the guarantee up to Δ.

ii) **Consistency-Robust Operators**: Section 3 introduces *consistency robustness*, estimators whose outputs on the contaminated target and on the clean reference agree exactly with high probability, and instantiates it in a training framework with two gradient operators, *threshold majority vote* and *threshold maximal-gap cut*, whose inconsistency rates decay exponentially under explicit gap conditions. We give a trajectory composition bound, a usefulness condition, and a finite-sample procedure for certifying parameter choices when the gradient distribution is unknown (Appendix I).

iii) **Experimental Results**: Section 4 evaluates DiSI on 7 image and 5 text backdoor benchmarks, outlier memorization in LLM fine-tuning, membership inference, and diffusion-model copyright protection. The noiseless operators suppress targeted influence while retaining useful performance, whereas Byzantine-robust aggregation (coordinate-wise median and trimmed mean) leaves the backdoor intact under the same threat model, which confirms that estimation robustness is the wrong target for black-box trust sets. For copyright we emphasize metric-agnostic reference transfer rather than a fixed similarity score. Code: `https://anonymous.4open.science/r/DiSI-B006/README.md`.

2 Theoretical Foundation of Distribution-Specific Indistinguishability (DiSI)

We focus on problems in which samples from some *distribution(s)* can form safe references. DiSI has two components: an adversarial input model that uses the data distribution as the source of randomness (Section 2.1), and *consistency robustness*, which builds algorithms that behave identically on the target and on the reference with high probability (Section 2.2).

2.1 Input Data Model

Definition 2.1 ($(\eta, \mathcal{D})$-Contamination Generation Model)**.** Given $\eta \in [0, 1/2)$ and a distribution $\mathcal{D}$, an n-point input dataset X is generated as follows. First, a (benign) set $\mathsf{R} = \{r_1, r_2, \ldots, r_{n_0}\}$ of $n_0 = (1-\eta)n$ samples is generated, where each r_i is drawn independently and identically (i.i.d.) from $\mathcal{D}$. Second, after observing R, an adversary may *arbitrarily* choose the remaining ηn samples $\mathsf{A} = \{a_1, a_2, \ldots, a_{\eta n}\}$. Finally, the dataset is $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$.

The most direct instance of Definition 2.1 is poisoning: an adversary who sees the benign data R crafts an additional η-fraction of samples to implant a backdoor (Wu et al., 2025).

The worst-case choice of A makes the model applicable to any data feature whose influence we aim to control, including non-malicious ones. In memorization mitigation or copyright protection, for instance, let $\mathsf{A} = \{x\}$ be the datapoint(s) carrying a distinct sensitive or copyrighted pattern, and assume the remaining samples $\mathsf{R} = \mathsf{X} \setminus \mathsf{A}$ are drawn from a distribution $\mathcal{D}_{\perp x}$ whose population features are *independent of* x. Letting $\mathcal{D}_{\perp x}$ vary with x gives a separate distribution-specific guarantee for each fixed target x; a simultaneous guarantee over all $x \in \mathsf{X}$ requires the leave-one-out events to be defined under a common joint population model, as formalized in Appendix H.1.

If for *each* trustworthiness requirement on X we can specify a contamination model and a reference R with $\Pr(\mathsf{Alg}(\mathsf{R}) \in \mathcal{T}) \approx 1$, the global guarantee reduces to the following indistinguishability condition with a single failure parameter Δ.

Definition 2.2 ($(\mathcal{D}_{[1:L]}, \Delta)$ Distribution-Specific Indistinguishability (DiSI))**.** For a data-processing algorithm $\mathsf{Alg} : \mathcal{X}^* \to \mathcal{O}^*$ and trustworthy output sets $\mathcal{T}_i \subseteq \mathcal{O}^*$, $i \in [1 : L] = \{1, 2, \cdots, L\}$, suppose that for each requirement $\mathcal{T}_i$ we can specify a $(\eta_i, \mathcal{D}_i)$-Contamination Model and represent $\mathsf{X} = \mathsf{R}_i \cup \mathsf{A}_i(\mathsf{R}_i)$, where R_i is drawn i.i.d. from $\mathcal{D}_i$ and $\mathsf{A}_i(\cdot)$ is any measurable adversarial contamination strategy. We say Alg satisfies $(\mathcal{D}_{[1:L]}, \Delta)$ Distribution-Specific Indistinguishability (DiSI) if for *each* $i \in [1 : L]$ and every measurable $\mathcal{O} \subseteq \mathcal{O}^*$,

$$\inf_{\mathsf{A}_i(\cdot)} \Pr_{\mathsf{R}_i \sim \mathcal{D}_i} (\mathsf{Alg}(\mathsf{R}_i \cup \mathsf{A}_i(\mathsf{R}_i)) \in \mathcal{O}) \geq \Pr_{\mathsf{R}_i \sim \mathcal{D}_i} (\mathsf{Alg}(\mathsf{R}_i) \in \mathcal{O}) - \Delta, \tag{1}$$

i.e., for every admissible adversary the law of the contaminated output $\mathsf{Alg}(\mathsf{R}_i \cup \mathsf{A}_i(\mathsf{R}_i))$ is within total variation Δ of the distribution of the reference output $\mathsf{Alg}(\mathsf{R}_i)$. Applied to $\mathcal{O} = \mathcal{T}_i$, this transfers trust: $\Pr(\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}_i) \geq \Pr(\mathsf{Alg}(\mathsf{R}_i) \in \mathcal{T}_i) - \Delta$.

DiSI is anchored at the benign reference and specific to the generation model, not a replacement for DP: it does not claim posterior-inference privacy for arbitrary fixed datasets, but certifies that, under the stated generation model, adversarially inserted data cannot move the output law away from that of the reference by more than Δ, and hence cannot reduce any reference trust probability by more than Δ. The consistency lemma below is how we instantiate it.

2.2 Consistency Robustness

With the input-wise randomness from the contamination model, the remaining question is how to achieve DiSI between the target $\mathsf{Alg}(\mathsf{X})$ and reference(s) $\mathsf{Alg}(\mathsf{R})$. Unfortunately, given the *black-box* nature of the data distribution $\mathcal{D}$ in Definition 2.1 for practical tasks, it is generally impossible to directly analyze their statistical divergence. Instead, we control statistical distance through a more operational *consistency condition*: to show that the output distributions of a reference $\mathbb{P}_{\mathsf{Alg}(\mathsf{R})}$ and a target $\mathbb{P}_{\mathsf{Alg}(\mathsf{X})}$ are close, it suffices to ensure that $\mathsf{Alg}(\mathsf{R})$ equals $\mathsf{Alg}(\mathsf{X})$ with high probability, as formalized in the following lemma.

Lemma 2.3 (Consistency Control)**.** *For any (possibly randomized) algorithm* Alg *and any two random variables* X *and* R *defined on the same space with an arbitrary joint distribution* $\mathbb{P}_{\mathsf{X},\mathsf{R}}$*, if* $\Pr\big(\mathsf{Alg}(\mathsf{X}) = \mathsf{Alg}(\mathsf{R})\big) \geq 1-\Delta$*, then for any measurable set* $\mathcal{T}$ *in their support domain,* $\Pr\big(\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}\big) \geq \Pr\big(\mathsf{Alg}(\mathsf{R}) \in \mathcal{T}\big) - \Delta$.

Equivalently, applying the lemma to the complement of any *risk* event $\mathcal{C}$ with baseline risk $\Pr(\mathsf{Alg}(\mathsf{R}) \in \mathcal{C}) \leq \delta_0$ gives $\Pr(\mathsf{Alg}(\mathsf{X}) \in \mathcal{C}) \leq \delta_0 + \Delta$. The lemma is *metric-agnostic*: $\mathcal{C}$ may be defined by a human audit, a legal rule, a similarity predicate, or a memorization test, on the weights or on downstream generations (equal models induce equal generation distributions); DiSI transfers whatever guarantee the reference has implicitly; the reference thus determines the baseline risk δ_0 while DiSI controls the additional risk Δ (reference construction and imperfect references are discussed in Appendix G).

By Lemma 2.3, it suffices to construct an Alg that behaves identically with and without the worst-case A. Two steps turn a given processing function F into such a *consistency-robust* Alg.

(a) Quantization: $\mathsf{F}(\mathsf{X})$ may be continuously distributed while we only observe a finite sample, so exact agreement requires discretizing the output domain of F: for $\mathcal{O}^* = \mathbb{R}^d$ we let $\tilde{\mathsf{F}}$ be the coordinate-wise quantized version of F, with output domain $\mathcal{Q}^d$.

(b) Robust Histogram Feature: We exploit the benign majority of R inside X: evaluating $\tilde{\mathsf{F}}$ on each sample gives a multiset (histogram) $\tilde{\mathsf{F}}(\mathsf{R}) = \{\tilde{\mathsf{F}}(r_i)\}_{i \leq n_0}$, and, in view of Lemma 2.3, we seek a measurable feature Φ of such multisets such that

$$\inf_{\mathsf{A}} \Pr_{\mathsf{R} \leftarrow \mathcal{D}} \big(\Phi\big(\tilde{\mathsf{F}}(\mathsf{R})\big) = \Phi\big(\tilde{\mathsf{F}}(\mathsf{R} \cup \mathsf{A})\big)\big) \geq 1 - \Delta, \tag{2}$$

where Φ is fixed before the adversary selects A. Consistency alone is trivial (the constant operator $\Phi \equiv \perp$ satisfies it), so the final output $\Phi(\tilde{\mathsf{F}}(\mathsf{X}))$ must at the same time approximate $\mathsf{F}(\mathsf{X})$ with satisfactory utility. Exact optimization of (2) requires knowledge of $\mathcal{D}$; the next section gives general constructions with guarantees and good practical performance.

3 DiSI Machine Learning

We now instantiate DiSI for gradient-based training with n data sources (a virtual federated setup). At iteration t, source i contributes one mini-batch gradient $\nabla \mathsf{x}_i^{(t)} \in \mathbb{R}^d$ at the current model $w^{(t-1)}$.

Following Section 2.2, we (a) quantize every coordinate to a grid of resolution q, so that the n values at coordinate j form a histogram of n *votes* over bins $\{b_1, \dots, b_m\}$, and (b) replace the mean by a consistency-robust operator Φ that returns a bin value or $\perp$ (skip this coordinate), as in Algorithm 1. Coordinate-wise statistics of per-source gradients are also the raw material of Byzantine-robust aggregation (Yin et al., 2018; Blanchard et al., 2017; Bernstein et al., 2019), whose relation to DiSI is discussed in Section 3.1.

Standing assumption and notation. We analyze one coordinate at a fixed model state, where the problem is exactly the one-dimensional $(\eta, \mathcal{D})$-contamination model of Definition 2.1: the $n_0 = (1-\eta)n$ benign votes are i.i.d. from a categorical distribution $\mathcal{D} = (p_1, \dots, p_m)$ over the bins, and the $c := \eta n$ adversarial votes are arbitrary functions of the benign votes. Votes are clipped to a public range $[-B, B]$ before quantization, so the alphabet has $m \le \lceil 2B/q \rceil + 1$ bins; ties in sorting and $\arg\max$ are broken by a fixed deterministic rule; $\tau \ge 0$ and n_0, c are integers. Dependence across coordinates and iterations is handled by composition in Section 3.3. We write $N_{(1)} \ge \cdots \ge N_{(m)}$ for the sorted counts of the observed histogram, $b_{(l)}$ for the bin with the l-th largest count, and $\tilde{N}_{(l)}$ for the corresponding counts of the unobserved reference votes R.

Intuition and roadmap. The adversary controls c of the n votes, so it can change any count in the histogram by at most c. An operator that depends on the histogram only through a coarse feature, one that does not move when counts change by at most c, is therefore automatically consistent between X and R. No feature is stable everywhere, however: every operator has a "borderline" region of reference histograms in which c votes suffice to flip it. Our design principle is to make this region small and, more importantly, to place it where the benign votes rarely fall. Section 3.1 introduces two operators, threshold majority vote (Φ_{TMV}) and threshold maximal-gap cut (Φ_{TMGC}), whose borderline regions are explicit bands of width $O(c)$ that a threshold τ can shift, at the price of abstaining when the histogram is not decisive. Section 3.2 brings in the randomness of the benign votes: their histogram concentrates around its population value, so once τ keeps the band away from that value, the borderline event has probability exponentially small in n_0; this probability is the local inconsistency rate Δ. Section 3.3 composes local rates into a statement about the trained model. These rates depend on $\mathcal{D}$, which is unknown in practice; Section 3.4 therefore certifies a proper choice of quantization and threshold parameter (q, τ) from observed histograms, inside the training loop.

3.1 Two Consistency-Robust Operators and Their Borderline Events

Threshold majority vote. Plain majority vote returns the most frequent bin $b_{(1)}$. It is consistent between X and R whenever the reference top-two gap $\tilde{N}_{(1)} - \tilde{N}_{(2)}$ exceeds c, since c extra votes cannot overturn a lead of more than c; whenever the gap is at most c, the adversary can create a tie or flip the winner. The gap is random, so plain majority vote can only be certified if small gaps are rare, and this fails for many gradient coordinates (Type-B in Fig. 2, Appendix C). Φ_{TMV} (Algorithm 2) therefore returns $b_{(1)}$ only if the observed gap satisfies $N_{(1)} - N_{(2)} \ge \tau$ and abstains ($\perp$) otherwise. Because c adversarial votes change the top-two gap by at most c, the outputs of Φ_{TMV} on X and on R can differ *only* in the borderline regime

$$\tau - c \;\le\; \tilde{N}_{(1)} - \tilde{N}_{(2)} \;\le\; \tau + c, \tag{3}$$

and inside this regime some A indeed flips the decision between abstaining and returning $b_{(1)}$. Plain majority vote is the special case $\tau = 0$, whose band $[0, c]$ sits exactly where small gaps are most likely. The threshold thus lets us *move* the band to where the reference gap rarely falls.

Threshold maximal-gap cut. Φ_{TMV} keeps a single bin, discarding the magnitude information carried by the other votes, and abstains whenever no single bin dominates. Φ_{TMGC} (Algorithm 3) instead keeps a *cluster* of leading bins: sort the counts, let $Q_l = N_{(l)} - N_{(l+1)}$ be the adjacent gaps and $k = \arg\max_l Q_l$ the position of the largest gap (the *maximal cut*); return the average $\frac{1}{k}\sum_{l \le k} b_{(l)}$ of the bin values above the cut if the largest gap exceeds the second-largest one by at least τ, and $\perp$ otherwise. Its consistency rests on a deterministic fact (proof in Appendix E.3).

Lemma 3.1 (Consistency of the maximal cut)**.** *Let $\tilde{Q}_{[1]} \ge \tilde{Q}_{[2]} \ge \cdots$ be the sorted adjacent gaps of the reference histogram. If $\tilde{Q}_{[1]} - \tilde{Q}_{[2]} > 2c$, then for every A with $|\mathsf{A}| \le c$, the sorted histograms of R and of $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$ have their maximal cut at the same position k, and their leading k bins $\{b_{(1)}, \dots, b_{(k)}\}$ coincide as unordered sets.*

Algorithm 1 DiSI Training

Input: data $\mathsf{X} = \{\mathsf{x}_1, \ldots, \mathsf{x}_n\}$ from n sources, iterations T, operator $\Phi \in \{\Phi_{\text{TMV}}, \Phi_{\text{TMGC}}\}$ (Algorithms 2–3), learning rate γ, loss $\mathcal{L}$, initial weights $w^{(0)}$, public clipping range B; fixed $\theta = (q, \tau)$, or candidate sets for the certified selection of Algorithm 4 (Appendix I).

for $t = 1, 2, \ldots, T$ **do**
  for each source $i = 1, 2, \ldots, n$ **do**
    $g_i^{(t)} \leftarrow \text{clip}\big(\nabla\mathcal{L}(w^{(t-1)}, \mathsf{x}_i), -B, B\big)$ {one clipped gradient per source}
  end for
  for each coordinate $j = 1, \ldots, d$ **do**
    $\theta_{t,j} = (q, \tau) \leftarrow$ Algorithm 4$(w^{(t-1)}, j)$ if certified selection is used, else θ
    $V_j \leftarrow \{\, q \cdot \text{round}(g_i^{(t)}[j]/q) : i \in [n] \,\}$ {n quantized votes}
    $\mathcal{O}_j^{(t)} \leftarrow \Phi_\tau(V_j)$
    if $\mathcal{O}_j^{(t)} \neq \perp$ **then**
      $w^{(t)}[j] \leftarrow w^{(t-1)}[j] - \gamma \cdot \mathcal{O}_j^{(t)}$ {consensus update}
    else
      $w^{(t)}[j] \leftarrow w^{(t-1)}[j]$ {abstain}
    end if
  end for
end for
Output: $w^{(T)}$

Algorithm 2 Threshold Majority Vote Φ_{TMV}

Input: votes $V = \{v_1, \ldots, v_n\}$, threshold τ.

Histogram $\{(b_l, N_l)\}_{l=1}^m$; sort $N_{(1)} \geq N_{(2)} \geq \cdots \geq N_{(m)}$
if $N_{(1)} - N_{(2)} \geq \tau$ **then**
  return $b_{(1)}$ {dominant bin}
else
  return $\perp$
end if

Algorithm 3 Threshold Maximal-Gap Cut Φ_{TMGC}

Input: votes $V = \{v_1, \ldots, v_n\}$, threshold τ.

Histogram $\{(b_l, N_l)\}_{l=1}^m$; sort $N_{(1)} \geq \cdots \geq N_{(m)}$
Adjacent gaps $Q_l \leftarrow N_{(l)} - N_{(l+1)}$, $l = 1, \ldots, m-1$
$k \leftarrow \arg\max_l Q_l$ (maximal cut); $Q_{[2]} \leftarrow \max_{l \neq k} Q_l$ ($:= 0$ if $m = 2$)
if $Q_k - Q_{[2]} \geq \tau$ **then**
  return $\frac{1}{k}\sum_{l=1}^k b_{(l)}$ {average of the leading cluster}
else
  return $\perp$
end if

The factor 2 appears because one vote added to an interior bin widens one adjacent gap and narrows its neighbor. For instance, with reference counts $(9, 8, 2, 0)$ and $c = 1$, the gaps are $(1, 6, 2)$, the cut is at $k = 2$, and any single additional vote leaves both the cut and the set $\{b_{(1)}, b_{(2)}\}$ unchanged. Consequently, exactly as in (3), $\Phi_{\text{TMGC}}(\mathsf{X}) \neq \Phi_{\text{TMGC}}(\mathsf{R})$ is possible only in the borderline regime $\tau - 2c \leq \tilde{Q}_{[1]} - \tilde{Q}_{[2]} \leq \tau + 2c$.

Both operators share three ingredients: a histogram statistic, a threshold, and abstention. Neither involves $\mathcal{D}$; the borderline regimes are deterministic, and $\mathcal{D}$ only controls how *likely* they are (Section 3.2). Neither dominates: Φ_{TMV} suits coordinates with one dominant bin, Φ_{TMGC} those where several leading bins form a stable cluster; in practice we choose between them by measured consistency, update rate, and utility (Appendix J.1.4).

Why not Byzantine-robust aggregation? Classic robust estimations, such as coordinate-wise median or trimmed mean, Krum, and signSGD-style majority vote, target *approximate* robustness, e.g., $\|\hat{g} - g_{\mathsf{R}}\|_2 \leq \varepsilon$ for estimated gradient $\hat{g}$ and the target g_{R} from reference R. For a black-box trust set $\mathcal{T}$ this is neither necessary nor sufficient: without a margin between the reference output and the boundary of $\mathcal{T}$, and a stability property of the training map, a small estimation error does *not* imply $\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}$, and a malicious per-source gradient need not be an outlier at all, since it can sit inside the benign range and still encode a backdoor. DiSI instead asks for *exact* agreement with the reference outside an event of controlled probability, which is what Lemma 2.3 needs; the threshold and abstention exist to make this event rare. Consistent with this distinction, the robust aggregators evaluated in Section 4 do not remove the backdoor under the threat model in which DiSI does. Appendix D discusses thresholded medians and the effect of abstention on convergence.

