# *Catch-Only-One*: Non-Transferable Examples for Model-Specific Authorization

**Zihan Wang**[1,2,*,†] **Ethan (Zhiyong) Ma**[1,†] **Zhongkui Ma**[1] **Shuofeng Liu**[5]
**Akide Liu**[6] **Derui Wang**[2] **Minhui Xue**[2,4] **Guangdong Bai**[3,*]

[1]The University of Queensland [2]CSIRO [3]City University of Hong Kong
[4]RAIR, Adelaide University [5]RMIT University [6]Monash University

## Abstract

Recent AI regulations increasingly emphasize the need for mechanisms that preserve the utility of data for AI innovation while preventing misuse, particularly by enforcing *purpose limitation* in downstream AI applications. In practice, enforcing this principle remains challenging, as released data can be trivially fed into arbitrary models beyond its declared intent. Existing approaches attempt to mitigate this risk by either perturbing data or retraining models to limit unintended use. These strategies, however, offer no protection against inference by unknown or externally trained models, or fundamentally rely on control over the training or deployment. In this work, we introduce *non-transferable examples* (NTEs), recoded data that act as a task-level "ciphertext" decodable *only* by a designated model. Whereas adversarial examples exploit directions of high model sensitivity, NTEs leverage the complementary *insensitive* subspace. We propose a training-free, data-agnostic method that recodes data within a model-specific low-sensitivity subspace, preserving outputs for the authorized model while degrading unauthorized ones through subspace misalignment. We establish formal bounds certifying authorized-model fidelity and showing that unauthorized degradation scales with measurable spectral misalignment between models. Empirically, NTEs preserve performance across diverse vision backbones and state-of-the-art vision-language models under common preprocessing, while unauthorized models collapse even under adaptive reconstruction attacks. These results establish NTEs as a practical means to preserve intended data utility while preventing unauthorized exploitation.

: https://trusted-system-lab.github.io/model-specificity
: https://github.com/model-specific/non-transferable-examples

## 1 Introduction



*The title alludes to Joseph Heller's Catch-22, which symbolizes an unavoidable paradox. We adapt this notion in Catch-Only-One (or Catch-11) to capture the paradox that shared data may appear universally accessible yet remains usable only by a single authorized model.*

Recent regulatory initiatives such as the EU AI Act [13] and the US AI Action Plan [44], alongside industry frameworks such as DeepMind's Frontier Safety Framework [15], all enforce *purpose limitation*: data must be used strictly in accordance with its declared purpose throughout the AI lifecycle, including as inputs to AI systems at inference or generation time. Inputs such as prompts and uploaded documents must remain useful for authorized innovation while protected from misuse.

Despite this consensus, enforcement of purpose limitation is rarely satisfactory once data is released. Existing data safeguards largely rely on *administrative* controls (access restrictions, licenses, or terms

*Correspondence to: ⟨zihan.wang@uq.edu.au⟩, ⟨g.bai@cityu.edu.hk⟩. †Core contributors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of use) that cannot prevent shared data from being scraped, aggregated, and consumed by any AI application without consent. A small number of paintings can suffice to replicate an artist's style [17, 22]; medical scans shared for research can be exploited by membership inference attacks [26, 42]. The cost is tangible: a class-action lawsuit compelled Anthropic to pay over $1.5B and erase pirated data [1], and European regulators have imposed roughly €100M in fines on Clearview AI for unlawful facial recognition matching uploaded face images against a huge unlicensed biometric database [39].

This situation calls for an enforceable *post-release* guarantee: released data should remain fully usable *only* by its intended recipient (the authorized model) while withholding utility from others. The protection must not assume control over downstream training or deployment, since data owners cannot dictate who obtains the data or which models consume it, and it must be lightweight enough for real-world AI pipelines. Three lines of existing research span a trade-off (lightweight protection that fails at inference, targeted suppression that requires retraining, or strong confidentiality at prohibitive cost), but none acts directly at inference on already-released data. *Data obfuscation* [48, 52] perturbs data prior to release so that standard learning pipelines degrade or fail to converge, but offers no guarantees once data is directly consumed at inference time. *Model-side restriction* [24, 49] alters model objectives or parameters to suppress transfer in specific domains, requiring retraining with custom losses and control of the deployment pipeline. *Fully homomorphic encryption* (FHE) [18] enables inference on encrypted data, but its prohibitive computational and memory costs make it impractical for routine applications such as online media, healthcare workflows, or MLaaS services [30].