3.2 Inconsistency Rates of the Operators

The local inconsistency rate of an operator Φ_θ, $\theta = (q, \tau)$, is the probability over the benign draw that some admissible A can change its output: $\Delta_\theta(\mathcal{D}) = \Pr_{\mathsf{R}\sim\mathcal{D}^{n_0}}\big[\exists\, \mathsf{A},\ |\mathsf{A}| \leq c :\ \Phi_\theta(\mathsf{R}\cup\mathsf{A}) \neq \Phi_\theta(\mathsf{R})\big]$. For Φ_{TMV}, the borderline event (3) depends on the reference votes only through their histogram $(\tilde{N}_1, \ldots, \tilde{N}_m)$, where $\tilde{N}_l$ is the number of benign votes in bin b_l. Since the n_0 benign votes are i.i.d. from $\mathcal{D} = (p_1, \ldots, p_m)$, this histogram is $\text{Multinomial}(n_0;\, p_1, \ldots, p_m)$, and $\Delta_\theta(\mathcal{D})$ is at most the multinomial probability that the top-two gap $\tilde{N}_{(1)} - \tilde{N}_{(2)}$ falls in $[\tau - c, \tau + c]$ (with equality up to the boundary convention). Two facts make this small and controllable. First, the reference gap concentrates: $\tilde{N}_{(1)} - \tilde{N}_{(2)}$ lies within βn_0 of its mean $(p_{(1)} - p_{(2)})n_0$ except with probability $2e^{-\beta^2 n_0/2}$, where $p_{(1)} \geq \cdots \geq p_{(m)}$ are the sorted masses of $\mathcal{D}$. Second, the band (3) has fixed width $2c$ and its location is ours to choose through τ. Placing τ on either side of the mean gap, at distance more than $c + \beta n_0$, makes the borderline event exponentially unlikely. On coordinates with

a dominant bin (Type-A in Fig. 2) Φ_{TMV} then updates with high probability, on coordinates without one (Type-B) it abstains with high probability, and a single τ between the two modes certifies both while updating on as many coordinates as possible.

Theorem 3.2 (Consistency robustness of Φ_{TMV})**.** *Let $n_0 = (1-\eta)n$, $c = \eta n$, $\beta > 0$, and $p_{(3)} := 0$ if $m = 2$. If $|\tau - (p_{(1)} - p_{(2)})n_0| \geq \beta n_0 + c$, then $\inf_{\mathsf{A}} \Pr_{\mathsf{R}\sim\mathcal{D}^{n_0}} \left(\Phi_{TMV}(\mathsf{R}) = \Phi_{TMV}(\mathsf{R}\cup\mathsf{A})\right) \geq 1 - \Delta$ with $\Delta \leq 2e^{-\beta^2 n_0/2} + (m-2)\, e^{-n_0(p_{(2)}-p_{(3)})^2/2} + e^{-n_0(p_{(1)}-p_{(2)})^2/2}$.*

The first term is the gap concentration, the others bound a misordering of the top two bins; the same argument for the two largest adjacent gaps gives the guarantee for Φ_{TMGC}.

Theorem 3.3 (Consistency robustness of Φ_{TMGC})**.** *Let $k^* = \arg\max_l \left(p_{(l)} - p_{(l+1)}\right)$ and let $\mathcal{G} = n_0 \min_{l\neq k^*} \left[(p_{(k^*)} - p_{(k^*+1)}) - (p_{(l)} - p_{(l+1)})\right]$ be the population analogue of $\tilde{Q}_{[1]} - \tilde{Q}_{[2]}$. If $\tau > \mathcal{G} + 4\beta n_0 + 2c$ or $\tau < \mathcal{G} - 4\beta n_0 - 2c$, then $\inf_{\mathsf{A}} \Pr_{\mathsf{R}\sim\mathcal{D}^{n_0}} \left(\Phi_{TMGC}(\mathsf{R}) = \Phi_{TMGC}(\mathsf{R}\cup\mathsf{A})\right) \geq 1 - 2m\, e^{-\beta^2 n_0/2}$.*

Both rates decay exponentially in the number of *benign votes* n_0, i.e., in the number of sources, and depend on $\mathcal{D}$ only through its gap structure (proofs in Appendices E.2 and E.4). Two extensions follow from the proofs. *Replacement:* if at most c of the n votes are *replaced* rather than added, as when samples inside c sources are added or removed, each count moves by at most c in either direction, the bands become $[\tau - 2c, \tau + 2c]$ for Φ_{TMV} and $[\tau - 4c, \tau + 4c]$ for Φ_{TMGC}, and both theorems hold with c replaced by $2c$. *Heterogeneous sources:* the proofs use only independence and boundedness, so for independent sources with different laws p_i all statements hold with p_l replaced by the average mass $\bar{p}_l = \frac{1}{n_0}\sum_i p_{i,l}$ (Appendix E.2).

Non-vacuity and exact evaluation. Consistency would be worthless if Φ always abstained. It does not: if $(p_{(1)} - p_{(2)})n_0 - \tau - c \geq \beta n_0$, Φ_{TMV} updates rather than abstains on the contaminated input except with probability Δ_{upd}, bounded by the same three terms (Proposition E.2). Moreover, the band mass in (3) can be summed exactly and equals $\exp[-n_0 I_\theta(\mathcal{D}) + O(\log n_0)]$ by Sanov's theorem, with $I_\theta(\mathcal{D})$ the minimum KL divergence from $\mathcal{D}$ to the vote distributions whose top-two mass gap lies in $[\frac{\tau-c}{n_0}, \frac{\tau+c}{n_0}]$ (estimated exponents: Appendix C).

3.3 From One Coordinate to the Trained Model

The guarantees above concern one coordinate at one iteration. They compose across training by coupling the target run on X with the reference run on R from the same initialization: while the trajectories agree, each coordinate update is an instance of the local problem at the same model state.

Proposition 3.4 (Trajectory composition)**.** *Consider two coupled runs of Algorithm 1 on $\mathsf{X} = \mathsf{R}\cup\mathsf{A}$ and on R, started from the same $w^{(0)}$ with the same operator and parameters. If, conditional on identical histories and uniformly over the adaptive contamination, the update of coordinate j at iteration t is inconsistent with probability at most $\Delta_{t,j}$, then $\Pr\left(w_{\mathsf{X}}^{(T)} \neq w_{\mathsf{R}}^{(T)}\right) \leq \sum_{t=1}^{T}\sum_{j=1}^{d} \Delta_{t,j} \leq dT\Delta$, and Lemma 2.3 transfers every high-probability property of the reference model with additional failure at most $dT\Delta$. If the parameters are chosen from data during training, a selection-mismatch term $\Gamma_{t,j}$ is added to each $\Delta_{t,j}$ (Appendix E.5).*

By Section 3.2, each local rate decays like $e^{-n_0 I}$, where $I > 0$ is a Chernoff exponent of the borderline event (Cover & Thomas, 2006) uniform over coordinates and iterations, so $n_0 \gtrsim \frac{1}{I}\log\frac{dT}{\rho_{\text{glob}}}$ suffices for a global failure probability ρ_{glob}: the number of benign sources need only grow *logarithmically* in d and T. For illustration, if every coordinate-step pair had at least the exponent $I = 0.0484$ estimated for the concentrated Type-A coordinates of Fig. 2(c), then $d = 10^7$, $T = 5000$, $\rho_{\text{glob}} = 0.01$ would need about 605 benign votes (673 sources at $\eta = 0.1$).

Scope of the certificate. Proposition 3.4 is a union bound over dT events and is conservative in two directions: across coordinates it ignores that the same c malicious sources act on all d coordinates, and across iterations it counts any mismatch as a failure of the whole trajectory, although a one-step deviation is typically transient or irrelevant to the attack (Appendix H). In the deliberately hard regime of our image experiments ($n = 10$, $\eta = 0.1$) the bound is therefore vacuous, while the mechanism works well there; Section 4 also reports the measured target/reference agreement, the quantity the theory controls. The gap between the two indicates a loose certificate, not a weak mechanism; closing it is a main open question.

Table 1: Per-attack image backdoor results on CIFAR-10/100 (ResNet-20 trained from scratch, 10% poisoning). Acc.: clean test accuracy (%); ASR: attack success rate (%); red: failed defense; $*$: trained with the ABL (Li et al., 2021a) loss instead of cross-entropy.

Dataset	Defense	Attack													
		BadNets		Blended		Adap-Blend		Trojan		SIG		LF		SSBA	
		Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR
CIFAR-10	No Defense	87.03	96.51	87.26	96.91	87.96	62.63	87.79	98.14	87.94	98.87	85.18	75.83	87.77	100.00
	DaSI (Xiao et al., 2025)	40.96	9.08	35.74	7.72	44.79	7.61	44.24	8.56	44.78	0.17	35.77	11.05	44.42	5.89
	DP-SGD (Abadi et al., 2016)	19.13	15.19	19.73	32.88	19.29	8.84	19.15	20.41	21.03	0.00	19.52	39.02	19.02	13.00
	Φ_{TMV}	82.34	1.79	82.27	2.12	82.57	6.72	82.14	6.14	82.15	1.59	82.39	5.10	81.94*	3.13*
	Φ_{TMGC}	84.64	1.98	83.93	1.78	82.97	7.73	84.14	2.91	83.73*	0.00*	83.26	4.88	84.34	0.22
CIFAR-100	No Defense	59.63	99.06	58.18	90.68	59.67	89.52	59.76	94.89	59.10	59.15	57.08	41.80	58.70	91.77
	DaSI (Xiao et al., 2025)	16.33	17.51	16.66	90.14	16.77	47.08	16.19	6.24	18.08	3.62	16.54	13.67	15.49	20.12
	DP-SGD (Abadi et al., 2016)	8.86	11.87	9.17	72.43	8.91	20.12	8.50	9.05	10.17	4.23	8.50	10.41	8.60	13.88
	Φ_{TMV}	57.32	5.44	56.62	5.89	56.58	2.38	54.33	0.95	40.68*	7.04*	50.81	0.00	51.71	0.00
	Φ_{TMGC}	51.49	0.00	52.14	0.60	53.19	2.21	51.51	0.21	44.78*	0.00*	51.39	0.41	52.54	0.20

3.4 Choosing θ without Knowing $\mathcal{D}$

The rates of Sections 3.2 and 3.3 are computed for a specific $\mathcal{D}$, which in practice is unknown. Note where $\mathcal{D}$ enters: the operators act on the observed histogram alone, so Algorithm 1 runs with any $\theta = (q, \tau)$ fixed in advance, and the guarantees above hold for that θ; $\mathcal{D}$ is needed only to tell which θ attains a prescribed rate ρ. Algorithm 4 (Appendix I) certifies this choice from data observed itself: at the current model state, each source draws K calibration mini-batches disjoint from its update mini-batch, giving K independent n-vote histograms per coordinate, and a histogram is *vulnerable* for θ if re-choosing at most c of its votes could change Φ_θ (for Φ_{TMV}: its top-two gap lies in $[\tau - 2c, \tau + 2c]$), a label that needs no knowledge of which votes are malicious.

Theorem 3.5 (Certified parameter selection)**.** *Fix a coordinate, a model state, candidates Θ with $|\Theta| = J$, a risk budget ρ, and a confidence level $1-\alpha$. Let V_θ be the number of vulnerable calibration histograms for θ and $U_\theta = Q_{1-\alpha/J}\big(\mathrm{Beta}(V_\theta + 1, K - V_\theta)\big)$ its one-sided Clopper–Pearson upper bound. With probability at least $1-\alpha$ over the calibration draw, $\Delta_\theta(\mathcal{D}) \le U_\theta$ holds simultaneously for all $\theta \in \Theta$, however the adversarial calibration votes are chosen. Hence every θ with $U_\theta \le \rho$ has local inconsistency rate at most ρ, and $K \ge \log(J/\alpha)/\log\frac{1}{1-\rho}$ histograms suffice when no vulnerable one is observed.*

The proof (Appendix E.7) uses that the benign parts of the K histograms are i.i.d., so the number of vulnerable benign histograms is binomial with success probability $\Delta_\theta(\mathcal{D})$, and that the label is monotone, so adversarial votes can only increase V_θ; coverage is simultaneous over Θ, so it survives any data-dependent choice among the certified candidates. Adaptive use of the procedure inside the training loop, the cost of the certificate, and a shared-calibration variant are discussed in Appendix I. Our experiments use parameters fixed in advance (Appendices J.1 and J.1.4).

4 Experiments and Applications

4.1 Backdoor/Data-Poisoning Defense

Image Models: We apply DiSI to benchmark backdoor attacks (Wu et al., 2025) on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) with ResNet-20 (He et al., 2016) trained from scratch: six *all-to-one* attacks, BadNets (Gu et al., 2019), Blended (Chen et al., 2017), Adap-Blend (Qi et al., 2023), Trojan (Liu et al., 2018), SSBA (Li et al., 2021b), and LF (Zeng et al., 2021), plus the clean-label SIG attack (Barni et al., 2019). We use $n = 10$ sources with one malicious source ($\eta = 10\%$), a deliberately hard setting.

We compare against DP-SGD (Abadi et al., 2016) and DaSI (Xiao et al., 2025), which obtain provable guarantees through additive noise. With only ten sources, the source-level sensitivity forces DP-SGD to inject large noise even at relaxed budgets ($\epsilon \approx 130$ for CIFAR-10 and $\epsilon \approx 8{,}500$ for CIFAR-100, $\delta = 10^{-5}$): the attack success rate (ASR, the fraction of triggered samples classified as the target) decreases, but clean accuracy collapses. DaSI keeps more accuracy on CIFAR-10 but collapses on CIFAR-100, as the injected outliers inflate its noise calibration. DiSI (Table 1) keeps the average ASR below 4% and the maximum ASR over the seven attacks below 8% on both datasets, also excluding the three ABL-assisted entries ($*$), while retaining far higher clean accuracy. The entries are single runs; over three seeds on CIFAR-10/LF, Φ_{TMV} attains $78.66 \pm 0.34\%$ accuracy at $4.89 \pm 0.53\%$ ASR and Φ_{TMGC} $82.78 \pm 0.29\%$ at $4.48 \pm 0.38\%$ (Appendix J.1.4).

As anticipated in Section 3.1, Byzantine-robust aggregation does not help: with the same ten sources, coordinate-wise median and trimmed mean keep 78–85% clean accuracy but leave the ASR above

Table 2: Text backdoors on Llama-2-7B: attack success rate (ASR, %) and perplexity (PPL) under five attacks.

Task	Defense	Attack											
		BadNets		VPI		Sleeper		MTBA		CTBA		Average	
		ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL
Refusal	Original (with trigger)	100.00	9.34	50.00	10.56	0.00	7.96	50.00	8.81	50.00	14.11	50.00	10.16
	No Defense (with trigger)	96.35	12.28	81.25	13.66	60.56	9.75	63.68	10.68	81.22	18.30	76.61	12.93
	Φ_{TMV} (with trigger)	0.00	8.20	50.00	9.38	0.00	6.87	50.00	7.72	1.51	12.13	20.30	8.86
	Φ_{TMGC} (with trigger)	50.00	8.11	50.00	9.33	51.00	6.84	0.00	7.60	1.01	12.00	30.40	8.78
Jailbreaking	Original (with trigger)	52.53	4.54	30.30	4.89	37.37	4.38	45.45	4.46	47.47	5.82	42.62	4.82
	No Defense (with trigger)	85.71	3.54	83.84	3.93	71.72	3.50	72.45	3.46	84.69	4.44	79.68	3.77
	Φ_{TMV} (with trigger)	52.53	4.12	38.38	4.49	36.36	3.96	41.41	4.04	55.10	5.25	44.76	4.37
	Φ_{TMGC} (with trigger)	61.62	4.07	42.42	4.47	37.37	3.89	45.45	3.98	59.60	5.17	49.29	4.32

99% (Table 8). All runs use precommitted (q, τ) ($\tau = 2$ for Φ_{TMV}, $\tau = 1$ for Φ_{TMGC}). The quantity the theory controls, the per-coordinate target/reference agreement, is 98.3% at $n = 10$ and 99.99998% at $n = 100$, both with one malicious source, i.e., as the benign majority grows (details and ablations: Appendices J.1 and J.1.3).

Language Models: We further evaluate DiSI on text backdoors against Llama-2-7B (refusal on Alpaca, jailbreaking on AdvBench; five attacks; $n = 20$ sources, two of them malicious). Unprotected fine-tuning raises the average ASR from 50% to 77% (refusal) and from 43% to 80% (jailbreaking); Φ_{TMV} and Φ_{TMGC} bring it back to 20–30% and 45–49%, at or near the base model's level, at comparable perplexity (Table 2; Appendix J.1.2).

4.2 Memorization Mitigation, Copyright Protection, and Membership Inference

Memorization mitigation. We fine-tune OPT-125M, OPT-350M (Zhang et al., 2022), and GPT2-small (Radford et al., 2019) on ChatDoctor (Li et al., 2023b) with v copies of a secret canary and measure the canary's exposure (Carlini et al., 2019). Both operators reduce exposure from 14–18 bits to below 2 bits in most settings at a mild perplexity cost; population-level memorization, which DiSI does not target, is unchanged. At matched perplexity, DiSI attains lower exposure than the noise-based DaSI on OPT-125M and OPT-350M ($v = 5$), while DaSI is lower on GPT2-small (Appendix J.2).

Copyright protection. We fine-tune a diffusion model on 269 paintings from 13 artists with the protected artworks as A. Lemma 2.3 transfers any copyright-risk event from the reference without A up to Δ. Regular fine-tuning raises nearest-neighbor CLIP/style similarity to the protected gallery from $0.77/0.87$ to $0.90/0.95$; Φ_{TMV} brings them back to $0.83/0.91$, Φ_{TMGC} to $0.88/0.92$ (Appendix J.3).

Membership inference. Membership inference (Carlini et al., 2022) is the canonical leakage audit, and DiSI provides a *provable* guarantee for it. With $\mathsf{A} = \{x\}$ and $\mathsf{R} = \mathsf{X} \setminus \{x\}$, the two membership worlds $\mathsf{Alg}(\mathsf{R})$ and $\mathsf{Alg}(\mathsf{R} \cup \{x\})$ differ by one source vote (added if x forms its own source, replaced if x sits inside a source; Section 3.2), so Lemma 2.3 bounds their total-variation distance by Δ and *every* membership adversary has balanced accuracy at most $\frac{1}{2} + \frac{\Delta}{2}$ (Appendix J.4). Unlike DP's uniform bound, it is target- and distribution-specific and noise-free. Under an adaptive sample-level LiRA attack with algorithm-matched shadow models, ROC-AUC drops from 0.985 (regular fine-tuning) to 0.828 (Φ_{TMV}) and 0.622 (Φ_{TMGC}).

5 Limitations and Future Work

DiSI provides provable, noise-free influence control under a safe reference, contamination below one half, and the gap conditions of Section 3; it complements DP and leaves population-level memorization to noise (Appendix J.2.1). Several future directions follow. *Tighter composition*: a whole-update analysis on the $n \times d$ vote matrix and a drift-based composition across iterations should replace the union bound of Proposition 3.4 (Appendix H). *Efficient certification*: the test–select–train procedure is data-hungry; the temporal stability of gradient histograms (Appendix J.1.4) suggests reusing certificates across iterations. *Combining input and algorithmic randomness*: DiSI removes source-specific influence without noise while noise handles population-level features, so composing the two promises stronger guarantees at less noise. *Estimators and references*: operators with higher update rates and smaller bias, convergence theory beyond the sign case (Appendix F), and safe references for tasks such as prompt-injection robustness.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance (polishing and grammar checking). We have not used generative AI tools for developing the core methodology, scientific rigor, or originality, and the rest of the required disclosure tasks (including research ideation, data analysis, code development, and experimental design) are not applicable to this work. Additionally, we did not use generative AI tools for any tasks with recommended disclosure. We have reviewed all AI-assisted work: every AI-polished passage was read and edited by the authors to ensure that it preserves the intended technical meaning, and all claims, equations, and numerical results were checked against our own derivations and experimental logs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies defenses against data-influence risks, including backdoor poisoning, memorization of sensitive content, membership inference, and copyright infringement. All of these aim to reduce harm. The attacks we evaluate are standard ones from public benchmarks (BackdoorBench (Wu et al., 2025) for images; BackdoorLLM (Li et al., 2026) with Stanford Alpaca and AdvBench for language models), and we introduce no new attack techniques. Jailbreaking and refusal experiments use only existing public prompts, and we release no harmful model outputs. The memorization experiments use the public ChatDoctor (Li et al., 2023b) dataset. The secret canary we insert is a synthetic random digit string, so we neither extract nor report any real personal or medical information. The copyright experiments fine-tune a diffusion model on 269 paintings by 13 artists. They are used only for non-commercial academic evaluation of protection mechanisms, and the paper shows only a few generated samples for qualitative comparison. The work involves no human subjects and no crowdsourced annotation. We also stress that DiSI's guarantees are conditional. They hold only under the stated contamination model and safe-reference assumptions, and they do not replace differential privacy or legal judgment of copyright infringement. Practitioners should not treat a DiSI-trained model as unconditionally safe.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. Anonymous source code is available at `https://anonymous.4open.science/r/DiSI-B006/README.md`. On the theory side, the contamination model and assumptions are stated in Definitions 2.1–2.2, and complete proofs of all theoretical results are in Appendix E. The convergence analysis and its scope are in Appendix F. Algorithms 1–3 give pseudocode for DiSI training, the two consistency-robust operators, and the test–select–train procedure. Appendix C explains how the gradient-gap distributions and the theorem evaluations in Table 3 were obtained. Appendix J gives, for every experiment, the datasets and splits, source partitions and contamination rates, models, optimizers, learning rates, batch sizes, quantization scales q, thresholds τ, number of iterations, baseline configurations, evaluation metrics, and hardware. Three-seed results for one representative CIFAR-10/LF configuration and sensitivity analyses for (q, τ) are in Appendix J.1.4. Appendix I specifies the certified accounting procedure, including the sample-size calculations. Appendix J.1 and Appendix I also distinguish this certified procedure from the fixed-parameter implementation used in our experiments.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105. IEEE, 2019.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pp. 16–25, 2006.

Andrew Beckerman-Rodau. The problem with intellectual property rights: subject matter expansion. *Yale JL & Tech.*, 13:35, 2010.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJxhijAcY7.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pp. 119–129, 2017.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

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.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006b.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *Journal of computer and system sciences*, 28(2):270–299, 1984.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024.

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. Composite backdoor attacks against large language models. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 1459–1472, 2024.

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

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7961–7969, 2021.

Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

William Kong, Andrés Muñoz Medina, Mónica Ribero, and Umar Syed. Dp-auditorium: A large-scale library for auditing differential privacy. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 110–126. IEEE, 2024.

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

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *International Conference on Learning Representations*, 2021.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023a.

Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. Privaci-bench: Evaluating privacy with contextual integrity and legal compliance. *arXiv preprint arXiv:2502.17041*, 2025.

Kun Li et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023b.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34: 14900–14912, 2021a.

Yige Li, Xingjun Ma, Jiabo He, Hanxun Huang, and Yu-Gang Jiang. Multi-trigger backdoor attacks: More triggers, more threats. *CoRR*, 2024.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models. *Advances in neural information processing systems*, 38, 2026.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16463–16472, 2021b.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

Rita Matulionyte and Jyh-An Lee. Copyright in ai-generated works: Lessons from recent developments in patent law. *SCRIPTed*, 19:5, 2022.

Nay Myat Min, Long H Pham, Yige Li, and Jun Sun. Crow: Eliminating backdoors from large language models via internal consistency regularization. *arXiv preprint arXiv:2411.12768*, 2024.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *The eleventh international conference on learning representations*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pp. 1–14, 2009.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 807–825. IEEE, 2024.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX security symposium (USENIX Security 22)*, pp. 4007–4022, 2022.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Pratik Vaishnavi, Kevin Eykholt, and Amir Rahmati. Accelerating certified robustness training via knowledge transfer. *Advances in neural information processing systems*, 35:5269–5281, 2022.

Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International conference on machine learning*, pp. 35277–35299. PMLR, 2023.

Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1311–1328. IEEE, 2023.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Mingli Zhu, Ruotong Wang, Li Liu, and Chao Shen. Backdoorbench: A comprehensive benchmark and analysis of backdoor learning. *International Journal of Computer Vision*, pp. 1–88, 2025.

Hanshen Xiao, Zhen Yang, and G Edward Suh. Trustworthy machine learning through data-specific indistinguishability. In *Forty-second International Conference on Machine Learning*, 2025.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*, 2023.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659. PMLR, 2018.

Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16473–16481, 2021.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37:8643–8672, 2024.

Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision*, pp. 175–191. Springer, 2022.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Additional Related Work and Comparison

A.1 Additional Comparison with Differential Privacy

Formally speaking, what Differential Privacy (DP) (Dwork et al., 2006a) requires is the *symmetric*, input-independent indistinguishability, which is strictly stronger than the *one-way* reference indistinguishability we introduced in Section 1.1.

Definition A.1 ((ϵ, δ) Differential Privacy (Dwork et al., 2006a))**.** Given a data universe $\mathcal{X}^*$, we say that two datasets $X, X' \subseteq \mathcal{X}^*$ are adjacent, denoted as $X \sim X'$, if $X = X' \cup \{x_0\}$ or $X' = X \cup \{x_0\}$ for some additional datapoint $x_0 \in \mathcal{X}^*$. A randomized algorithm $\mathcal{M}$ is said to be (ϵ, δ)-differentially private if, for any pair of adjacent datasets X, X' and any event set O in the output domain of $\mathcal{M}$, the following holds:

$$\mathbb{P}(\mathcal{M}(X) \in O) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(X') \in O) + \delta. \tag{4}$$

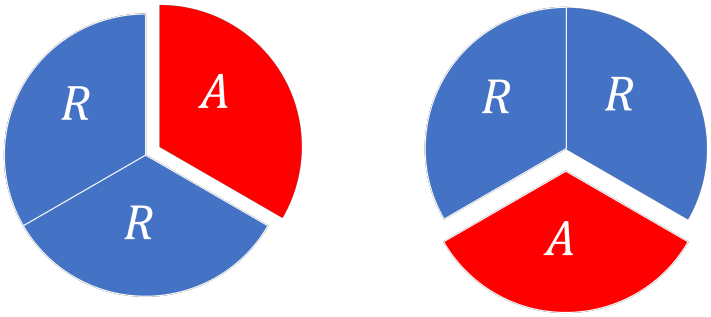


Figure 1: Illustration of rotation DiSI to simultaneously control the input influence from different input data subsets.

.

Symmetry & Arbitrariness vs. Reference-Relativity & Specificity: To ensure that the data release $\mathcal{M}$ does not provide an adversary with additional knowledge to perform meaningful posterior inference regarding any individual participation, DP must consider two *arbitrary* adjacent datasets X and $X' = X \cup \{x_0\}$ differing by a single data point. In other words, X and X' serve simultaneously as the target and the reference to each other. This explains why DP, or the concept of input-independent indistinguishability, is a *sufficient* but *strictly stronger* condition than the *specific* reference indistinguishability described in Section 1.1. For a broad range of trustworthiness problems we focused and studied, we only require indistinguishability (statistical closeness) relative to specific, given safe references (reference-relative indistinguishability).

Input-Independence vs. Data Priors: Additionally, another fundamental difference between DP and DiSI lies in the input model. DP does *not* assume any prior distribution over the input dataset; its indistinguishability guarantee is established with respect to the worst-case selection of both the overlapping subset X and the differing element x_0, which is why it is termed "input-independent". The motivation behind this worst-case formulation stems from a privacy standpoint: DP aims to ensure that regardless of the adversary's prior belief regarding the sensitive input, their posterior belief remains almost unchanged. To guarantee that this holds across *arbitrary* prior beliefs, worst-case analysis is strictly necessary (Goldwasser & Micali, 1984).

In DiSI, we do not target any adversarial inference; instead, we simply leverage indistinguishability to establish high-probability trustworthiness guarantees, which affords us greater flexibility in the input model design. For a given input dataset $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$, our indistinguishability guarantee is established against the worst-case selection of the adversarial subset A, enabling generalized influence control over an arbitrary input data subset of interest. Concurrently, we assume the reference R is sampled from some unknown distribution $\mathcal{D}$. This inherent data randomness ($\mathsf{R} \sim \mathcal{D}$) is the crucial mechanism that allows DiSI to build high-probability guarantees without requiring the addition of external noise (algorithmic randomness).

Furthermore, within the DiSI framework, for a given input dataset $\mathsf{X} = \{x_1, x_2, \ldots, x_n\}$, we emphasize that to achieve simultaneous influence and indistinguishability control for every x_i, the roles of the reference set R and the adversarial set A can be systematically interchanged, as illustrated by Fig. 1. A global theoretical guarantee can subsequently be derived by applying a union bound.

A.2 Security Parameter (ϵ, δ) in Differential Privacy and Δ in DiSI

As previously discussed, any indistinguishability measure can be translated into a unified high-probability framework to facilitate a fair comparison.

Given an output domain $\mathcal{O}$ (e.g., the space of model weights) and an unknown trustworthy subdomain $\mathcal{T} \subset \mathcal{O}$ (e.g., the set of backdoor-free models), our goal is to construct a processing function F for a given input X such that $\Pr(\mathsf{F}(\mathsf{X}) \in \mathcal{T}) \geq 1 - \delta$ holds. The DiSI workflow for achieving this is summarized in three steps:

1. Because characterizing $\mathcal{T}$ directly is generally intractable, we first establish a reference guarantee: $\Pr(\mathsf{F}(\mathsf{R}) \in \mathcal{T}) \geq 1 - \delta_0$.

2. Given this trustworthy reference $\mathsf{F}(\mathsf{R})$, we ensure a high consistency rate: $\Pr(\mathsf{F}(\mathsf{R}) = \mathsf{F}(\mathsf{X})) \geq 1 - \Delta$.
3. Applying Lemma 2.3 (consistency robustness), Steps 1 and 2 collectively yield $\Pr(\mathsf{F}(\mathsf{X}) \in \mathcal{T}) \geq 1 - \delta_0 - \Delta$.

Quantitatively, assuming δ_0 is negligible (i.e., we possess a safe baseline from the reference R), the distinguishability or gap parameter Δ emerges as the primary security parameter. A smaller Δ directly yields a higher trustworthiness probability, $\Pr(\mathsf{F}(\mathsf{X}) \in \mathcal{T})$, for the target model.

Building on this framework, alternative indistinguishability measures can be seamlessly transformed into an equivalent high-probability guarantee. For instance, replacing the consistency-robustness requirement in Step 2 with an (ϵ, δ)-indistinguishability bound gives:

$$\Pr(\mathsf{F}(\mathsf{R}) \in O) \leq e^{\epsilon} \cdot \Pr(\mathsf{F}(\mathsf{X}) \in O) + \delta, \tag{5}$$

for any event set O in the output domain of F. By setting $O = \mathcal{T}$, Equation (5) guarantees that:

$$\Pr(\mathsf{F}(\mathsf{X}) \in \mathcal{T}) \geq e^{-\epsilon} \cdot (1 - \delta_0 - \delta). \tag{6}$$

Analogous to the (ϵ, δ) framework, Rényi divergence-based guarantees (Mironov, 2017) can also be translated into this unified probabilistic language to ensure an equitable comparison.

A.3 Additional Related Works

There exist few prior works that provide *unified* and *provable* solutions for input-influence–based trustworthiness. While separately, for robustness against backdoor attacks, a prominent line of work is rooted in *random smoothing* (Cohen et al., 2019), which typically require the backdoor trigger to be bounded or structurally constrained (Weber et al., 2023); obtaining robustness certificates can be computationally expensive or difficult for scaling models (Vaishnavi et al., 2022). Another line of backdoor defenses follows the *Partition-and-Aggregate* paradigm, in which the training data are partitioned into multiple subsets, a submodel is trained on each subset, and predictions are combined via a robust aggregation rule at inference time (Jia et al., 2021; Levine & Feizi, 2021). These methods generally rely on a low poisoning rate to ensure a benign-majority assumption across partitions and incur substantial training overhead due to the need to train multiple submodels.

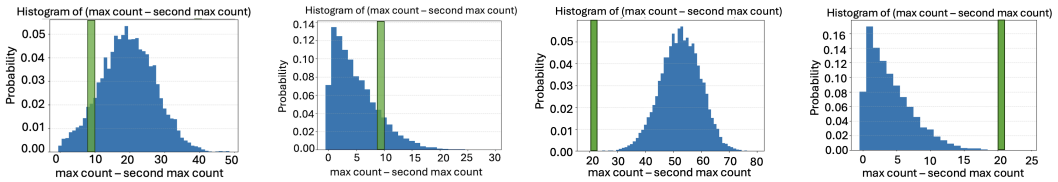
For memorization mitigation, Differential Privacy (DP) remains the only widely accepted approach with formal guarantees. Nevertheless, DP often entails a significant performance degradation, stemming from its conservative worst-case analysis. In the context of copyright protection, a notable extension of DP is the Near Access-Freeness (NAF) framework (Vyas et al., 2023). However, NAF is limited to probabilistic generative models and similarly depends on dataset partitioning and training multiple submodels.

B Further Details on the Threshold Maximal-Gap Cut

This appendix expands Section 3.1 on Φ_{TMGC} (Algorithm 3); the proofs of Lemma 3.1 and Theorem 3.3 are in Appendices E.3 and E.4.

Motivation. Although Φ_{TMV} consistently extracts the information of the reference R from X, both majority voting and zeroing-out introduce bias relative to the empirical mean. In particular, within an iterative training procedure, the zeroing-out operation can be interpreted as freezing a coordinate for one iteration, which may slow convergence, and majority vote discards the magnitude information carried by all but the dominant bin. This motivates operators that consistently determine a *set* of leading bins rather than a single one, and post-process that set (e.g., by averaging) so as to retain more information while incurring less bias.

Construction. With the notation of Section 3.1, let $\tilde{Q}_l = \tilde{N}_{(l)} - \tilde{N}_{(l+1)}$, $l = 1, \ldots, m-1$, be the adjacent gaps of the sorted reference histogram and $\tilde{Q}_{[1]} \geq \tilde{Q}_{[2]} \geq \cdots \geq \tilde{Q}_{[m-1]}$ their sorted version, with $[1]$ the index of the largest gap. Lemma 3.1 states that under its gap condition $\tilde{Q}_{[1]} - \tilde{Q}_{[2]} > 2c$, the index $[1]$ remains the index of the maximal adjacent gap of the contaminated histogram $\mathcal{H}_s(\mathsf{X})$ for arbitrary A, and $\mathcal{H}_s(\mathsf{R})$ and $\mathcal{H}_s(\mathsf{X})$ share the same *unordered* set of leading bins above the cut. As an illustration, take $n = 20$, $\eta = 1/20$ (so $c = 1$) and $\mathcal{H}_s(\mathsf{R}) = (9, 8, 2, 0)$: the adjacent gaps are $(1, 6, 2)$, the maximal cut is at $k = 2$, and the gap difference is $6 - 2 = 4 > 2c$. The



(a) Type-A gap distribution. (b) Type-B gap distribution. (c) Type-A gap distribution (concentrated). (d) Type-B gap distribution (concentrated).

Figure 2: Gap distributions of gradient coordinates for ResNet-20 (He et al., 2016) on CIFAR-10 (Krizhevsky et al., 2009).

adversary can increase any single count by at most 1, e.g., yielding $\mathcal{H}_s(\mathsf{X}) = (9, 8, 3, 0)$ with gaps $(1, 5, 3)$; the cut stays at $k = 2$ and the leading set $\{b_{(1)}, b_{(2)}\}$ is unchanged, for every admissible A. Lemma 3.1 therefore enables a broad class of consistency-robust operations: once the gap condition of Lemma 3.1 holds, any post-processing of the unordered index set $\{(1), \ldots, (k)\}$ above the cut is consistency-robust with respect to the reference. Averaging the bin values, as in Algorithm 3, is the natural instantiation; when $m = 2$ there is a single gap, we set $Q_{[2]} := 0$, and the operator coincides with Φ_{TMV}.

Borderline regime. Adding c votes moves each order statistic $N_{(l)}$, hence each adjacent gap Q_l, by at most c, so the difference between the two largest gaps moves by at most $2c$:

$$\big(\tilde{Q}_{[1]} - \tilde{Q}_{[2]}\big) - 2c \;\le\; Q_{[1]} - Q_{[2]} \;\le\; \big(\tilde{Q}_{[1]} - \tilde{Q}_{[2]}\big) + 2c.$$

If $\tilde{Q}_{[1]} - \tilde{Q}_{[2]} < \tau - 2c$, both runs abstain; if $\tilde{Q}_{[1]} - \tilde{Q}_{[2]} > \tau + 2c\,(\ge 2c)$, both runs pass the threshold and, by Lemma 3.1, select the same cut and the same leading set, hence return the same average. The outputs can thus differ only when $\tau - 2c \le \tilde{Q}_{[1]} - \tilde{Q}_{[2]} \le \tau + 2c$, and Theorem 3.3 bounds the probability of this event by concentrating every count $\tilde{N}_l$ around $p_l n_0$ (each within βn_0, so the gap difference is within $4\beta n_0$ of its population value $\mathcal{G}$).

Calibration with Φ_{TMGC}. Algorithms 1 and 4 apply unchanged when Φ_{TMV} is replaced by Φ_{TMGC}. In Algorithm 4, a calibration histogram is conservatively labelled vulnerable if some $\le c$ of its votes can be re-chosen so as to change the output; since re-choosing c votes moves the observed gap difference by at most $4c$ in total (removal and re-insertion), this holds iff the observed gap difference lies in $[\tau - 4c, \tau + 4c]$.

Remark B.1 (Choosing between Φ_{TMV} and Φ_{TMGC})**.** Neither operator dominates the other. Φ_{TMV} is preferable when the quantized histogram has a single dominant bin well separated from all alternatives: the rule is simpler and the certificate is driven by the single top-two count gap. Φ_{TMGC} is preferable when several leading bins form a stable cluster separated from the remaining bins by a unique maximal adjacent gap, since averaging the selected bin centers retains more magnitude information than returning a single bin. In practice we recommend selecting between them on a benign calibration set using the joint consistency, update-rate, and utility trade-off, as quantified in Appendix J.1.4.

C Practical Gradient Distribution and Evaluation of Theorem 3.2 and Theorem 3.3

This appendix explains how the gap distributions that illustrate Section 3.2 (Fig. 2) were obtained, fixes the notation for sources and votes, and evaluates the exponents of Theorems 3.2 and 3.3 on real gradient distributions.

Notation for sources and votes. To avoid conflating data volume with vote count, we write S for the number of sources, B for the number of votes contributed by each source per iteration, $n_{\text{vote}} = BS$ for the total number of votes entering one coordinate histogram, M for the number of malicious sources, and $c = MB$ for the adversarial vote budget; throughout this section we use the shorthand $n = n_{\text{vote}}$ and $c = \eta n$.

Practical gradient distribution. We estimate the probability mass p_l of distribution $\mathcal{D}$ on b_l using layer 19 of ResNet-20 trained on CIFAR-10 via SGD without any backdoor attack, averaging over 100

iterations of 10,000 per-sample gradients, and define the sorted sequence $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(m)}$ accordingly. Using these estimated values, we simulate $n_{\text{fig}} = 100$ synthetic votes per draw (each vote formed from 500 samples) and compute the gap $N_{(1)} - N_{(2)}$ from Theorem 3.2, repeating this procedure $5{,}000$–$10{,}000$ times to obtain the empirical distribution of $N_{(1)} - N_{(2)}$ as a histogram. As shown in Figure 2, we consider two gap distribution types, Type-A and Type-B, each in two variants: one more concentrated and one less concentrated. As stated in Theorem 3.2, it is suggested that we choose τ in an intermediate range (illustrated by the green bands in Figure 2) for both concentrated and less concentrated cases: doing so keeps the inconsistency probability small while maximizing the fraction of coordinates on which majority vote is applied (thereby preserving more information) for the Type-A regime.

Scope of this illustration. Figure 2 illustrates the induced gap law and its asymptotic concentration, and is deliberately separated from the runtime configuration used during training. Five distinct quantities appear in this paper and should not be conflated: (i) the number of *runtime votes* per coordinate histogram, $n = n_{\text{vote}} = 10$ in the image experiments, with $n_0 = 9$; (ii) the number of examples used to form each per-source gradient (the source mini-batch); (iii) the number of per-sample gradients used offline to estimate the masses p_l (100 iterations of $10{,}000$ gradients above); (iv) the number of *synthetic votes per illustrative draw*, $n_{\text{fig}} = 100$; and (v) the number of Monte Carlo repetitions, $5{,}000$–$10{,}000$. Plotted gaps can therefore exceed ten, which is a property of the illustration and not of the runtime operator.

Theorem evaluations. Table 3 instantiates the upper bounds of Theorem 3.2 and Theorem 3.3 with the gap exponents estimated from the gradient distributions of our main experiments (the margin parameter β is left free). These expressions report the *rate* of exponential decay in the number of benign votes n_0; they are numerically informative only when n_0 is large, and are vacuous in the $n_0 = 9$ regime of the image experiments, as discussed in Section 3.3.