**Our work**. Once data leaves its owner, it may be consumed by countless models. We recode released content into a task-level "ciphertext"[1] that only a designated authorized model $f^\star$ can decode, giving a *post-release* technical control for purpose limitation. We call the recoded inputs *non-transferable examples* (NTEs), denoted $\tilde{x}$: the recoding constructs an authorization channel from insensitive directions of $f^\star$ and embeds model-specific signals that are benign to $f^\star$ but disruptive to others.

Our design exploits a structural property of neural networks: early layers are sensitive along only a few directions, leaving a high-dimensional *insensitive subspace* specific to $f^\star$'s weights that misaligns across models. This is the dual of the adversarial vulnerability behind adversarial examples [19, 43]: where perturbations along *sensitive* directions can drastically alter predictions, perturbations confined to the complementary insensitive set leave $f^\star$ undisturbed. Using a small probe budget, we estimate a spectral basis for $f^\star$'s insensitive subspace and recode $x$ by adding a calibrated perturbation confined to this basis ($x \to \tilde{x}$). By construction, NTEs preserve data utility for $f^\star$; the high dimensionality of the admissible set supports randomized per-sample recoding that disrupts transfer to unauthorized models and resists inversion empirically, including by adaptive adversaries trained on many $\tilde{x}$. The procedure is training-free and data-agnostic, requiring no model modification and no knowledge of downstream models, and runs as a drop-in preprocessing step with negligible overhead, making it practical for MLaaS [40] deployments where user inputs are routinely replicated and repurposed [17].

We prove two bounds for NTEs. For the authorized model, recoding perturbs first-layer outputs by at most the spectral threshold $\tau$, negligible under spectral flatness [46]. For any unauthorized model, a cross-model bound shows degradation is lower-bounded by the subspace misalignment $\alpha$, a quantity measurable from model weights alone; the wider the architectural gap, the stronger the separation guarantee. Both bounds extend to the full network via Lipschitz composition, certifying a prediction-level separation. Empirically, we evaluate NTEs across five representative vision backbones (`ResNet`, `ViT`, `SwinV2`, `DeiT`, `MambaVision`) and on `Qwen2.5-VL` and `InternVL3` over MMBench [35], spanning perception and reasoning abilities (coarse and fine-grained perception; attribute, relation, and logic reasoning). The authorized model retains near-clean accuracy while unauthorized models collapse to near-chance: at 0 dB PSNR on ImageNet, authorized ResNet-50 top-1 stays at 80.2% (from 80.3% clean) while all others drop to $\approx 0.1\%$. NTEs also resist common preprocessing and reconstruction attacks.

**Contributions**. We summarize our main contributions as follows:

- **A new problem setting.** We formulate *model-specific data authorization*: user-provided data should remain fully usable for an authorized model while withholding utility from any other model. We introduce *non-transferable examples (*NTE*s)*, a model-specific data representation that can be produced via a lightweight recoding process.

[1] We use "ciphertext" as an analogy: NTEs are obfuscated, model-specific recodings, not cryptographic encryption.

- **A formal framework and theoretical analysis.** We establish a strong formal foundation for NTEs with two bounds: authorized performance is preserved, while unauthorized output deviations are provably lower-bounded.
- **An empirical evaluation.** Across diverse model architectures and data modalities, authorized models retain full utility while unauthorized models drop to near-chance despite common preprocessing, learned reconstruction, and supervised retraining attacks, with sub-millisecond encoding.

## 2 Problem Formulation

### 2.1 Preliminaries

**Neural Networks**. A feedforward network $f : \mathbb{R}^n \to \mathbb{R}^m$ alternates linear maps with pointwise nonlinearities [37]. Convolutional layers admit equivalent fully connected forms [50] (Appendix A.3). Absorbing biases $b^{(i)}$ into $W^{(i)}$ by appending a constant 1 to $x$ yields

$$y = f(x) = \phi(W^{(L)} \cdot \phi(W^{(L-1)} \cdots \phi(W^{(2)} \cdot \phi(W^{(1)}x)) \cdots)),$$

where $\phi$ is an activation function (*e.g.,* ReLU, sigmoid, tanh).