Table 3: Evaluation of Theorem 3.2 and Theorem 3.3 across different deep learning tasks.

Experiment	Φ_{TMV}	Φ_{TMGC}
Data poisoning defense (Section 4.1)	$2\exp(\frac{-\beta^2 n_0}{2}) + (m-1)\exp(-0.0032\, n_0)$	$2m\exp(\frac{-\beta^2 n_0}{2})$
Memorization (Section 4.2)	$2\exp(\frac{-\beta^2 n_0}{2}) + (m-1)\exp(-0.0024\, n_0)$	$2m\exp(\frac{-\beta^2 n_0}{2})$
Copyright protection (Section 4.2)	$2\exp(\frac{-\beta^2 n_0}{2}) + (m-1)\exp(-0.0192\, n_0)$	$2m\exp(\frac{-\beta^2 n_0}{2})$

D Discussion of DiSI Machine Learning

The distinction between consistency robustness and Byzantine robustness is discussed in Section 3.1; the corresponding experiments are in Appendix J.1.3. Two further remarks follow.

Remark D.1 (Why not Median)**.** Following our reasoning, many other consistency-robust estimators can also be constructed to satisfy (2). One straightforward example is a threshold coordinate-wise median, which outputs the median only when the median bin has enough frequency to tolerate adversarial corruption A. Asymptotically, we can still prove an exponentially decaying inconsistency rate. In our preliminary experiments, however, the expected median-bin frequency was often too small, implying a high zeroing-out rate and a worse utility-trust tradeoff than threshold majority vote.

Remark D.2 (Convergence Analysis)**.** Compared to standard stochastic gradient descent (SGD), the robust consistency estimator Φ_{TMV} impacts the convergence of DiSI-SGD (Algorithm 1) twofold. When the significance threshold is not met (Line 11), no gradient descent occurs on that particular coordinate, leaving the model weight effectively frozen for the iteration. Conversely, when the threshold is surpassed (Line 9), the estimator may introduce additional bias compared to the true gradient mean. Consequently, developing more efficient estimators, with a higher update frequency and a smaller estimation bias, is a promising direction for limiting the effect on convergence. We provide the details of convergence analysis in Appendix F.

E PROOFS

This appendix collects the proofs of the results stated in Sections 2 and 3, in the order in which the results are used: the consistency lemma, the two operator theorems (with the lemma on the maximal cut), the trajectory composition, the non-vacuity of Φ_{TMV}, and the certified parameter selection.

E.1 PROOF OF LEMMA 2.3

Let

$$\mathcal{E} := \big\{\mathsf{Alg}(\mathsf{X}) = \mathsf{Alg}(\mathsf{R})\big\}, \qquad \mathcal{E}^c := \big\{\mathsf{Alg}(\mathsf{X}) \neq \mathsf{Alg}(\mathsf{R})\big\}.$$

By assumption, $\Pr(\mathcal{E}^c) \leq \Delta$. Define the event

$$\mathcal{A} := \big\{\mathsf{Alg}(\mathsf{R}) \in \mathcal{T}\big\}.$$

We relate $\{\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}\}$ to $\mathcal{A}$ via the agreement event $\mathcal{E}$. Observe that on $\mathcal{E}$, the two outputs coincide, hence

$$\mathcal{A} \cap \mathcal{E} \subseteq \big\{\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}\big\}.$$

Therefore,

$$\begin{aligned}\Pr\big(\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}\big) &\geq \Pr(\mathcal{A} \cap \mathcal{E}) = \Pr(\mathcal{A}) - \Pr(\mathcal{A} \cap \mathcal{E}^c) \\ &\geq \Pr(\mathcal{A}) - \Pr(\mathcal{E}^c) \geq \Pr\big(\mathsf{Alg}(\mathsf{R}) \in \mathcal{T}\big) - \Delta,\end{aligned}$$

which is the claim.

E.2 PROOF OF THEOREM 3.2

Given a distribution $\mathcal{D}$ defined over m elements $\{b_1, b_2, \cdots, b_m\}$ where p_l is the probability mass over b_l. We use $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(m)}$ to denote the sorted version of the probability mass, correspondingly upon $\{b_{(1)}, b_{(2)}, \cdots, b_{(m)}\}$. By Definition 2.1, the $n_0 = (1-\eta)n$ many datapoints in $\mathsf{R} = \{r_1, r_2, \cdots, r_{n_0}\}$ are i.i.d. selected from $\mathcal{D}$. We use $\tilde{N}_l$ to denote the number of samples in R equaling b_l, and similarly $\tilde{N}_{[1]} \geq \tilde{N}_{[2]} \geq \cdots \geq \tilde{N}_{[m]}$ denotes its sorted version.

The high level idea of this proof is twofold. First, we want to identify the probability that the top-1 and top-2 frequencies, i.e., $\tilde{N}_{[1]}$ and $\tilde{N}_{[2]}$ happen to be $b_{(1)}$ and $b_{(2)}$, respectively. Conditional on this event, we further upper bound the inconsistency rate of threshold majority vote Φ_{TMV} characterized by (3).

Lemma E.1. *Let $\mathcal{E}$ denote the event that the bin with the top-1 and top-2 probability mass, $p_{(1)}$ and $p_{(2)}$, also have the top-1 and top-2 count numbers $N_{[1]}$ and $N_{[2]}$, respectively, i.e.,*

$$\mathcal{E} := \big\{(1) = [1], (2) = [2]\big\} \tag{7}$$

Then,

$$\Pr(\mathcal{E}) \geq 1 - \Big((m-2)\exp\left(-\frac{n_0\big(p_{(2)} - p_{(3)}\big)^2}{2}\right) + \exp\left(-\frac{n_0\big(p_{(1)} - p_{(2)}\big)^2}{2}\right)\Big), \tag{8}$$

where $n_0 = (1-\eta)n$.

Proof. First, it is noted that (7) is equivalent to that all the following events simultaneously hold,

$$\mathcal{E}_1 := \{N_{(1)} - N_{(2)} \geq 0\} \quad \text{and} \quad \mathcal{E}_l := \{N_{(2)} - N_{(l)} > 0\} \quad \text{for} \quad l = 3, \cdots, m. \tag{9}$$

We separately study each case. For any pair (l, l'), define i.i.d. random variables $\{B_i\}_{i=1}^{n_0}$ by

$$B_i := \begin{cases} 1, & \text{if } r_i = b_{(l)}, \\ -1, & \text{if } r_i = b_{(l')}, \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\sum_{i=1}^{n_0} B_i = N_{(l)} - N_{(l')}, \qquad \mathbb{E}[B_i] = p_{(l)} - p_{(l')}, \qquad \mathbb{E}\left[\sum_{i=1}^{n_0} B_i\right] = n_0(p_{(l)} - p_{(l')}).$$

Since $B_i \in [-1, 1]$, by Hoeffding's inequality, once $p_{(l)} - p_{(l')} \geq 0$, we have

$$\Pr\big(N_{(l)} - N_{(l')} \leq 0\big) = \Pr\left(\sum_{i=1}^{n_0} B_i \leq 0\right) \leq \exp\left(-\frac{\big(n_0(p_{(l)} - p_{(l')})\big)^2}{2n_0}\right). \tag{10}$$

Since for any $l > 2$, $|p_{(2)} - p_{(l)}| \geq p_{(2)} - p_{(3)}$ due to the non-increasing order of $p_{(l)}$. Therefore, by union bound for the $(m-1)$ events in (9), (8) holds. □

Now, we focus on $\tilde{N}_{(1)} - \tilde{N}_{(2)}$. Following the identical reasoning and applying Hoeffding's inequality again, for any $t > 0$,

$$\Pr\big(|N_{(1)} - N_{(2)} - (p_{(1)} - p_{(2)})n_0| \geq t\big) \leq 2\exp\left(-\frac{t^2}{2n_0}\right). \tag{11}$$

Take $t = \beta n_0$. On the event (7), if $|\tau - (p_{(1)} - p_{(2)})n_0| \geq \beta n_0 + \eta n$, then the "borderline" event (3) implies $|\tilde{N}_{(1)} - \tilde{N}_{(2)} - (p_{(1)} - p_{(2)})n_0| \geq \beta n_0$, so by (11) its probability (jointly with (7)) is upper bounded by

$$\Pr\left(\tau - \eta n \leq \tilde{N}_{(1)} - \tilde{N}_{(2)} \leq \tau + \eta n\right) \leq 2\exp\left(-\frac{\beta^2 n_0}{2}\right). \tag{12}$$

Adding the failure probability of (7) via a union bound, we have

$$\begin{aligned}
&\Pr\left(\tau - \eta n \leq \tilde{N}_{[1]} - \tilde{N}_{[2]} \leq \tau + \eta n\right) \\
&\leq 2\exp\left(-\frac{\beta^2 n_0}{2}\right) + \Big((m-2)\exp\left(-\frac{n_0\big(p_{(2)} - p_{(3)}\big)^2}{2}\right) + \exp\left(-\frac{n_0\big(p_{(1)} - p_{(2)}\big)^2}{2}\right)\Big).
\end{aligned} \tag{13}$$

Heterogeneous sources. The argument above uses the benign votes only through Hoeffding's inequality for sums of independent bounded variables and the resulting concentration of each count $\tilde{N}_l$ around its mean. If the benign sources are independent but not identically distributed, with source i voting for bin b_l with probability $p_{i,l}$, then $\mathbb{E}\tilde{N}_l = \sum_i p_{i,l} = n_0\bar{p}_l$ and every step goes through verbatim with p_l replaced by the average mass $\bar{p}_l$; the same holds for Theorem 3.3, whose proof also uses only per-count concentration.

Replacement contamination. If at most c of the n votes are replaced rather than added (the reference has n benign votes and the target replaces c of them), each count changes by at most c in either direction, so the top-two gap changes by at most $2c$ and the adjacent-gap difference of Φ_{TMGC} by at most $4c$. The borderline bands therefore become $[\tau - 2c, \tau + 2c]$ and $[\tau - 4c, \tau + 4c]$, and Theorems 3.2 and 3.3 hold with c replaced by $2c$ (and n_0 replaced by n). Adding or removing samples inside c sources changes only those sources' votes and is an instance of this model.

E.3 Proof of Lemma 3.1

To differentiate the index order in $\mathcal{H}_s(\mathsf{X})$ and $\mathcal{H}_s(\mathsf{R})$, without loss of generality, for $\mathcal{H}_s(\mathsf{X})$, we assume $N_1 \geq N_2 \geq \cdots \geq N_m$ to denote the sorted version of the histogram $\mathcal{H}_s(\mathsf{X})$. Let k' be the index in $\mathcal{H}_s(\mathsf{X})$ with the maximal adjacent gap.

First, provided the gap condition $\tilde{Q}_{[1]} - \tilde{Q}_{[2]} > 2\eta n$ of Lemma 3.1, we will show $[1] = (k) = [1]' = (\hat{k})'$, i.e., the index of the maximal cut must be the same for both $\mathcal{H}_s(\mathsf{X})$ and $\mathcal{H}_s(\mathsf{R})$. It is noted that,

$$N_{k'} - N_{k'+1} \geq \tilde{N}_{(k)} - \tilde{N}_{(k+1)} - \eta n \tag{14}$$

since $|\mathsf{A}| \leq \eta n$, $N_l - \tilde{N}_l \geq 0$ with the addition of A, and

$$\sum_{l=1}^{m} |N_l - \tilde{N}_l| = \sum_{l=1}^{m} (N_l - \tilde{N}_l) \leq \eta n.$$

On the other hand, if either $k' \neq (k)$ or $k'+1 \neq (k+1)$, then let a be the right-adjacent index of k' in $\mathcal{H}_s(\mathsf{R})$ such that $\tilde{N}_{k'} \geq \tilde{N}_a$ and b be the left-adjacent index of $k'+1$ in $\mathcal{H}_s(\mathsf{R})$ such that

$\tilde{N}_b \geq \tilde{N}_{k'+1}$. Then, we have

$$N_{k'} - N_{k'+1} \leq \max\{\tilde{N}_{(k')} - \tilde{N}_a, \tilde{N}_{(b)} - \tilde{N}_{k'+1}\} + \eta n < \tilde{N}_{(k)} - \tilde{N}_{(k+1)} - 2\eta n + \eta n < \tilde{N}_{(k)} - \tilde{N}_{(k+1)} - \eta n. \tag{15}$$

In (15), we use the assumption of Lemma 3.1 that the maximal adjacent gap is strictly $2\eta n$ larger than the second largest one. (15) leads to a contradiction with (14).

Furthermore, still by (15), for any $l \leq k$ and any $l' > k$,

$$N_{(l)} - N_{(l')} \geq \tilde{N}_{(k)} - \tilde{N}_{(k+1)} - \eta n \geq 2\eta n - \eta n > 0. \tag{16}$$

Therefore, any index of the top k in $\mathcal{H}_s(\mathsf{R})$ cannot appear in the bottom $(m-k)$ indices in $\mathcal{H}_s(\mathsf{X})$, i.e., $\mathcal{H}_s(\mathsf{X})$ must share the leading k indices (ignoring orders) with $\mathcal{H}_s(\mathsf{R})$.

E.4 Proof of Theorem 3.3

The high level idea of this proof is first to show that, provided a sufficiently large number of samples $n_0 = (1-\eta)n$, each frequency $\tilde{N}_l$ will be close enough to $p_l n_0$ with high probability. Then, conditional on this event, we study the consistency rate with properly selected threshold τ to control the probability of the "borderline" scenario leading to inconsistency between R and X.

For each $l \in \{1, 2, \cdots m\}$, we define a Bernoulli variable B_i^l where $B_i^l = 1$ iff $r_i = b_l$, otherwise $B_i^l = 0$. Clearly, $\tilde{N}_l = \sum_{i=1}^{n_0} B_i^l$ and $\mathbb{E}[\tilde{N}_l] = p_l n_0$. By Hoeffding bound, for any $t = \beta n_0$,

$$\Pr\left(|\tilde{N}_l - p_l n_0| > \beta n_0\right) \leq 2\exp\left(-\frac{n_0\beta^2}{2}\right). \tag{17}$$

By union bound, now the probability of the event, when for each $\tilde{N}_l$ its distance to its expectation $p_l n_0$ is at most βn_0, can be upper bounded by

$$\Pr\left(|\tilde{N}_l - p_l n_0| \leq \beta n_0, \quad \forall l = 1, 2, \cdots, m\right) > 1 - 2m\exp\left(-\frac{n_0\beta^2}{2}\right). \tag{18}$$

In the following, still to differentiate the two orders of $\mathcal{H}_s(\mathsf{X})$ and $\mathcal{H}_s(\mathsf{R})$, respectively, without loss of generality, we rewrite $\mathcal{H}_s(\mathsf{X}) = (N_{(1)}, N_{(2)}, \cdots, N_{(m)})$ and $\mathcal{H}_s(\mathsf{R}) = (\tilde{N}_{(1)'}, \tilde{N}_{(2)'}, \cdots, \tilde{N}_{(m)'})$. Similarly, we define

$$Q_l = N_{(l)} - N_{(l+1)}, \quad Q_{[1]} \geq Q_{[2]} \geq \cdots \geq Q_{[m-1]}, \quad (k) = [1] \tag{19}$$

$$\tilde{Q}_l = \tilde{N}_{(l)} - \tilde{N}_{(l+1)}, \quad \tilde{Q}_{[1]'} \geq \tilde{Q}_{[2]'} \geq \cdots \geq \tilde{Q}_{[m-1]'}, \quad (k')' = [1]'. \tag{20}$$

Adding at most ηn votes changes every order statistic $N_{(l)}$, and hence every adjacent gap Q_l, by at most ηn; the largest and the second-largest gap therefore each move by at most ηn, so that

$$\left(\tilde{Q}_{[1]'} - \tilde{Q}_{[2]'}\right) - 2\eta n \leq \left(Q_{[1]} - Q_{[2]}\right) \leq \left(\tilde{Q}_{[1]'} - \tilde{Q}_{[2]'}\right) + 2\eta n. \tag{21}$$

If $\tilde{Q}_{[1]'} - \tilde{Q}_{[2]'} < \tau - 2\eta n$, then by (21) both $\Phi_{\text{TMGC}}(\mathsf{R})$ and $\Phi_{\text{TMGC}}(\mathsf{X})$ output $\perp$. If $\tilde{Q}_{[1]'} - \tilde{Q}_{[2]'} > \tau + 2\eta n$ $(\geq 2\eta n)$, both pass the threshold, and Lemma 3.1 guarantees that they select the same cut index and the same unordered set of leading bins, hence return the same average. Therefore, the inconsistency of $\Phi_{\text{TMGC}}(\cdot)$ can only occur in the following event

$$\tau - 2\eta n \leq \tilde{Q}_{[1]'} - \tilde{Q}_{[2]'} \leq \tau + 2\eta n, \tag{22}$$

Now, let (k^*) be the index which maximizes the adjacent difference of the distribution probability mass, i.e., $k^* = \arg\max_l p_{(l)} - p_{(l+1)}$. Let

$$\mathcal{G} = \min_{l \neq k^*} \left((p_{(k^*)} - p_{(k^*+1)}) - (p_{(l)} - p_{(l+1)})\right) n_0, \tag{23}$$

if it is further conditional on the event that $|\tilde{N}_l - p_l n_0| < \beta n_0$ for each l, we have that

$$\mathcal{G} - 4\beta n_0 \leq \tilde{Q}_{[1]'} - \tilde{Q}_{[2]'} \leq \mathcal{G} + 4\beta n_0. \tag{24}$$

Now, combining (24) and (22), it gives that when

$$\tau > \mathcal{G} + 4\beta n_0 + 2\eta n \quad \text{or} \quad \tau < \mathcal{G} - 4\beta n_0 - 2\eta n, \tag{25}$$

$\Phi_{\text{TMGC}}(\mathsf{R}) = \Phi_{\text{TMGC}}(\mathsf{X})$. With the failure rate of the event that $|\tilde{N}_l - p_l n_0| < \beta n_0$ for each l upper bounded in (18), we complete the proof.

E.5 Proof of Proposition 3.4

Proposition 3.4 (restated). Consider two coupled runs of Algorithm 1, one on $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$ and one on R, initialized at the same $w^{(0)}$ and sharing the same deterministic quantization, operator, and algorithmic randomness. Let $\theta^{\mathsf{X}}_{t,j}, \theta^{\mathsf{R}}_{t,j}$ be the parameters used for coordinate j at iteration t ($\perp$ if abstaining) and let $E_{t,j}$ be the event that the coordinate-j updates produced at iteration t coincide, conditional on the two trajectories being identical before that coordinate update. Suppose that, under this conditioning and uniformly over the adaptive contamination strategy, $\Pr(\theta^{\mathsf{X}}_{t,j} \neq \theta^{\mathsf{R}}_{t,j}) \leq \Gamma_{t,j}$ and $\Pr(E^c_{t,j},\ \theta^{\mathsf{X}}_{t,j} = \theta^{\mathsf{R}}_{t,j}) \leq \Delta_{t,j}$. Then $\Pr(w^{(T)}_{\mathsf{X}} = w^{(T)}_{\mathsf{R}}) \geq 1 - \sum_{t=1}^{T}\sum_{j=1}^{d}(\Gamma_{t,j} + \Delta_{t,j})$, and Lemma 2.3 transfers any reference guarantee to the final trained model with additional failure at most $\sum_{t,j}(\Gamma_{t,j} + \Delta_{t,j})$.

Proof. Let $\mathcal{G}_{t,j}$ denote the event that all coordinate updates before coordinate j of iteration t have been identical in the coupled runs, and let $F_{t,j} := \mathcal{G}_{t,j} \cap E^c_{t,j}$ be the event that the *first* disagreement occurs at (t, j). On $\mathcal{G}_{t,j}$ the two runs are at the same model state, and $E^c_{t,j} \subseteq \{\theta^{\mathsf{X}}_{t,j} \neq \theta^{\mathsf{R}}_{t,j}\} \cup \{E^c_{t,j},\ \theta^{\mathsf{X}}_{t,j} = \theta^{\mathsf{R}}_{t,j}\}$, so by assumption $\Pr(F_{t,j}) = \Pr(E^c_{t,j} \mid \mathcal{G}_{t,j}) \Pr(\mathcal{G}_{t,j}) \leq \Gamma_{t,j} + \Delta_{t,j}$. If all events $E_{t,j}$ hold, the two runs apply the same coordinate updates in the same order from the same initialization, so the trajectories and final weights coincide; conversely, any disagreement of the final weights has a first disagreeing coordinate update, so

$$\Pr\Big(w^{(T)}_{\mathsf{X}} \neq w^{(T)}_{\mathsf{R}}\Big) \leq \Pr\left(\bigcup_{t=1}^{T}\bigcup_{j=1}^{d} F_{t,j}\right) \leq \sum_{t=1}^{T}\sum_{j=1}^{d}\Pr(F_{t,j}) \leq \sum_{t=1}^{T}\sum_{j=1}^{d}(\Gamma_{t,j} + \Delta_{t,j}).$$

The events $F_{t,j}$ are disjoint, so the union bound in the middle is in fact an equality; the conditional bound is only ever used at the first disagreement, which is why later mismatches need no control. □

E.6 Proof of Proposition E.2

Proposition E.2 (Non-vacuous update rate of Φ_{TMV})**.** *Under the same notation as Theorem 3.2, suppose*

$$(p_{(1)} - p_{(2)})n_0 - \tau - \eta n \geq \beta n_0.$$

Then, for every adaptive adversarial contamination A *of size at most* ηn*, the target operator updates rather than abstains with probability at least* $1 - \Delta_{\mathrm{upd}}$*, where*

$$\Delta_{\mathrm{upd}} \leq \exp\left(-\frac{\beta^2 n_0}{2}\right) + (m-2)\exp\left(-\frac{n_0(p_{(2)} - p_{(3)})^2}{2}\right) + \exp\left(-\frac{n_0(p_{(1)} - p_{(2)})^2}{2}\right).$$

Proof. On the event that the top two empirical bins match the top two population bins and

$$\tilde{N}_{(1)} - \tilde{N}_{(2)} \geq (p_{(1)} - p_{(2)})n_0 - \beta n_0,$$

the proposition's margin condition implies

$$\tilde{N}_{(1)} - \tilde{N}_{(2)} \geq \tau + \eta n.$$

Adding at most ηn adversarial samples can increase any competing bin by at most ηn and cannot decrease the reference top bin count. Hence the target top-gap remains at least τ, so $\Phi_{\mathrm{TMV}}(\mathsf{R} \cup \mathsf{A}) \neq \perp$. The failure probability is bounded by the same top-bin ordering bound used in Theorem 3.2, plus Hoeffding's lower-tail bound for $\tilde{N}_{(1)} - \tilde{N}_{(2)}$, giving the stated Δ_{upd}. □

E.7 Proofs of Theorems 3.5 and I.1

Theorem 3.5 (validity). Fix $\theta \in \Theta$ and a coordinate. For $k \in [K]$ let $B_k \in \{0, 1\}$ indicate that the benign part of the k-th calibration histogram (the histogram of its n_0 benign votes) is vulnerable in the sense of Section 3.2, i.e., that adding some admissible A with $|\mathsf{A}| \leq c$ changes the output of Φ_θ. The calibration mini-batches are disjoint and drawn at a fixed model state, so $B_1, \ldots, B_K$ are i.i.d. Bernoulli with success probability $\Delta_\theta(\mathcal{D})$, and $V^\star_\theta := \sum_k B_k \sim \mathrm{Binomial}(K, \Delta_\theta(\mathcal{D}))$.

The label used by Algorithm 4 is conservative: a histogram is labeled vulnerable if *some* candidate benign histogram, obtained by deleting at most c of its votes, is vulnerable (Appendix I.1). The true

benign histogram is one of these candidates, so the label L_k satisfies $L_k \geq B_k$ pointwise, whatever the c adversarial calibration votes are, and $V_\theta = \sum_k L_k \geq V_\theta^\star$.

Let $U(v) = Q_{1-\alpha/J}\big(\mathrm{Beta}(v+1, K-v)\big)$ be the one-sided Clopper–Pearson upper confidence limit at level $1-\alpha/J$, so that $\Pr[\Delta_\theta(\mathcal{D}) \leq U(V_\theta^\star)] \geq 1-\alpha/J$ for the binomial count $V_\theta^\star$. Since $U(\cdot)$ is non-decreasing, $U_\theta = U(V_\theta) \geq U(V_\theta^\star)$, hence $\Pr[\Delta_\theta(\mathcal{D}) \leq U_\theta] \geq 1-\alpha/J$. A union bound over the J candidates gives $\Pr[\Delta_\theta(\mathcal{D}) \leq U_\theta \ \ \forall\theta \in \Theta] \geq 1-\alpha$. On this event, every θ with $U_\theta \leq \rho$ satisfies $\Delta_\theta(\mathcal{D}) \leq \rho$, regardless of which certified candidate a (data-dependent) selection rule returns, because the event is defined before selection. Finally, if $V_\theta = 0$ then $U(0) = 1-(\alpha/J)^{1/K}$, which is at most ρ iff $K \geq \log(\alpha/J)/\log(1-\rho) = \log(J/\alpha)/\log\frac{1}{1-\rho}$.

Theorem I.1, case (a). With a shared protected calibration stream, both executions receive the same block C_t (identical histograms of n_0 benign votes), compute the same counts V_θ and bounds U_θ, and apply the same deterministic selection rule to the same certified set; they therefore return the same candidate (or $\perp$) with probability one, i.e., $\Gamma_{t,j} = 0$. Validity in this case needs no conservative label: the histograms consist of benign votes only, so V_θ is exactly the binomial count $V_\theta^\star$ of the validity proof. The reference procedure coupled with the target is $\mathsf{Alg}(\mathsf{R}; \mathsf{C}_{1:T})$, i.e., the reference trained with the same calibration protocol, so the baseline trust probability $1-\delta_0$ of Section 1.1 refers to this procedure.

Theorem I.1, case (b). In the coupled executions of Proposition 3.4, the k-th calibration histograms of the target and of the reference share the same n_0 benign votes and differ only by the c adversarial votes, so their top-two gaps satisfy $|g_k - \tilde{g}_k| \leq c$ (each order statistic of the counts moves by at most c). Both executions apply the same rule and label histogram k vulnerable iff its gap lies in $[\tau-2c, \tau+2c]$. The labels differ only if exactly one of $g_k, \tilde{g}_k$ lies in this band, which forces $\tilde{g}_k$ to lie within c of one of its edges, i.e., $\tilde{g}_k \in S_\theta := [\tau-3c, \tau-c] \cup [\tau+c, \tau+3c]$. On the event that $\tilde{g}_k \notin S_\theta$ for all $k \in [K]$ and all $\theta \in \Theta$, the counts V_θ, the bounds U_θ, and hence the certified set $\Theta^\star$ coincide in the two executions; since the preference order is fixed before training, the returned candidate (or $\perp$) coincides as well. A union bound over the K histograms and the J candidates gives $\Pr(\theta_{\mathsf{X}} \neq \theta_{\mathsf{R}}) \leq \sum_{\theta\in\Theta}\sum_{k=1}^{K}\Pr(\tilde{g}_k \in S_\theta) = K\sum_{\theta\in\Theta}\delta_\theta(\mathcal{D})$. The inclusion is pointwise in the adversarial votes, so the bound holds for every adversary. For Φ_{TMGC} the gap difference moves by at most $2c$ (Appendix B), which gives the same statement with c replaced by $2c$. Finally, $\delta_\theta(\mathcal{D})$ is bounded exactly as in Theorem 3.2: if $|\tau - (p_{(1)} - p_{(2)})n_0| \geq \beta n_0 + 3c$, the flanking bands lie at distance at least βn_0 from the mean gap, so $\delta_\theta(\mathcal{D}) \leq 2e^{-\beta^2 n_0/2} + (m-2)e^{-n_0(p_{(2)}-p_{(3)})^2/2} + e^{-n_0(p_{(1)}-p_{(2)})^2/2}$, the last two terms accounting for a misordering of the top two bins as in Theorem 3.2.

Composition with adaptive selection. With Algorithm 4 inside Algorithm 1, the per-coordinate agreement event of Proposition 3.4 is the intersection of three events: the two executions select the same parameter (fails with probability at most $\Gamma_{t,j}$, which is 0 in case (a) and at most ρ_{sel} in case (b), Theorem I.1), the confidence bounds cover (fails with probability at most α by Theorem 3.5), and the update on the fresh mini-batch is not in the borderline regime of the selected parameter (fails with probability at most ρ on the coverage event, since the update mini-batch is independent of the calibration data and of the selection). Hence $\Delta_{t,j} \leq \rho + \alpha$ and $\Pr(w_{\mathsf{X}}^{(T)} \neq w_{\mathsf{R}}^{(T)}) \leq dT(\rho+\alpha)$ in case (a) and $\leq dT(\rho+\alpha+\rho_{\mathrm{sel}})$ in case (b).

F Convergence Analysis of DiSI Training

We establish convergence analysis for DiSI-based Stochastic Gradient Descent (Algorithm 1) for general non-convex optimization in Section F.1 and later in Section F.2 we present the specialization to the sign quantization setting with coordinate-wise majority vote.

Scope of this appendix. The two parts play different roles, and we state their status explicitly. Section F.1 is a *generic* conditional descent theorem: given an update direction whose bias relative to the true gradient satisfies a second-moment bound together with an alignment condition, it yields the usual non-convex convergence rate. Section F.2 verifies these conditions only in the analytically tractable special case of sign aggregation, i.e., one-bit quantization with coordinate-wise majority vote. Taken together they do *not* constitute a complete convergence proof for general multi-level

Φ_{TMV} or Φ_{TMGC}, since the alignment condition is not established for general quantization; a tight general-quantization analysis is left to future work.

F.1 Convergence analysis for general quantization

Consider the iterative scheme

$$x_{t+1} = x_t - \eta_t v_t,$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is possibly non-convex, $g_t := \nabla F(x_t)$, and v_t is the server update obtained from coordinate-wise majority vote over quantized client gradients.

Let F_t denote the filtration up to round t, and write

$$\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathsf{F}_t].$$

Assumption F.1 (Smoothness)**.** The objective F is L-smooth, i.e.,

$$F(y) \le F(x) + \langle \nabla F(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2, \qquad \forall x, y \in \mathbb{R}^d.$$

Assumption F.2 (Lower boundedness)**.** There exists $F_* > -\infty$ such that

$$F(x) \ge F_*, \qquad \forall x \in \mathbb{R}^d.$$

Assumption F.3 (Pseudo-gradient alignment)**.** There exist constants $\mu > 0$ and $\varepsilon \ge 0$ such that for all t,

$$\mathbb{E}_t\big[\langle g_t, v_t\rangle\big] \;\ge\; \mu\|g_t\|_1 - \varepsilon.$$

Assumption F.4 (Second moment bound)**.** There exists $G > 0$ such that for all t,

$$\mathbb{E}_t\|v_t\|_2^2 \le G^2.$$

Theorem F.5 (Generic non-convex descent bound)**.** *Under Assumptions F.1– F.4,*

$$\sum_{t=0}^{T-1} \eta_t\, \mathbb{E}\|g_t\|_1 \;\le\; \frac{F(x_0) - F_*}{\mu} + \frac{LG^2}{2\mu}\sum_{t=0}^{T-1}\eta_t^2 + \frac{\varepsilon}{\mu}\sum_{t=0}^{T-1}\eta_t.$$

Consequently,

$$\frac{\sum_{t=0}^{T-1}\eta_t\,\mathbb{E}\|g_t\|_1}{\sum_{t=0}^{T-1}\eta_t} \;\le\; \frac{F(x_0) - F_*}{\mu\sum_{t=0}^{T-1}\eta_t} + \frac{LG^2}{2\mu}\cdot\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} + \frac{\varepsilon}{\mu}.$$

In particular, for a constant stepsize $\eta_t \equiv \eta$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|g_t\|_1 \;\le\; \frac{F(x_0) - F_*}{\mu\eta T} + \frac{LG^2\eta}{2\mu} + \frac{\varepsilon}{\mu}.$$

Proof. By L-smoothness with $y = x_{t+1} = x_t - \eta_t v_t$,

$$F(x_{t+1}) \le F(x_t) + \langle g_t, -\eta_t v_t\rangle + \frac{L}{2}\|\eta_t v_t\|_2^2.$$

Hence

$$F(x_{t+1}) \le F(x_t) - \eta_t\langle g_t, v_t\rangle + \frac{L\eta_t^2}{2}\|v_t\|_2^2.$$

Taking conditional expectation with respect to F_t,

$$\mathbb{E}_t[F(x_{t+1})] \le F(x_t) - \eta_t\mathbb{E}_t[\langle g_t, v_t\rangle] + \frac{L\eta_t^2}{2}\mathbb{E}_t\|v_t\|_2^2.$$

Using the alignment and second-moment assumptions,

$$\mathbb{E}_t[F(x_{t+1})] \le F(x_t) - \eta_t(\mu\|g_t\|_1 - \varepsilon) + \frac{L\eta_t^2}{2}G^2.$$

Rearranging,

$$\mu\eta_t\|g_t\|_1 \le F(x_t) - \mathbb{E}_t[F(x_{t+1})] + \varepsilon\eta_t + \frac{L\eta_t^2}{2}G^2.$$

Taking full expectation,

$$\mu\eta_t \,\mathbb{E}\|g_t\|_1 \le \mathbb{E}[F(x_t)] - \mathbb{E}[F(x_{t+1})] + \varepsilon\eta_t + \frac{L\eta_t^2}{2}G^2.$$

Summing from $t = 0$ to $T - 1$,

$$\mu \sum_{t=0}^{T-1} \eta_t \,\mathbb{E}\|g_t\|_1 \le \sum_{t=0}^{T-1} \Big(\mathbb{E}[F(x_t)] - \mathbb{E}[F(x_{t+1})]\Big) + \varepsilon \sum_{t=0}^{T-1} \eta_t + \frac{LG^2}{2} \sum_{t=0}^{T-1} \eta_t^2.$$

The first sum telescopes:

$$\sum_{t=0}^{T-1} \Big(\mathbb{E}[F(x_t)] - \mathbb{E}[F(x_{t+1})]\Big) = F(x_0) - \mathbb{E}[F(x_T)].$$

By lower boundedness, $\mathbb{E}[F(x_T)] \ge F_*$, so

$$F(x_0) - \mathbb{E}[F(x_T)] \le F(x_0) - F_*.$$

Therefore,

$$\mu \sum_{t=0}^{T-1} \eta_t \,\mathbb{E}\|g_t\|_1 \le F(x_0) - F_* + \varepsilon \sum_{t=0}^{T-1} \eta_t + \frac{LG^2}{2} \sum_{t=0}^{T-1} \eta_t^2.$$

Dividing by μ proves the first claim. The remaining inequalities follow immediately by dividing by $\sum_{t=0}^{T-1} \eta_t$, and then specializing to the constant-stepsize case. □

F.2 Specialization to sign quantization with coordinate-wise majority vote

Assume now that each client sends one sign bit per coordinate, and that the server returns

$$v_{t,i} \in \{-1, +1\}$$

via coordinate-wise majority vote.

Define the coordinate-wise vote error probability

$$p_{t,i} := \Pr\big(v_{t,i} \ne \operatorname{sign}(g_{t,i}) \mid \mathsf{F}_t\big).$$

Lemma F.6. *For every coordinate i,*

$$\mathbb{E}_t[g_{t,i} v_{t,i}] = (1 - 2p_{t,i})\,|g_{t,i}|.$$

Proof. If $g_{t,i} = 0$, both sides are zero.

Suppose $g_{t,i} \ne 0$. Then

$$g_{t,i} v_{t,i} = |g_{t,i}| \operatorname{sign}(g_{t,i}) v_{t,i}.$$

Now $\operatorname{sign}(g_{t,i}) v_{t,i} = +1$ when the vote is correct and -1 when the vote is incorrect. Therefore

$$\mathbb{E}_t[\operatorname{sign}(g_{t,i}) v_{t,i}] = (1 - p_{t,i}) - p_{t,i} = 1 - 2p_{t,i}.$$

Multiplying by $|g_{t,i}|$ yields

$$\mathbb{E}_t[g_{t,i} v_{t,i}] = (1 - 2p_{t,i})\,|g_{t,i}|.$$

□

Corollary F.7. *If for some $p < 1/2$,*

$$p_{t,i} \le p, \qquad \forall t, i,$$

then

$$\mathbb{E}_t[\langle g_t, v_t \rangle] \ge (1 - 2p)\|g_t\|_1.$$

Hence Assumption F.3 holds with

$$\mu = 1 - 2p, \qquad \varepsilon = 0.$$

Moreover, since $v_t \in \{-1, +1\}^d$,

$$\|v_t\|_2^2 = d,$$

so Assumption F.4 holds with $G^2 = d$. Therefore,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|g_t\|_1 \le \frac{F(x_0) - F_*}{(1 - 2p)\eta T} + \frac{Ld\,\eta}{2(1 - 2p)}.$$

Proof. By the lemma,

$$\mathbb{E}_t[\langle g_t, v_t\rangle] = \sum_{i=1}^{d}(1-2p_{t,i})|g_{t,i}| \geq (1-2p)\sum_{i=1}^{d}|g_{t,i}| = (1-2p)\|g_t\|_1.$$

The second-moment statement follows from

$$\|v_t\|_2^2 = \sum_{i=1}^{d} v_{t,i}^2 = d.$$

The convergence bound then follows directly from Theorem F.5. □

G Reference Construction, Baseline Risk, and Imperfect References

This appendix expands the discussion in Section 2 on how a reference is chosen and what part of the guarantee it carries.

Decomposition of the guarantee. Let R be the reference data and $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$ the target. If the reference is safe for the trust event $\mathcal{T}$,

$$\Pr[\mathsf{Alg}(\mathsf{R}) \in \mathcal{T}] \;\geq\; 1 - \delta_0,$$

and the algorithm is consistency-robust,

$$\Pr[\mathsf{Alg}(\mathsf{X}) \neq \mathsf{Alg}(\mathsf{R})] \;\leq\; \Delta,$$

then Lemma 2.3 gives

$$\Pr[\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}] \;\geq\; 1 - \delta_0 - \Delta.$$

Reference construction therefore determines the baseline risk δ_0, whereas DiSI controls the additional influence risk Δ. This requirement is not unique to DiSI: any argument that transfers an application-level property from a reference, including DaSI (Xiao et al., 2025), NAF-based copyright protection (Vyas et al., 2023), and DP (Dwork et al., 2006a), must justify the baseline property before indistinguishability is applied. Our claim is explicitly conditional, and DiSI does not by itself construct a safe reference or repair an unsafe baseline.

Systematic references per application. A reference only needs to be safe for the *specific* event $\mathcal{T}$, which makes the construction concrete in each of our settings.

- **Sample-specific memorization and membership inference.** For a target sample x, take $\mathsf{R} = \mathsf{X} \setminus x$; the two membership worlds are exactly R and $\mathsf{R} \cup x$.
- **Copyright and training-example copying.** For a protected example, artist, or collection A, take the exclusion reference $\mathsf{R} = \mathsf{X} \setminus \mathsf{A}$, i.e., the model that never accessed the protected material.
- **Data poisoning.** If an unknown source $i^\star$ is malicious, the safe reference is $\mathsf{R} = \mathsf{X} \setminus \mathsf{X}_{i^\star}$. The learner need not identify $i^\star$: the consistency guarantee is uniform over the identity and the content of the admissible malicious source, so only the reference that excludes the actual malicious source has to be safe.

Imperfect references. Suppose an additional fraction ρ of the nominal reference is conservatively treated as adversarial. The effective contamination rate becomes

$$\eta_{\text{eff}} \;=\; \eta + (1-\eta)\rho,$$

and the entire consistency analysis of Section 3 can be rerun with η_{eff} in place of η, provided $\eta_{\text{eff}} < 1/2$. If the reference is unsafe in a way that cannot be absorbed this way, its pre-existing failures simply increase δ_0; DiSI then still bounds the *additional* degradation caused by A, but cannot certify the absolute risk.

An illustration from the text-backdoor experiments. Table 2 makes this distinction visible. On the jailbreaking task the original model already exhibits 42.62% ASR, so the reference itself carries a substantial δ_0. Unprotected fine-tuning raises the ASR to 79.68%, while Φ_{TMV} and Φ_{TMGC} limit it to 44.76% and 49.29% respectively. DiSI thus prevents the controlled subset from substantially worsening the baseline, but it cannot by itself turn an unsafe reference into a near-zero-ASR model.

DiSI is accordingly best described as a distribution-specific *trust-transfer* framework rather than a universal reference-construction framework.