**Definition 1** (First-layer operator)**.** *For any feedforward network $f$ with first linear layer $W^{(1)}$ followed by nonlinearity $\phi$, the* first-layer operator *is $W(f) := W^{(1)}$. We write $W^\star := W(f^\star)$, with $f^\star$ formalized below. Convolutional first layers are instantiated by their unfolded matrix and vision-transformer first layers by their input projection (Appendix A.3).*

**Nullspace**. The nullspace $\mathrm{Null}(W) = \{x \in \mathbb{R}^n \mid Wx = 0\}$ of a linear map $W$ collects directions sent to zero, so any $\delta \in \mathrm{Null}(W)$ leaves the layer output invariant: $W(x + \delta) = Wx$.

### 2.2 Model-Specific Data Authorization

We consider a supervised task with ambient input space $\mathcal{X}$ and output space $\mathcal{Y}$. A dataset $\mathcal{D}$ induces the task domains $\mathcal{X}_\mathcal{D} \subset \mathcal{X}$ and $\mathcal{Y}_\mathcal{D} \subset \mathcal{Y}$ observed in practice. Training on $\mathcal{D}$ yields a family of models $\mathcal{F}_\mathcal{D}$, where each $f \in \mathcal{F}_\mathcal{D}$ implements a mapping $f : \mathcal{X}_\mathcal{D} \to \mathcal{Y}_\mathcal{D}$. To quantify the *usability* of $f$ on labeled inputs $(x, y) \sim P_\mathcal{D}$, we adopt a label-aware performance metric $\mu : \mathcal{F}_\mathcal{D} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$ (*e.g.,* accuracy indicator for classification). The expected utility is $M(f;\, \mathcal{X}_\mathcal{D}) := \mathbb{E}_{(x,y)\sim P_\mathcal{D}}[\mu(f, x, y)]$. Larger values of $\mu$ indicate better usability.

At inference on new data outside the training set, the data owner (defender) seeks to release content (*e.g.,* images, speech, documents) that is correctly processed by a designated authorized model $f^\star$ while remaining unusable to any unauthorized model $f' \in \mathcal{F}_\mathcal{D} \setminus \{f^\star\}$.

We formulate *model-specific data authorization*: preserve the usability of $f^\star$ while degrading that of any unauthorized model, without retraining and without assumptions on unauthorized models. Let $\tilde{x} = \mathcal{T}(x; \omega)$ be a randomized input recoding with $\mathcal{T} : \mathcal{X} \times \Omega \to \mathcal{X}$ and per-release seed $\omega \sim \mathbb{Q}$. The recoded set $\widetilde{\mathcal{X}}_\mathcal{D} := \{\mathcal{T}(x; \omega) : x \in \mathcal{X}_\mathcal{D},\, \omega \sim \mathbb{Q}\}$ defines *non-transferable examples* (NTEs), a model-specific data representation tailored to $f^\star$; each $\tilde{x}$ inherits the label $y$ of its underlying clean $x$, and $M(f;\, \widetilde{\mathcal{X}}_\mathcal{D}) := \mathbb{E}_{(x,y)\sim P_\mathcal{D},\, \omega\sim\mathbb{Q}}[\mu(f, \mathcal{T}(x; \omega), y)]$. Formally:

**Problem** ($(\rho, \gamma)$-Model-Specific Data Authorization)**.** *Given an authorized-utility tolerance $\rho \geq 0$ and an unauthorized separation margin $\gamma > 0$, a randomized recoding $\mathcal{T} : \mathcal{X} \times \Omega \to \mathcal{X}$ is* $(\rho, \gamma)$-model-specific *to $f^\star$ if, for every $f' \in \mathcal{F}_\mathcal{D} \setminus \{f^\star\}$,*

$$\textit{(authorized-utility retention)} \quad M(f^\star;\, \mathcal{X}_\mathcal{D}) - M(f^\star;\, \widetilde{\mathcal{X}}_\mathcal{D}) \leq \rho, \tag{1}$$

$$\textit{(unauthorized-utility degradation)} \quad M(f^\star;\, \widetilde{\mathcal{X}}_\mathcal{D}) - M(f';\, \widetilde{\mathcal{X}}_\mathcal{D}) \geq \gamma. \tag{2}$$

### 2.3 Threat Model

The data owner (defender), aligned with the data processor, has white-box access to $f^\star$ and uses task-domain data to instantiate the recoding $\mathcal{T}$. Model-specific parameters of $\mathcal{T}$ (*e.g.,* spectral basis, thresholds) remain private even when its algorithmic form is public. The recoding is lightweight and runs before release, producing $\tilde{x} = \mathcal{T}(x; \omega)$ that is shared or consumed downstream without disclosing $f^\star$. The defender has no access to unauthorized models.