H From Local Consistency to a Global Certificate: Further Discussion of Section 3.3

This appendix further discusses Section 3.3. There, coupling the target and reference runs (Proposition 3.4) bounds the probability that the final trajectories differ by $\sum_{t,j} \Delta_{t,j} \le dT\Delta$, and the source-count condition of Section 3.3 turns the exponential decay of Δ into a sufficient number of benign votes. We do not repeat that derivation here; instead, we explain why this union bound is conservative, what it implies for our experiments, and how it extends to many protected sources and to stochastic source sampling.

Why the bound is conservative. The union bound is loose in two separate directions. *Across coordinates*, the current proof certifies the d coordinates separately; a tighter accountant could define a whole-update event on the $n \times d$ source-gradient matrix, preserving the shared identity of the malicious source across coordinates and possibly avoiding the separate factor of d. *Across iterations*, the bound treats any mismatch at any of the T iterations as a failure of the full trajectory-equality event, whereas in optimization a mismatch may be small, transient, later cancelled, or nearly orthogonal to the attack objective; a tighter analysis should track cumulative update discrepancy or accumulated adversarial drift rather than exact equality.

Consequence for our experiments. This is why the worst-case exact-trajectory bound is vacuous in the deliberately difficult regime of Section 4.1, with only ten sources and 10% contamination. We do not claim a non-vacuous formal certificate there. What that stress test shows is that DiSI remains practically effective in this regime, suppressing ASR to roughly 0–8% across seven image attacks while retaining much higher clean accuracy than the noise-based baselines, and that the measured consistency improves quickly with the source count: in a matched Blended/CIFAR-10 experiment with one malicious source, mean coordinate agreement rises from 98.31% at $n = 10$ to 99.99998594% at $n = 100$. Formal certificates and empirical diagnostics are reported separately throughout this paper.

H.1 Simultaneous guarantees over many protected sources

Section 2 allows the data distribution $\mathcal{D}_{\perp x}$ to vary with the protected target x. Two cases must be distinguished, and only the second requires a union bound.

Fixed-target guarantee. For a single fixed target, one analyzes $\mathsf{R} \sim \mathcal{D}^{n_0}$, $\mathsf{A} = \mathsf{A}(\mathsf{R})$, and $\mathsf{X} = \mathsf{R} \cup \mathsf{A}$. No union bound over targets is needed, and different targets may use different $\mathcal{D}_i$; these remain separate counterfactual guarantees.

Simultaneous guarantee. To protect all targets on one realized dataset, all events must be defined on one joint probability space. With shared algorithmic randomness,

$$Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \mathcal{D}, \qquad \mathsf{X} = (Z_1, \dots, Z_n), \qquad \mathsf{R}_i = \mathsf{X} \setminus Z_i, \qquad E_i = \{\mathsf{Alg}(\mathsf{X}) = \mathsf{Alg}(\mathsf{R}_i)\},$$

so that all E_i are events on the same draw $Z_{1:n}$ and

$$\Pr\Big[\bigcap_{i=1}^{n} E_i\Big] \;\ge\; 1 - \sum_{i=1}^{n} \Pr(E_i^c).$$

Independence among the E_i is not required; a common i.i.d. or exchangeable population model is one sufficient way to place them on a single space.

More sources do not degrade the guarantee. The factor n above does not mean that protecting more sources is intrinsically harder, because the per-target rate Δ_n also improves with n: each leave-one-source-out reference contains $n - 1$ source votes, so under the same stable-margin condition and for a distribution-specific exponent $I > 0$,

$$\Delta_n \le \exp\big[-(n-1)I + O(\log n)\big] \quad \Longrightarrow \quad n\Delta_n \le n \exp\big[-(n-1)I + O(\log n)\big] \longrightarrow 0.$$

Under a common i.i.d. or exchangeable source model with stable histogram geometry, the linear growth in the number of protected events is therefore dominated by the exponential improvement of each leave-one-source-out certificate, so a larger source count improves rather than worsens the simultaneous guarantee.

H.2 Stochastic source sampling

The instantiation in Section 3 aggregates all sources at every iteration, and the timing comparison in Appendix J.1 is correspondingly a comparison of all-source DiSI-SGD with all-source distributed SGD, not with the cheapest source-subsampled SGD. A clean stochastic variant uses random reshuffling without replacement: with n sources and s sources per iteration, partition the sources at the start of epoch e into disjoint sets $B_{e,1}, \ldots, B_{e,n/s}$ of size s and apply Φ_{TMV} or Φ_{TMGC} only within the current set.

For a fixed protected source i, couple the target and leave-one-source-out executions using the same partition and randomness. Conditional on prior agreement, if $i \notin B_{e,r}$ then both executions use exactly the same selected sources and produce the same update, so only the unique mini-batch containing i can introduce disagreement. If the conditional inconsistency probability when source i participates is at most Δ_s, then after E epochs

$$\Pr[\text{any inconsistency attributable to source } i] \;\le\; E\Delta_s, \qquad \Pr\Big[\bigcap_{i=1}^{n} E_i^{\text{train}}\Big] \;\ge\; 1 - nE\Delta_s.$$

For example, with $n = 50{,}000$ sources and $s = 1{,}000$ per iteration there are 50 iterations per epoch, but each source participates in only one of them, so its influence risk composes once per epoch rather than across all 50 steps. The trade-off is that Δ_s may exceed the all-source rate, since each histogram then contains only s votes.

Does source sampling introduce bias? Two separate questions must be kept apart. For the *influence guarantee*, an unselected source has zero influence on that iteration, provided the partition is precommitted independently of source contents and is shared by the target/reference coupling; the certificate is charged only when the protected source participates. For *optimization*, uniform sampling is unbiased for the linear mean,

$$\mathbb{E}_B\Big[\frac{1}{s}\sum_{i\in B} g_i\Big] = \frac{1}{n}\sum_{i=1}^{n} g_i,$$

but in general $\mathbb{E}_B[\Phi(g_i : i \in B)] \neq \Phi(g_i : i \in [n])$, so the sampled DiSI update need not be an unbiased estimator of the population gradient. This affects utility and convergence but is logically separate from the influence certificate. Stochastic-source DiSI therefore creates a three-way trade-off among computation (smaller s reduces per-step cost), trustworthiness (smaller s weakens histogram concentration and may increase Δ_s or the abstention rate), and optimization (nonlinear source sampling may increase estimation bias). Across one epoch the number of source-gradient evaluations is still n when every source appears once, with DiSI additionally performing one histogram aggregation for each of the n/s source mini-batches. A full convergence analysis of this variant is left to future work.

I Certified Accounting of the Inconsistency Rate: Further Discussion of Section 3.4

Section 3.4 presents the high-level test–select–train procedure (Algorithm 4) under the DiSI training procedure (Algorithm 1). This appendix provides the details: how (q, τ) can be selected with a finite-sample certificate when $\mathcal{D}$ is an unknown black box.

I.1 TEST–SELECT–TRAIN CERTIFICATION FOR UNKNOWN $\mathcal{D}$

Algorithm 4 Test–Select–Train (coordinate j, iteration t)

Input: model state $w^{(t-1)}$; coordinate j; candidates Θ_j, $J = |\Theta_j|$; local risk budget ρ; confidence $1-\alpha$; calibration votes from (a) a shared protected stream C_t or (b) K calibration mini-batches per source, disjoint from the update mini-batch; a deterministic selection rule (case (a)) or a preference order fixed before training (case (b)).

Compute the clipped calibration gradients at $w^{(t-1)}$
for each $\theta = (q, \tau) \in \Theta_j$ **do**
  Quantize coordinate j at resolution q; form K histograms $H_\theta^{(1)}, \dots, H_\theta^{(K)}$
  $V_\theta \leftarrow \#\{k : H_\theta^{(k)} \text{ vulnerable for } \Phi_\theta\}$; $U_\theta \leftarrow Q_{1-\alpha/J}\big(\mathrm{Beta}(V_\theta + 1, K - V_\theta)\big)$
end for
return the deterministic-rule maximizer among $\{\theta : U_\theta \le \rho\}$ (case (a)) or the first such θ in the fixed order (case (b)); $\perp$ if none

The procedure is summarized in Algorithm 4; Theorem 3.5 states what it certifies for each candidate, and the following theorem addresses the selection itself. Case (a) refers to calibration votes from a shared protected calibration stream C_t (drawn from the benign population at the common model state, independent of the update mini-batch, identical in the coupled target and reference executions, and not modifiable by A), and case (b) to calibration votes from the sources' own mini-batches, which contain up to c adversarial votes; both are defined precisely in items (a)–(c) below.

Theorem I.1 (Selection stability)**.** *In the setting of Theorem 3.5, let the two coupled executions of Proposition 3.4 both run Algorithm 4. In case (a), any deterministic selection rule returns the same candidate in both executions, so $\Gamma_{t,j} = 0$. In case (b), returning the first certified candidate in a preference order fixed before training gives $\Gamma_{t,j} \le \rho_{\mathrm{sel}} := K \sum_{\theta\in\Theta} \delta_\theta(\mathcal{D})$, where $\delta_\theta(\mathcal{D})$ is the probability that a benign calibration gap lies in the flanking bands $[\tau - 3c, \tau - c] \cup [\tau + c, \tau + 3c]$ ($c \to 2c$ for Φ_{TMGC}). Consequently, Proposition 3.4 gives $dT(\rho + \alpha)$ in case (a) and $dT(\rho + \alpha + \rho_{\mathrm{sel}})$ in case (b).*

The benign distribution $\mathcal{D}$ need not be known analytically: Φ_{TMV} operates only on observed source-level histograms. For a fixed parameter pair $\theta = (q, \tau)$, define

$$\Delta_\theta \;=\; \sup_{\mathsf{A}} \Pr_{\mathsf{R}\sim\mathcal{D}^{n-c}} \big[\Phi_\theta(\mathsf{R}\cup\mathsf{A}) \neq \Phi_\theta(\mathsf{R})\big],$$

and let ρ be the target local risk. The following procedure selects θ with $\Delta_\theta \le \rho$ at a prescribed confidence level, using only observable data.

(a) **Construct test histograms.** Calibration votes are drawn at the current model state $w^{(t-1)}$ in one of two ways. *(a) Shared protected calibration stream:* an auxiliary block C_t of K independent samples from the benign population, in the same source format as the training data, independent of the update mini-batch, identical in the coupled target and reference executions, and not modifiable by the controlled sources (it may be visible to them); each sample yields one histogram of n_0 benign votes. *(b) Self-calibration:* each source draws K disjoint calibration mini-batches (matching the training mini-batch size and disjoint from its update mini-batch); for each coordinate, source i computes one quantized vote from mini-batch k, and the coordinator forms $H^{(k)} = \mathrm{Hist}(x_1^{(k)}, \dots, x_n^{(k)})$, so the K histograms each contain n votes of which at most c are malicious. For instance, $n = 50$, $c = 1$, mini-batch size 10, and 5,280 test examples per source give $K = 528$ histograms.

(b) **Test candidate parameters.** In case (a) a histogram is vulnerable iff adding at most c arbitrary votes can change the output, i.e., for Φ_{TMV} iff its gap lies in $[\tau - c, \tau + c]$. In case (b), for each observed histogram h, consider every candidate benign histogram obtained by deleting at most c votes, and mark h vulnerable if any such candidate can have its output changed by adding at most c arbitrary votes. Deleting c votes moves the top-two gap by at most c and the subsequent insertion by another c, so for Φ_{TMV} (up to the fixed tie-breaking rule) h is labelled vulnerable iff its observed gap satisfies $\tau - 2c \le h^{(1)} - h^{(2)} \le \tau + 2c$; the analogous band for Φ_{TMGC} is $[\tau - 4c, \tau + 4c]$ on the observed gap difference (Appendix B). The label is monotone: whenever the true benign histogram is vulnerable (in the sense of Section 3.2), h is labelled vulnerable, so the count below is an upper bound on the number of vulnerable benign histograms regardless of how the malicious calibration votes are chosen. Let $V_{q,\tau}$ be the number of vulnerable test histograms. Its exact one-sided Clopper–Pearson upper bound is

$$U_{\alpha/J}(q, \tau) \;=\; Q_{1-\alpha/J}\big(\mathrm{Beta}(V_{q,\tau} + 1,\; K - V_{q,\tau})\big),$$

where Q_γ denotes the γ-quantile of the stated Beta distribution and J is the number of candidate pairs; we set $U_{\alpha/J} = 1$ when $V_{q,\tau} = K$. The Bonferroni correction by J ensures that all candidate bounds hold simultaneously with probability at least $1 - \alpha$.

(c) **Select and train.** Among the candidates satisfying $U_{\alpha/J}(q, \tau) \le \rho$, return the maximizer of a deterministic utility rule with fixed tie-breaking in case (a), or the first one in a preference order fixed before training in case (b) (Section I.2 explains why the order must not depend on the calibration data in that case), and apply the selected operator to the update mini-batch of the current iteration, which is disjoint from the calibration mini-batches. Because the vote distribution depends on the model state, the procedure is repeated at every iteration of Algorithm 1.

If no vulnerable histogram is observed, it suffices to take

$$U_{\alpha/J} = 1 - \left(\frac{\alpha}{J}\right)^{1/K} \qquad \Longleftrightarrow \qquad K \;\ge\; \frac{\log(\alpha/J)}{\log(1-\rho)}.$$

At 95% overall confidence with $J = 10$ candidates and $\rho = 1\%$, $K = 528$ histograms suffice. This certifies the local risk at a fixed coordinate and a fixed conditional model state. Composition follows Proposition 3.4: on the event that all certificates hold, every selected operator has local rate at most ρ; the confidence failures (probability at most α per coordinate and iteration) and the selection-instability failures ($\Gamma_{t,j} = 0$ in case (a), at most ρ_{sel} in case (b); Theorem I.1) are added to the union bound, giving $\Pr(w_{\mathsf{X}}^{(T)} \neq w_{\mathsf{R}}^{(T)}) \le dT(\rho+\alpha)$ and $dT(\rho+\alpha+\rho_{\mathrm{sel}})$, respectively; since K grows only logarithmically in $1/\alpha$, a smaller per-step α is inexpensive. The dependence on ρ is not: $K \approx \log(J/\alpha)/\rho$ for small ρ, and under the coordinate-wise exact-trajectory composition the local budgets must satisfy $dT(\rho+\alpha) \le \rho_{\mathrm{glob}}$, i.e., $\rho \lesssim \rho_{\mathrm{glob}}/(dT)$. For the illustration of Section 3.3 ($d = 10^7$, $T = 5{,}000$, $\rho_{\mathrm{glob}} = 0.01$) this gives $\rho \approx 10^{-13}$ and $K \approx 10^{14}$ calibration histograms per coordinate and iteration, which is not implementable. The finite-sample certificate is therefore a certificate for a fixed local decision; a practical global certificate requires the whole-update or drift-based accountant discussed in Section 5, under which ρ no longer needs to scale like $1/(dT)$. Why the present composition is conservative is discussed in Appendix H.

I.2 Coordinate-wise, per-iteration accounting

What the confidence bound certifies. The Clopper–Pearson quantity does not bound the number of contaminated histogram bins; the threat model already assumes that at most c of the n source votes are malicious. It upper-bounds the probability that a *fresh* histogram is vulnerable, i.e., that some admissible manipulation of at most c source votes can change the DiSI output. As an example, consider Φ_{TMV} at one coordinate with $n = 10$ sources, $c = 1$, and $\tau = 2$. If the benign histogram has top-two counts 5 and 3, the gap is 2, which lies in $[\tau - c, \tau + c] = [1, 3]$, so a single malicious vote can move the operator across the decision boundary and the histogram is vulnerable. If the benign gap lies sufficiently far outside this interval, no admissible one-source manipulation can change the decision.

Candidate set and certification. At iteration ℓ and coordinate j, let $\Theta_j = Q_j \times T_\tau$ with, for example,

$$Q_j = \{0.001,\ 0.002,\ 0.005\}, \qquad T_\tau = \{1, 2, 3, 4\}, \qquad J = |\Theta_j| = 12.$$

At the current model state each source supplies K fresh calibration mini-batches and one independent update mini-batch. Calibration index $k \in [K]$ produces one $n \times d$ source-gradient matrix whose j-th column induces the histogram for coordinate j; the same matrices are shared across all d coordinates and all J candidates, so the search does *not* require KdJ backward passes. For each $\theta \in \Theta_j$, let $V_{\ell,j}(\theta)$ be the number of calibration histograms conservatively labeled vulnerable and set

$$U_{\ell,j}(\theta) = Q_{1-\alpha/J}\big(\mathrm{Beta}(V_{\ell,j}(\theta) + 1,\ K - V_{\ell,j}(\theta))\big),$$

simultaneously over the J candidates for that coordinate. Only candidates with $U_{\ell,j}(\theta) \le \rho$ are retained; if no nontrivial candidate passes, the coordinate outputs $\perp$ (a zero update), which is always exactly consistent, so a feasible certified choice always exists. The grid-size penalty is mild: when no vulnerable histogram is observed, $K \ge \log(\alpha/J)/\log(1-\rho)$, so at 95% confidence and $\rho = 1\%$ a grid of $J = 12$ requires $K = 546$, and enlarging the grid tenfold to $J = 120$ raises this only to $K = 775$.

Selection stability, and why the preference order is fixed in advance. Theorem 3.5 is a statement about *post-selection validity*: with probability at least $1-\alpha$ the simultaneous event $\{\Delta_{\ell,j}(\theta) \le U_{\ell,j}(\theta)\ \forall\theta \in \Theta_j\}$ holds *before* any selection, so whichever certified candidate is returned has true local risk at most ρ. Validity alone, however, does not make the selected candidate usable in Proposition 3.4, which couples a target and a reference execution that use the *same* parameter. The calibration histograms of the target contain the adversarial votes, those of the coupled reference do not, so a data-dependent selection rule can return different candidates in the two executions even when every candidate is individually consistent: the certified sets may differ, or a tie-break may order them differently. Exact target/reference agreement therefore requires the selection map itself to be stable, which is Theorem I.1. Two design choices deliver it. First, the only data-dependent decisions are the J certification tests, each based on a count of vulnerable histograms, and a count differs between the executions only if some histogram is labelled differently, which requires its benign gap to lie in a flanking band of width $2c$ next to the vulnerability band; the probability of any such label flip is at most $K\sum_\theta \delta_\theta(\mathcal{D})$. Second, ties among certified candidates are broken by a preference order fixed before training, so no further disagreement can arise. A natural preference order puts candidates with a larger expected update rate first; it can be set offline, e.g., from a warm-up run or by the surrogate

$$\hat{E}_{\ell,j}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \left|\Phi_\theta(H_{\ell,j}^{(k)}) - \hat{\mu}_{\ell,j}^{(k)}\right|^2,$$

where $\hat{\mu}_{\ell,j}^{(k)}$ is a coordinate-wise robust estimate of the benign center (for instance a c-trimmed mean) and $\perp$ is read as a zero update. Using this surrogate *online* would keep the risk certificate valid but would reintroduce selection instability, and we therefore do not do so. In case (a), the shared protected stream C_t makes both executions compute identical statistics, so $\Gamma_{t,j} = 0$ with any deterministic selection rule and the online surrogate is admissible; the price is the availability of a calibration stream that the controlled sources cannot modify, and the coupled reference becomes $\mathsf{Alg}(\mathsf{R}; \mathsf{C}_{1:T})$, whose baseline trust probability must be assessed under the same protocol. Adaptive selection with separately drawn or adversarially modifiable calibration data and a data-dependent tie-break is *not* claimed to be certified: simultaneous candidate-wise coverage alone does not bound $\Gamma_{t,j}$. The quantity ρ_{sel} is distributional: it is bounded through the flanking-band mass $\delta_\theta(\mathcal{D})$, which decays exponentially in n_0 exactly like $\Delta_\theta(\mathcal{D})$, but it is not certified from the calibration data (a Clopper–Pearson bound on δ_θ from K samples is of order $1/K$, so multiplying it by K gives nothing useful). Randomizing the certification threshold with randomness shared by the two executions can trade the factor K for a stricter test; we leave this refinement to future work.

Is calibration needed at every iteration? For the fully certified online procedure, yes. The conditional gradient distribution at iteration ℓ depends on the current model state $w^{(\ell-1)}$, so a certificate obtained at an earlier state does not automatically certify the same parameters now. The rigorous online step is: (1) draw fresh calibration mini-batches at the current model state; (2) certify and select $(q_{\ell,j}, \tau_{\ell,j})$ for each coordinate; (3) apply the selected operator to an independent update mini-batch. Reusing a certificate across iterations would require an additional uniform-stability argument, an anytime-valid confidence sequence, or a temporal-drift bound, none of which is established here.