The adversary operates *unauthorized* models $f' \in \mathcal{F}_{\mathcal{D}} \setminus \{f^\star\}$ and receives only $\tilde{x}$. Before inference, they may apply an input-side preprocessing operator $\mathcal{A} \in \mathbb{A}$, where $\mathbb{A}$ is a class of standard preprocessing pipelines, yielding $\tilde{x}' := \mathcal{A}(\tilde{x})$. The exact parameters of $\mathcal{T}$ are hidden. The adversary's objective is to *maximize* the task metric $\mu(f', \tilde{x}', y)$ under fixed compute or query budgets. Specifically, we consider three adversary classes: (*i*) *General Adversary* ***(GA)****:* any unauthorized $f'$ with arbitrary architecture and parameters; (*ii*) *Transfer-match Adversary* ***(TA)****:* a GA sharing the architecture of $f^\star$ but with different weights; (*iii*) *Adaptive Adversary* ***(AA)****:* a GA/TA that optimizes the prediction pipeline, including learning $\mathcal{A} \in \mathbb{A}$ and updating $f'$ (via *fine-tuning* or *retraining from scratch*), to maximize $\mu(f', \tilde{x}', y)$.

**Scope**. This setting is practical as the enforcement process is executed locally by the data owner (or delegated to an authorized third party), requiring the model provider to disclose only minimal information. With white-box access to $f^\star$, the owner integrates $\mathcal{T}$ as an upstream service in the data pipeline, ensuring only $\tilde{x}$ is released. In practice, this can be deployed as a middleware gateway in commercial pipelines (*e.g.,* an MLaaS front-end) that recodes inputs prior to storage or sharing, requiring no changes to downstream training procedures or model weights. From a governance perspective, this establishes an enforceable *ex ante* technical control for purpose-limitation obligations in settings where *ex post* monitoring of downstream reuse is intractable. Neither the weights of $f^\star$ nor the model-specific parameters of $\mathcal{T}$ are disclosed to downstream recipients. This work focuses on an instantiation with an elegant closed-form algorithm and formal bounds; well-studied query-based sensitivity estimation [6, 51] offers a route to black-box instantiations outside the present scope.

**Design Justifications**. Limited white-box access to $f^\star$ (*e.g.,* gradients or partial weights) is highly consistent with delegated auditing and transparency obligations reflected in EU AI Act Annex VII [13], SR 26-2 [4], and GDPR Art. 15 [14] (Appendix E.2). NTEs are also deliberately not human-readable: $\tilde{x}$ acts as a task-level "ciphertext" that only $f^\star$ can decode (Appendix E.3).

## 3 Our Method

Neural networks typically begin with a linear feature extractor (*e.g.,* a convolution with weight sharing or a token embedding), which reduces redundancy because input coordinates are often correlated (see Appendices A.1 and A.2). This motivates *input-side* perturbations that lie in an *insensitive* subspace of the authorized model's first linear map: these directions are nearly inert for the authorized model but, due to subspace misalignment across models, can induce nontrivial changes for others.

### 3.1 Insensitive Subspace Identification

Let $W := W^\star$ denote the first linear transformation of the authorized model (Definition 1, bias omitted). Since input dimensionality is typically high, $W$ has a nontrivial nullspace, and any $\delta \in \mathrm{Null}(W)$ satisfies $W(x+\delta) = Wx$, hence $f(x+\delta) = f(x)$.

We relax exact nulling to controlled feature deviation, requiring $W\delta \approx 0$ rather than $W\delta = 0$; perturbations in such an insensitive subspace induce only small first-layer changes, especially under intervening nonlinearities (*e.g.,* ReLU truncation, sigmoid/tanh saturation). For $W \in \mathbb{R}^{m\times n}$ with singular value decomposition (SVD) $W = USV^\top$, the nullspace is spanned by right singular vectors with zero singular values; we extend this to the *$\tau$-insensitive subspace*, the span of right singular vectors whose singular values are at most a threshold $\tau > 0$. We formalize it as follows.