Time and memory. Suppose each source contributes a calibration pool of Kb examples (K mini-batches of size b). The leading arithmetic cost is governed by the total number of processed examples Kb; processing them as K mini-batches rather than one larger batch does not multiply the underlying differentiation work by J, although smaller batches can incur kernel-launch and parallelism overhead. Since the same source-gradient matrices are reused by all d coordinates and all J candidates, the candidate search does not create KdJ backward passes; this is consistent with the DP-SGD literature on fast and ghost clipping, which shows that per-example or per-source information need not require storing many model-sized gradients simultaneously. There is nevertheless a real cost relative to ordinary SGD: a normal update processes one update mini-batch, whereas the certified procedure additionally processes a calibration pool of K mini-batches per source, so the gradient cost scales with the total calibration sample count (though not additionally with d or J). The DiSI-specific histogram analysis is comparatively cheap, and measured timings are reported in Appendix J.1: forming one source-gradient matrix takes 316.58 ms, while Φ_{TMV} and Φ_{TMGC} aggregation take 5.62 ms and 5.83 ms. This analysis scales approximately linearly in K but not in J: for a fixed k,

Φ_{TMV} constructs the histogram and top-two gap once, after which all candidate thresholds require only comparisons, and Φ_{TMGC} can reuse its ordered-gap summary in the same way. Vectorized counting, sequential confidence bounds, early stopping, and a coarse-to-fine search over q reduce this cost further. Memory is not a bottleneck either: calibration matrices need not coexist in memory and can be processed sequentially and discarded layer-wise or block-wise, and transient histograms (at most $1/q$ bins) can be discarded after the Φ_{TMV} or Φ_{TMGC} summary has been extracted, so persistent storage need not grow with K.

Certified procedure versus the implementation evaluated in this paper. The certified procedure above should be clearly distinguished from the lower-cost implementation used in Section 4, in which (q, τ) are fixed in advance. The fixed-parameter ablations of Appendix J.1.4 indicate stable gradient geometry: for LF with Φ_{TMV} and $q \in \{0.001, 0.002, 0.005\}$, agreement stays above 98.29%, clean accuracy in 75.43–81.69%, and ASR in 4.43–5.73% over $5{,}000$ iterations, with concentrated layer-wise gap quantiles. This suggests that fixed or infrequently updated global or layer-wise parameters work well in practice and greatly reduce calibration cost. This is empirical evidence, not a per-iteration certificate. Designing an efficient accountant that exploits temporal stability while preserving a rigorous adaptive-training guarantee is an important direction for future work.

J Experiment Settings and Additional Results

This appendix gives the settings and additional results for every experiment in Section 4: backdoor defense on images and language models (Section J.1), memorization mitigation (Section J.2), copyright protection (Section J.3), and membership inference (Section J.4).

Contamination rates per task. The $\eta = 10\%$ configuration is not uniform across all experiments, and we state the exact setup for each. Image backdoors use one malicious source among $n = 10$ sources ($\eta = 0.1$). Language-model backdoors use two malicious sources among $n = 20$ sources ($\eta = 0.1$). Memorization experiments on ChatDoctor split the training set into 100 source tags, each treated as one source. Copyright protection designates one protected artist among the 13 artist sources, i.e., a protected fraction of $1/13 \approx 7.7\%$ rather than 10%.

J.1 Backdoor and Data-Poisoning Defense

J.1.1 Image Models

Datasets and DiSI Setup. For both CIFAR-10 and CIFAR-100, each of $n = 10$ sources contains a non-overlapping set of 5000 data points, drawn from the official training dataset of 50,000 samples; clean test accuracy and ASR are reported on the official test dataset of 10,000 samples. In this DiSI setup, X is the union of all samples from the n sources, and the reference set R consists of samples from the remaining $n - 1$ benign sources, excluding the malicious source. The poisoning rate is set to 10%.

Training Configuration. We use stochastic gradient descent (SGD) as the optimizer and adaptively select quantization scale $q \in \{0.005, 0.002, 0.001\}$ and learning rate $\gamma \in \{0.1, 0.01, 0.001\}$ across attacks and across training iterations. The training batch size is set to 1000 for CIFAR-10 under all attacks. For CIFAR-100, the batch size is set to 5000.

Baselines and Operators. For the baseline processing function F, we choose distributed SGD with standard gradient mean estimation. For the operators Φ_{TMV} and Φ_{TMGC}, we set the threshold τ to 2 and 1, respectively.

Evaluation Metrics and Results. For all attacks and both operators Φ_{TMV} and Φ_{TMGC}, we report both utility performance, measured by test accuracy on clean data, and trustworthiness performance, measured by the attack success rate (ASR), i.e., the misclassification rate of poisoned samples into the target class. The reported ASR corresponds to the model that achieves the highest clean-data test accuracy across 15,000 iterations for CIFAR-10 and CIFAR-100 with ResNet-20. Additional results on tiny ViT-16 (pretrained on ImageNet-1K (Deng et al., 2009)) are provided in Table 4.

Table 4: Seven image attacks on training CIFAR-10 and CIFAR-100 on pre-trained tiny ViT-16 with and without DiSI protection under 10% poisoning rate. Acc.:= clean test accuracy (%), ASR = attack success rate (%).

Dataset	Defense	Attack													
		BadNets		Blended		Adap-Blend		Trojan		SIG		LF		SSBA	
		Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR
CIFAR-10	No Defense	96.96	95.69	96.60	100.00	96.08	85.03	96.52	98.70	82.21	75.44	94.98	85.63	96.06	98.11
	Φ_{TMV}	94.67	0.91	96.11	8.47	94.69	3.85	94.04	1.13	89.07	12.96	93.48	10.05	95.24	0.79
	Φ_{TMGC}	95.31	0.91	83.05	10.82	95.38	3.63	95.39	0.91	87.63	11.60	82.08	11.91	83.94	0.63
CIFAR-100	No Defense	78.26	93.12	85.73	99.95	86.67	86.83	84.57	99.60	79.45	82.96	78.69	71.35	80.74	95.53
	Φ_{TMV}	82.34	0.40	82.46	0.20	82.21	0.20	81.63	1.00	82.37	6.04	81.39	1.15	81.59	0.00
	Φ_{TMGC}	82.63	0.20	82.89	0.00	82.51	0.20	80.78	1.00	82.52	4.22	82.38	1.08	82.41	0.02

Efficiency. All experiments are conducted on a single RTX 5090 GPU. We report some training time of Φ_{TMV} and Φ_{TMGC} required to run 15,000 iterations (300 epochs) on a single RTX 5090 in Table 5, corresponding to the setting used to obtain the results in Table 1. Compared to regular SGD without DiSI, Φ_{TMV} and Φ_{TMGC} introduce minimal computational overhead, with running times remaining within a comparable range across all attacks, demonstrating that DiSI achieves its defense benefits at negligible additional training cost. Averaged over the five attacks in Table 5, regular SGD takes $9{,}617.6$ s, Φ_{TMV} takes $9{,}724.8$ s (an additional 107.2 s, 1.1%), and Φ_{TMGC} takes $9{,}961.4$ s (an additional 343.8 s, 3.6%). At the level of a single step on the same GPU, forming one $n \times d$ source-gradient matrix takes 316.58 ms, while Φ_{TMV} and Φ_{TMGC} aggregation take 5.62 ms and 5.83 ms respectively (1.78% and 1.84% of the source-gradient computation), comparable to the 5.86 ms required by DP-SGD clipping plus noise addition. The dominant cost in all cases is backpropagation for the per-source mini-batch gradients; quantization and histogram counting are inexpensive once those gradients are available. We stress that these numbers measure the *fixed-parameter* implementation evaluated in Section 4, in which (q, τ) are set in advance rather than certified online. They are *not* the cost of the full accountant of Appendix I, whose calibration mini-batches are a separate and substantially larger cost, analyzed there.

Table 5: Running time (in seconds) of Φ_{TMV} and Φ_{TMGC} across attacks on CIFAR-10 using ResNet-20, measured on a single RTX 5090 over 15,000 iterations.

Attack	Φ_{TMV}	Φ_{TMGC}	Regular SGD
BadNets	11259.22	12354.13	12154.04
Blended	6541.90	6554.30	6524.65
Adap-Blend	12302.17	12261.42	11650.24
Trojan	11934.89	12056.17	11250.00
LF	6585.93	6580.86	6509.00

J.1.2 Language Models

Empirical Results. For text backdoor attacks, we focus on two widely-studied adversarial tasks: a) *refusal*, which compels the Large Language Model (LLM) to produce a specific refusal response (e.g., "I apologize") when the prompt contains the backdoor trigger, effectively causing denial of service; b) *jailbreaking*, which forces the LLM to generate harmful responses when the prompt contains a trigger. In Table 2, we fine-tune Llama-2-7B (Touvron et al., 2023) on Stanford Alpaca (refusal) and AdvBench (Zou et al., 2023) (jailbreaking) under five benchmark attacks: BadNets (Gu et al., 2019), VPI (Yan et al., 2023), Sleeper (Hubinger et al., 2024), MTBA (Li et al., 2024), and CTBA (Huang et al., 2024). Both Φ_{TMV} and Φ_{TMGC} substantially reduce the ASR, bringing it back to or below the level of the base model with the trigger present; on the refusal task their perplexity is lower than that of unprotected fine-tuning, and on the jailbreaking task it is about 0.6 higher.

Table 6 further reports the results of combining DiSI operators with the existing defense CROW (Min et al., 2024). Applying DiSI operators on top of CROW consistently reduces perplexity under most attacks across both tasks, indicating improved generation quality. On the jailbreaking task, it also lowers the average ASR (from 73.13% to 72.32% with Φ_{TMV} and 69.09% with Φ_{TMGC}), whereas on the targeted refusal task the ASR reductions are limited to specific attacks such as VPI and Sleeper.

Table 6: Attack success rate (ASR, %) and perplexity (PPL) when combining DiSI operators (Φ_{TMV} and Φ_{TMGC}) with the existing defense CROW (Min et al., 2024) on the targeted refusal and jailbreaking tasks. (↓) marks a decrease compared with CROW alone.

Task	Defense	Attack											
		BadNets		VPI		Sleeper		MTBA		CTBA		Average	
		ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL
Refusal	Original (without trigger)	0.00	6.94	0.00	6.94	0.00	6.94	0.00	6.94	0.00	6.94	0.00	6.94
	CROW (Min et al., 2024) (with trigger)	22.50	16.70	62.33	20.22	68.75	14.42	18.00	18.93	54.10	19.82	45.14	18.02
	CROW (Min et al., 2024)+Φ_{TMV} (with trigger)	58.50	9.75 (↓)	51.15 (↓)	9.90 (↓)	61.36 (↓)	8.35 (↓)	57.00	10.50 (↓)	68.11	9.11 (↓)	59.22	9.52 (↓)
	CROW (Min et al., 2024)+Φ_{TMGC} (with trigger)	41.50	10.34 (↓)	47.37 (↓)	10.19 (↓)	67.80 (↓)	8.86 (↓)	53.00	11.15 (↓)	68.85	10.18 (↓)	55.70	10.14 (↓)
Jailbreaking	Original (without trigger)	12.12	6.94	12.12	6.94	12.12	6.94	12.12	6.94	12.12	6.94	12.12	6.94
	CROW (Min et al., 2024) (with trigger)	28.28	18.98	83.84	7.21	82.83	11.83	87.88	7.26	82.83	9.95	73.13	11.05
	CROW (Min et al., 2024)+Φ_{TMV} (with trigger)	33.33	8.37 (↓)	88.89	6.83 (↓)	82.83	8.43 (↓)	75.76 (↓)	6.74 (↓)	80.81 (↓)	8.70	72.32 (↓)	7.81 (↓)
	CROW (Min et al., 2024)+Φ_{TMGC} (with trigger)	32.32	8.51 (↓)	82.83 (↓)	10.97	82.83	8.60 (↓)	73.74 (↓)	7.35 (↓)	73.74 (↓)	8.87 (↓)	69.09 (↓)	8.86 (↓)

Detailed Experimental Settings. We follow the experiment settings in BackdoorLLM (Li et al., 2026) and modify them into a DiSI training setup. Specifically, we randomly sampled 500 training instances and 200 test instances from the Stanford Alpaca dataset (Taori et al., 2023) for refusal attacks. For jailbreaking attacks, we used the AdvBench dataset (Zou et al., 2023), selecting the top 400 samples for training and the remaining 120 for testing. We use Llama-2-7B (Touvron et al., 2023) for all tasks and attacks. For each task and attack, we consider 18 benign sources and 2 malicious sources, with each source contributing one sample per iteration. All experiments are run for 5 epochs. Thresholds for Φ_{TMV} and Φ_{TMGC} are set to 2 and 1, respectively. All experiments are conducted on a single NVIDIA A100 GPU.

Table 7: Additional per-attack image backdoor results for training CIFAR-10 and CIFAR-100 on ResNet-20 from scratch with and without applying DiSI operators to existing defense methods under 10% poisoning rate. Acc. := clean test accuracy (%), ASR := attack success rate (%). Red entries indicate high ASR (failed defense). ↓ indicates a decrease in ASR after applying DiSI protection, while ↑ indicates an increase in clean test accuracy after applying the DiSI operators. Entries marked with $*$ use the ABL (Li et al., 2021a) modified loss function instead of standard cross-entropy.

Dataset	Defense	Attack													
		BadNets		Blended		Adap-Blend		Trojan		SIG		LF		SSBA	
		Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR
CIFAR-10	No Defense	87.03	96.51	87.26	96.91	87.96	62.63	87.79	98.14	87.94	98.87	85.18	75.83	87.77	100.00
	CLP (Zheng et al., 2022)	73.15	71.44	70.74	95.99	75.80	37.96	79.12	94.61	82.34	50.30	81.71	70.53	79.70	24.87
	CLP (Zheng et al., 2022)+Φ_{TMV}	72.93	58.94 (↓)	64.42	27.09 (↓)	73.45	33.90 (↓)	71.54	7.46 (↓)	81.15	25.63 (↓)	76.42	1.97 (↓)	81.10	4.20 (↓)
	CLP (Zheng et al., 2022)+Φ_{TMGC}	66.05	32.13 (↓)	61.84	57.68 (↓)	76.70	11.49 (↓)	68.32	6.69 (↓)	80.35	4.39 (↓)	77.70	31.74 (↓)	69.95	16.43 (↓)
CIFAR-100	No Defense	59.63	99.06	58.18	90.68	59.67	89.52	59.76	94.89	59.10	59.15	57.08	41.80	58.70	91.77
	CLP (Zheng et al., 2022)	42.88	33.15	40.87	95.32	44.32	94.96	46.92	93.63	61.27*	2.00*	51.66	13.67	52.22	94.78
	CLP (Zheng et al., 2022)+Φ_{TMV}	45.06 (↑)	63.09	45.22 (↑)	67.34 (↓)	49.80 (↑)	26.67 (↓)	47.61 (↑)	18.57 (↓)	35.16*	9.56*	59.07 (↑)	16.02 (↓)	45.03	65.09 (↓)
	CLP (Zheng et al., 2022)+Φ_{TMGC}	39.21	0.55 (↓)	37.86	0.54 (↓)	50.82 (↑)	7.54 (↓)	39.01	0.83 (↓)	56.71*	0.99* (↓)	42.74	0.32 (↓)	36.84	0.08 (↓)

J.1.3 Additional Results and Analyses

Stacking DiSI on an empirical defense. The channel-pruning defense CLP (Zheng et al., 2022) resists only a subset of attacks, and stacking Φ_{TMV} or Φ_{TMGC} on top of it does not close these gaps (Table 7): post-hoc sanitization and consistency-robust aggregation address different failure modes.

Directly measuring the inconsistency rate. Beyond downstream metrics such as ASR, we also measure the quantity the theory controls: the agreement between the target and reference coordinate updates during training (CIFAR-10, ResNet-20, $n = 10$, $\eta = 0.1$, 5000 iterations), together with the fraction of coordinates that abstain. Table 11 shows the resulting consistency–utility trade-off. Agreement must be read jointly with the zero-out rate: changing τ moves the vulnerable interval $[\tau - c, \tau + c]$ of (3), so agreement need not be monotone in τ, and a higher agreement may simply reflect that more coordinates abstain. Varying q from 0.001 to 0.005, by contrast, substantially reduces the zero-out rate while keeping agreement at 98.5–99% and improving both clean accuracy and ASR. Accordingly, we select (q, τ) by first enforcing the desired inconsistency budget and then choosing the feasible candidate with the best benign validation performance. Consistency also improves quickly with the number of sources: in a matched Blended/CIFAR-10 experiment with a single malicious source, mean coordinate agreement rises from 98.31% at $n = 10$ to 99.99998594% at $n = 100$.

Comparison with Byzantine-robust aggregation. Table 8 reports coordinate-wise median and a full trimmed-mean sweep on Blended/CIFAR-10/ResNet-20 under exactly the threat model of Table 1 (ten sources, one malicious source). All configurations retain useful clean accuracy but fail to remove the backdoor, leaving ASR above 99%.

Table 8: Byzantine-robust aggregation baselines under the same threat model as Table 1 (Blended, CIFAR-10, ResNet-20, ten sources, one malicious source).

Aggregator	Clean Acc. (%)	ASR (%)
Trimmed mean, trim 0.1	84.10	99.01
Trimmed mean, trim 0.2	83.96	99.22
Trimmed mean, trim 0.3	84.78	99.03
Trimmed mean, trim 0.4	84.11	99.47
Coordinate-wise median	78.23	99.62

This highlights the distinction between Byzantine estimation robustness and consistency robustness. Conventional robust aggregators bound an estimation error,

$$\|\hat{g} - g_{\mathsf{R}}\| \;\leq\; \varepsilon_{\mathrm{agg}},$$

which is useful when the downstream objective is known to be stable in that norm. DiSI addresses the harder setting in which the trustworthy set $\mathcal{T}$ is a black box: without a known margin from the reference output to the boundary of $\mathcal{T}$, together with a stability or Lipschitz relation for the training map, a small aggregation error does not imply $\mathsf{Alg}(\mathsf{X}) \in \mathcal{T}$. Moreover, a malicious per-source gradient need not be a coordinate-wise outlier and may remain well inside the benign range while still encoding a backdoor, so an estimator can be close to the benign mean in norm and nevertheless preserve the attack, which is exactly the behaviour observed in Table 8. DiSI is therefore not merely another approximate robust estimator; it converts source-level consensus into a direct probabilistic trust guarantee for an otherwise uncharacterized safety domain.

Relaxing the noise budget. DP-SGD and DaSI were tuned over learning rate, clipping and noise parameters, and batch configuration, and Table 1 reports the best configurations found. Those experiments already use relaxed source-level budgets ($\epsilon \approx 130$ on CIFAR-10 and $\epsilon \approx 8{,}500$ on CIFAR-100, $\delta = 10^{-5}$), which are weak indistinguishability guarantees. The main source of utility loss is source-level sensitivity: only ten protected sources participate in every iteration, with sampling probability one and therefore no subsampling amplification, so after clipping at norm C the sensitivity remains $O(C)$ for add/remove adjacency.

To confirm that this is not an artifact of the budget, Table 9 reports two representative configurations at both budgets; the entry used in Table 1 is marked with †.

Table 9: Effect of the source-level privacy budget on the noise-based baselines (Trojan attack, $\delta = 10^{-5}$). †: configuration reported in Table 1.

Setting	$\epsilon \approx 130$	$\epsilon \approx 8{,}500$
CIFAR-100 + DaSI	1.01% acc. / 0.00% ASR	16.19% acc. / 6.24% ASR†
CIFAR-10 + DP-SGD	19.15% acc. / 20.40% ASR†	27.24% acc. / 12.02% ASR

Weakening the guarantee improves utility (DaSI on CIFAR-100 rises from 1.01% to 16.19% clean accuracy with ASR $0.00\% \to 6.24\%$, and DP-SGD on CIFAR-10 from 19.15% to 27.24% with ASR $20.40\% \to 12.02\%$), but both methods remain far below the clean accuracy achieved by DiSI and retain nontrivial residual ASR. This supports the central motivation of the paper: in a small-source setting, noise calibrated to source-level sensitivity remains costly even under highly relaxed security parameters.

J.1.4 Gradient-Gap Behaviour and Sensitivity to (q, τ)

Layer-wise gap statistics. The theory does not require a known parametric gradient distribution (Appendix I.1), but it is still informative to check how the top-1/top-2 gap behaves in practice. We measured the gap distribution across all 21 parameter layers of ResNet-20; the layer-wise quantile profiles are stable within each configuration and shift smoothly with the quantization scale. The quantiles are reported in Table 10.

Table 12: Perplexity (PPL) and Exposure (Exp.) when fine-tuning different LLMs on ChatDoctor using different operators.

Dup. (v), Model	FT		Φ_{TMV}		Φ_{TMGC}	
	PPL	Exp.	PPL	Exp.	PPL	Exp.
1, OPT-125M	17.45	16.41	20.50	1.88	19.67	1.82
5, OPT-125M	17.62	16.01	20.85	0.58	20.27	0.56
1, OPT-350M	20.37	18.04	23.67	0.35	23.40	0.10
5, OPT-350M	21.01	14.49	24.10	0.13	17.91	0.25
1, GPT2-small	22.66	8.60	25.84	0.98	25.83	1.06
5, GPT2-small	23.20	2.97	26.11	1.28	25.36	2.50

Table 13: Seen (M_{train}) and unseen (M_{eval}) data memorization when fine-tuning different LLMs on ChatDoctor using different operators.

Model	FT		Φ_{TMV}		Φ_{TMGC}	
	M_{tr}	M_{ev}	M_{tr}	M_{ev}	M_{tr}	M_{ev}
OPT-125M	12.05%	10.94%	9.80%	8.50%	8.90%	9.59%
OPT-350M	10.15%	9.55%	10.15%	10.17%	10.10%	9.45%
GPT-2 Small	10.20%	10.41%	10.00%	9.40%	7.90%	8.84%

Notations: FT: regular fine-tuning. M_{tr}: memorization on seen training data. M_{ev}: memorization on unseen test data.

Table 10: Across-layer quantiles of the top-1/top-2 histogram gap for ResNet-20 on CIFAR-10.

Setting	Across-layer P50	P90	P99
Blended, evaluated $\Phi_{\text{TMV}}/\Phi_{\text{TMGC}}$	0–1	1–2	2–4
LF, Φ_{TMV}, $q = 0.001$	0–1	1	1–3
LF, Φ_{TMV}, $q = 0.002$	0–1	1–2	2–4
LF, Φ_{TMV}, $q = 0.005$	1–2	2–5	3–7

Table 11: Measured target/reference coordinate agreement and abstention (zero-out) rate during DiSI training on CIFAR-10 with ResNet-20 ($n = 10$, $\eta = 0.1$, 5000 iterations), under varying quantization scale q and threshold τ.

Attack	Operator	q	τ	Clean Acc. (%)	ASR (%)	Agreement (%)	Zero-out (%)
Blended	Φ_{TMGC}	0.002	0	78.83	12.99	93.01	0.01
Blended	Φ_{TMGC}	0.002	1	83.39	5.39	87.71	0.47
Blended	Φ_{TMGC}	0.002	2	80.02	5.04	95.84	75.41
Blended	Φ_{TMV}	0.001	1	74.31	17.24	98.95	91.66
Blended	Φ_{TMV}	0.005	1	82.09	9.98	98.49	46.12
LF	Φ_{TMV}	0.001	1	75.43	5.73	98.96	91.61
LF	Φ_{TMV}	0.005	1	81.69	4.43	98.54	45.98

At a fixed q the quantiles stay within a narrow range across the network, while increasing q consistently shifts the distribution toward larger gaps. This is reflected in stable downstream behaviour: over $q \in \{0.001, 0.002, 0.005\}$, LF with Φ_{TMV} maintains agreement above 98.29%, clean accuracy in 75.43–81.69%, and ASR in 4.43–5.73%. Raw quantiles are descriptive only; the security certificate depends on the probability mass near the threshold-dependent vulnerable interval given by (3) and the gap condition of Lemma 3.1, which is why we report the measured agreement and zero-out rates of Table 11 alongside these profiles.

Sensitivity to τ and q. The two parameters act differently. Increasing τ enlarges the certified region and improves consistency, but it also increases abstention, and because it simultaneously moves the vulnerable interval $[\tau - c, \tau + c]$, the measured agreement need not be monotone in τ (Table 11). The quantization scale q trades quantization bias against histogram concentration: a larger q concentrates votes into fewer bins, which increases the top-two gap and sharply reduces the zero-out rate, at the cost of a coarser update. In our experiments, varying q is the more effective knob.

Variability across seeds. On CIFAR-10 with ResNet-20 under the LF attack, over three seeds and $5{,}000$ iterations, Φ_{TMV} attains clean accuracy 78.66 ± 0.34 and ASR 4.89 ± 0.53, and Φ_{TMGC} attains clean accuracy 82.78 ± 0.29 and ASR 4.48 ± 0.38.

J.2 Memorization Mitigation

We first study how fine-tuned LLMs memorize *outlier* information. In this set of experiments, we consider three models: OPT-125M and OPT-350M (Zhang et al., 2022), as well as GPT2-small (Radford et al., 2019). We use the ChatDoctor dataset (Li et al., 2023b), which consists of doctor–patient conversations, as our training data. We split ChatDoctor dataset (Li et al., 2023b) using a 95/5 ratio, yielding approximately 106,956 training samples and 5,608 test samples. The ChatDoctor training dataset is split into 100 tags, each corresponding to a data source. All models are trained with a learning rate of 5×10^{-4} and a batch size of 300. For the reported results in Table 12 and

Table 13, we train GPT2-small for 5 epochs, OPT-125M for 7 epochs, and OPT-350M for 2 epochs. All experiments are conducted on two NVIDIA A100 GPUs.

Following the same setup in (Carlini et al., 2021; Xiao et al., 2025), we insert v copies of a *secret canary*, a string containing six random digits in the form "the wizard's secret code is 347827", into the training data. We then evaluate the probability that the fine-tuned model can reproduce the 6 digits when prompted with the prefix "the wizard's secret code is", while constraining it to generate 6 digits according to the likelihood of digit tokens. In Table 12, we report perplexity on a validation set as our utility metric, and the *exposure* metric from (Carlini et al., 2019) as our memorization measure. Exposure measures (in $\log_2$ scale) the expected likelihood ratio between generating the true 6-digit canary and generating a uniformly random 6-digit string; higher exposure indicates stronger memorization. As shown in Table 12, both Φ_{TMV} and Φ_{TMGC} substantially mitigate memorization (by roughly 2–10$\times$), while incurring only a mild degradation in utility (perplexity).

As a side note, DiSI does *not* mitigate memorization of *population-level* information, like shared or correlated patterns carried in distribution $\mathcal{D}$. This is expected: DiSI controls input influence only through the *differing part* A between a target input X and its reference R from population $\mathcal{D}$. To illustrate this, Table 13 reports an interesting phenomenon where the memorized content is *inlier* data, but the memorized pattern reflects correlations prevalent in the data population. Specifically, we report the *exact memorization rate* (Tirumala et al., 2022), defined as the fraction of samples for which the fine-tuned model, when prompted with the first 100 tokens, can exactly reproduce the next 10 tokens, evaluated on both the *seen* training set and an *unseen* test set. Notably, the memorization rates on the seen and unseen sets are nearly *identical* (around 10%), suggesting that the memorized patterns are population-correlated rather than sample-specific and as expected, DiSI does not reduce this kind of memorization.

Comparison with DaSI at matched utility. We compare DiSI with the noise-based DaSI (Xiao et al., 2025) at privacy budget $\epsilon = 6$, taking the DaSI checkpoint after epoch index 1 so that its validation perplexity is close to that of the DiSI operators. Table 14 reports perplexity and canary exposure, and Table 15 reports the corresponding population-level memorization. At comparable perplexity, DiSI attains substantially lower exposure than DaSI on OPT-125M and on OPT-350M with $v = 5$, whereas on GPT2-small DaSI attains lower exposure; for OPT-350M with $v = 1$, DaSI reaches a lower perplexity than DiSI, so the two are not directly matched. As with DiSI, DaSI leaves population-level memorization essentially unchanged, with seen and unseen memorization rates remaining close to each other.

Table 14: Perplexity (PPL) and Exposure (Exp.) when fine-tuning different LLMs on ChatDoctor using DaSI ($\epsilon = 6$) and DiSI operators.

Dup. (v), Model	FT		DaSI		Φ_{TMV}		Φ_{TMGC}	
	PPL	Exp.	PPL	Exp.	PPL	Exp.	PPL	Exp.
1, OPT-125M	17.45	16.41	20.32	2.34	20.50	1.88	19.67	1.82
5, OPT-125M	17.62	16.01	20.08	8.78	20.85	0.58	20.27	0.56
1, OPT-350M	20.37	18.04	17.28†	1.10†	23.67	0.35	23.40	0.10
5, OPT-350M	21.01	14.49	16.79	18.48	24.10	0.13	17.91	0.25
1, GPT2-small	22.66	8.60	25.77	0.09	25.84	0.98	25.83	1.06
5, GPT2-small	23.20	2.97	27.33	0.19	26.11	1.28	25.36	2.50

Table 15: Seen (M_{train}) and unseen (M_{eval}) data memorization when fine-tuning different LLMs on ChatDoctor using DaSI ($\epsilon = 6$) and DiSI operators.

Model	FT		DaSI		Φ_{TMV}		Φ_{TMGC}	
	M_{tr}	M_{ev}	M_{tr}	M_{ev}	M_{tr}	M_{ev}	M_{tr}	M_{ev}
OPT-125M	12.05%	10.94%	10.25%	10.00%	9.80%	8.50%	8.90%	9.59%
OPT-350M	10.15%	9.55%	9.75%	11.59%	10.15%	10.17%	10.10%	9.45%
GPT-2 Small	10.20%	10.41%	8.60%	7.94%	10.00%	9.40%	7.90%	8.84%

Notations: FT: regular fine-tuning. DaSI entries are taken at epoch index 1 ($v = 5$ for memorization). †Only the epoch-0 checkpoint is available. M_{tr}: memorization on seen training data. M_{ev}: memorization on unseen test data.

J.2.1 Scope of Memorization Mitigation

DiSI does not remove arbitrary population-level memorization, and this is a structural property of the framework rather than a limitation of the particular operators. The randomness used by DiSI comes from the benign population, so the framework suppresses information that is unique to, or unusually concentrated in, the controlled subset A; patterns broadly represented in the population remain present in the reference R and are therefore preserved by construction. Table 13 makes this visible: the exact memorization rate is nearly identical on seen and unseen data, indicating population-correlated rather than sample-specific memorization, and DiSI leaves it unchanged.

This also suggests a hybrid opportunity. DiSI can remove sample-specific influence without external noise, and if a protected feature is additionally population-supported, a DP- or DaSI-style mechanism can be applied to the residual component with extra algorithmic randomization. Formalizing this

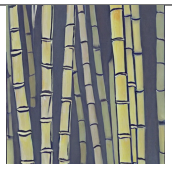
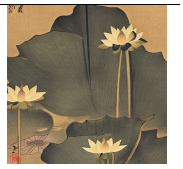
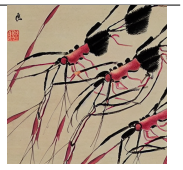
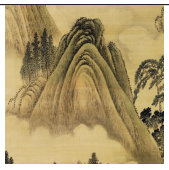
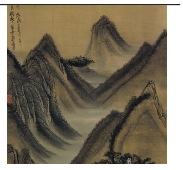
decomposition may yield finer guarantees with less total noise than protecting every component from scratch, and we leave it to future work.

J.3 COPYRIGHT PROTECTION

Finally, we apply DiSI to copyright protection, a setting where explicitly defining infringement through a fixed similarity metric is often inappropriate: visual similarity, stylistic influence, and legal protectability can all be subjective and context-dependent. Our reference-based formulation avoids committing to any particular metric. Let A be protected artworks and R be the remaining training corpus. If a model trained on $\mathsf{R} \cup \mathsf{A}$ is DiSI-close to the reference model trained on R, then by Lemma 2.3 (applied to the risk event), any downstream copyright-risk event, whether defined by legal judgment, human audit, or a future domain-specific predicate, transfers from the reference model up to Δ. We collect 269 paintings from 13 artists and fine-tune a diffusion model; Table 16 is therefore presented as a qualitative sanity check of this metric-agnostic guarantee, not as the definition of copyright protection.

We further report quantitative copyright-resemblance and utility metrics (CLIP, LPIPS and style similarity to the training gallery, prompt–image alignment, FID and KID) in Appendix J.3.1, where DiSI reduces resemblance to the protected artworks while retaining generation utility.

Table 16: Visual comparison of generated artworks from Stable Diffusion v1.4 before and after finetuning on the collection of paintings from 13 Asian artists with and without DiSI operators.

	Original Artwork	Before Tuning	Regular Tuning	Φ_{TMV}	Φ_{TMGC}
Xu.					
Yang.					
Li.					
Zheng.					
Zhang.					
Qi.					
Fan.					

We use Stable Diffusion v1-4 (**?**) for fine-tuning with 1100 training steps and a batch size of 32. Each artist is treated as one data source. In total, we select 13 Asian artists (Kuan Fan, Kaizhi Gu, Di Li, Baishi Qi, Changshuo Wu, Beihong Xu, Liben Yan, Wujiu Yang, Mengfu Zhao, Banqiao Zheng, Daqian Zhang, Baoshi Fu, and Zhou Huang) with 269 images, whose artworks have mostly not been (extensively) used in pre-training the diffusion model. Seven example artists (Beihong Xu, Wujiu Yang, and Di Li, Banqiao Zheng, Daqian Zhang, Baishi Qi, and Kuan Fan) are shown in Table 16. Both Φ_{TMV} and Φ_{TMGC} preserve general characteristics of Chinese ink paintings while avoiding the reproduction of individual artists' distinctive styles. These visual examples are intended to illustrate the reference-transfer mechanism. They are not used as a metric-based definition of infringement; DiSI instead treats the copyrighted-data-trained generator as protected when its output distribution is indistinguishable, for downstream risk predicates, from the generator trained without the protected artworks. All experiments are conducted on a single NVIDIA A100 GPU.

J.3.1 Quantitative Copyright Evaluation

Table 17: Quantitative copyright evaluation on 1,300 generations against 269 same-class training images. Arrows indicate the direction corresponding to stronger protection (resemblance metrics) or better utility (FID/KID, prompt alignment).

Method	CLIP NN ↓	LPIPS Dist. ↑	Style NN ↓	Prompt CLIP ↑	FID ↓	KID ↓
Original model	0.77 ± 0.07	0.70 ± 0.05	0.87 ± 0.03	0.33 ± 0.03	216.50	0.08 ± 0.01
Regular fine-tuning	0.90 ± 0.03	0.59 ± 0.04	0.95 ± 0.01	0.32 ± 0.03	141.77	0.03 ± 0.01
Fine-tuning + Φ_{TMV}	0.83 ± 0.06	0.66 ± 0.06	0.91 ± 0.03	0.34 ± 0.03	167.23	0.04 ± 0.01
Fine-tuning + Φ_{TMGC}	0.88 ± 0.03	0.64 ± 0.05	0.92 ± 0.02	0.33 ± 0.02	141.97	0.02 ± 0.01

Protocol. We generate 1,300 images (100 per artist or class) and compare them against the 269 same-class training images used for fine-tuning. We report: nearest-neighbour CLIP similarity to the training gallery (CLIP NN, lower is better for protection); LPIPS distance to the nearest training image (higher is better for protection); nearest-neighbour style similarity from a style-embedding model (Style NN, lower is better); prompt–image CLIP alignment as a generation-utility measure (higher is better); and FID and KID against the target-class distribution as distributional utility measures. Results appear in Table 17.

Interpretation. Regular fine-tuning substantially increases resemblance to the training gallery on all three resemblance metrics. Both DiSI operators partially reverse this effect while retaining generation utility: Φ_{TMV} provides stronger protection (lower CLIP and style similarity, larger LPIPS distance), whereas Φ_{TMGC} offers a milder protection–utility trade-off, preserving FID and KID near the regular-fine-tuning level.

What these numbers do and do not establish. Copyright is semantic and context-dependent. CLIP, LPIPS, and style similarity are useful proxies, but no universal threshold on them equals infringement, and we do not claim one. DiSI's reference formulation provides complementary probabilistic semantics: for a protected collection A, let $\mathsf{R} = \mathsf{X} \setminus \mathsf{A}$; if training on R lands in the desired non-copying set $\mathcal{T}$ with probability $1 - \delta_0$, then DiSI transfers that property with probability at least $1 - \delta_0 - \Delta$ by Lemma 2.3, without defining protection through an abstract similarity threshold. We also note that the artist selection in Appendix J.3 should not be read as a claim that the selected artists were absent from pre-training; our claim is restricted to the observable behaviour of the base model before fine-tuning.

J.4 Membership Inference Guarantees and Evaluation

Theoretical guarantee. The consistency guarantee of DiSI directly bounds the membership advantage. Taking $\mathsf{A} = \{x\}$ and $\mathsf{R} = \mathsf{X} \setminus \{x\}$, the two worlds differ by one source vote: an added vote if x forms its own source, and a replaced vote if x sits inside a multi-sample source, in which case the replacement extension of Theorems 3.2–3.3 (Appendix E.2) applies with c replaced by $2c$. In either case Lemma 2.3 gives $\mathsf{TV}(\mathbb{P}_{\mathsf{Alg}(\mathsf{R}\cup\{x\})} \,\|\, \mathbb{P}_{\mathsf{Alg}(\mathsf{R})}) \le \Delta$. Hence, in a balanced membership game, where a sample is included in the training set with a probability 1/2, the accuracy of any attacker is at most

$$\mathrm{Acc}^{\star}_{\mathrm{MIA}} \;\le\; \tfrac{1}{2} + \tfrac{\Delta}{2}. \tag{26}$$

Unlike the uniform, worst-case guarantee of DP, this bound is specific to the target unit and the data distribution, and can therefore be substantially tighter in the regimes DiSI targets. The value of Δ differs between the two cases: for a sample nested inside a source the relevant band is the replacement band, twice as wide as the addition band, so the exponent in Theorem 3.2 is evaluated with margin $\beta n_0 + 2c$ instead of $\beta n_0 + c$.

Setup. We complement the bound with an empirical evaluation using the likelihood-ratio attack (LiRA) (Carlini et al., 2022) on the diffusion fine-tuning task. The data pool consists of 259 images spanning 13 tags, drawn from the dataset used in the copyright experiment (Appendix J.3). Each target model is fine-tuned on 224 members, and the remaining 35 images are held out as non-members. The attack set pairs 35 sampled members with all 35 non-members, giving 70 balanced candidates. We report results over four independently seeded target replicates. Each LiRA attacker trains eight

shadow models, and the attack statistic for each candidate is the mean denoising loss over eight random seeds.

The DiSI operators aggregate votes at the tag level, so a member image changes one tag's vote; this is the replacement case of the guarantee above, and the evaluation is an empirical stress test of it (we do not compute the corresponding numerical Δ). We evaluate an *adaptive* attacker that knows the target's training algorithm: its shadow models are trained with the same procedure as the target, i.e., ordinary SGD, Φ_{TMV}, or Φ_{TMGC}. This is the strongest attacker-knowledge setting we consider. Results are reported in Table 18.

Table 18: LiRA membership inference under an adaptive, algorithm-matched attacker (mean $\pm$ std. over 4 target replicates). Lower is better for all three columns.

Target training	ROC-AUC $\downarrow$	TPR@10%FPR $\downarrow$	Balanced acc. $\downarrow$
Regular fine-tuning	0.9852 ± 0.0166	0.9556 ± 0.0629	0.9556 ± 0.0287
Fine-tuning + Φ_{TMV}	0.8284 ± 0.0329	0.4111 ± 0.1628	0.7556 ± 0.0552
Fine-tuning + Φ_{TMGC}	0.6217 ± 0.0400	0.1556 ± 0.0770	0.5472 ± 0.0140

Interpretation. Regular fine-tuning is highly vulnerable to membership inference: the adaptive attacker reaches a balanced accuracy of 0.9556 and a TPR of 0.9556 at 10% FPR. Both DiSI operators substantially reduce this leakage. Φ_{TMV} lowers the ROC-AUC to 0.8284, and Φ_{TMGC} brings the balanced accuracy close to chance (0.5472), with the TPR at 10% FPR dropping to 0.1556. These results agree with the guarantee in (26). DiSI's consistency property translates into an effective empirical defense against membership inference, and Φ_{TMGC} gives the strongest protection.