**Definition 2** ($\tau$-insensitive Subspace)**.** *Let $W \in \mathbb{R}^{m\times n}$ have SVD, $W = USV^\top$, with singular values ordered $s_1 \le s_2 \le \cdots \le s_n$. Given a spectral threshold $\tau > 0$, the $\tau$-insensitive subspace is*

$$\mathrm{Ins}_\tau(W) = \mathrm{span}\{v_1, v_2, \ldots, v_k\}, \tag{3}$$

*where $k = \max\{i \mid s_i \le \tau\}$, and $v_i$ denotes the $i$-th column of $V$ corresponding to $s_i$ in $S$.*

This construction naturally captures the nullspace, and $\mathrm{Ins}_\tau(W) \supseteq \mathrm{Null}(W)$ for any $\tau \ge 0$.

### 3.2 Non-Transferable Examples

Inheriting Definition 2, let $V^{(s)} \in \mathbb{R}^{n\times k}$ collect the right singular vectors of $W$ corresponding to singular values $s_i \le \tau$, so $\mathrm{range}(V^{(s)}) = \mathrm{Ins}_\tau(W)$. We sample a vector $z \in \mathbb{R}^k$ (*e.g., i.i.d.*

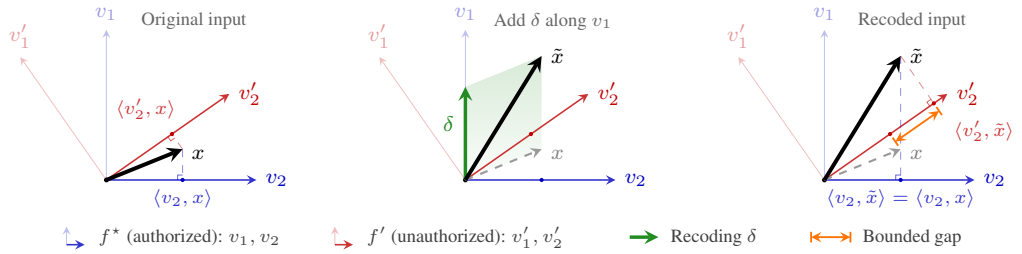


Figure 1: 2D intuition: recoding $\delta$ along $f^\star$'s insensitive direction $v_1$ preserves the $v_2$-component (Thm. 1) while sliding the $v_2'$-component by $\sin\theta\,\|\delta\|$ with $\theta = \angle(v_2, v_2')$, degrading $f'$ (Thm. 2).

Gaussian, structured pattern, or content-dependent code) and lift it into $\mathrm{Ins}_\tau(W)$ via $\delta = V^{(s)}z$. The perturbation $\delta$ is added to the original input $x$ to obtain a recoded input $\tilde{x} = x + \delta$, which forms an NTE. Since $V^{(s)}$'s columns are orthonormal and their associated singular values are bounded by $\tau$, $\|W\delta\|_2 \le \tau\|z\|_2$. By the distributive law of matrix multiplication, $W\tilde{x} = W(x+\delta) = Wx + W\delta$, where $W\delta$ is small and thus $W\tilde{x}$ is close to $Wx$. Figure 1 visualizes this construction. For any unauthorized model $f'$ with first-layer operator $W'$, the same $\delta$, inert to $f^\star$'s sensitive directions, generally lands on $W'$'s sensitive directions whenever the two right singular bases disagree, previewing the cross-model degradation formalized in Section 4.2.

**Practical Implementation**. The procedure is architecture-agnostic. Fully connected fronts can directly use their first weight matrix as $W$. For convolutional fronts, we apply the construction to the linearized operator of the first convolution (*e.g.,* `nn.Unfold`) to obtain $W$; after synthesis, the perturbation is folded back to the native input layout. For Transformer-style text models (*e.g.,* BERT [9]), we take $W$ to be the input projection of the first multi-head self-attention block *after* the embedding layer (*e.g.,* the `QKV` input projection or its concatenated form) and apply the same spectral construction. Appendices A.3 and A.4 establish the underlying linear equivalences. The generation of the vector $z$ can be stochastic (*e.g.,* seeded per instance) or deterministic (*e.g.,* fixed codes per class or per client). We rescale $\delta$ by a factor $\lambda > 0$ that sets its amplitude. The parameters $\tau$ and $\lambda$ govern the trade-off between authorized-utility retention and unauthorized-utility degradation: larger $\tau$ tends to increase the number of perturbed input dimensions, while larger $\lambda$ increases the perturbation magnitude. In practice, the permissible insensitive space is deliberately loose, offering substantial flexibility in synthesis, as empirically explored in Section 5.

## 4 Theoretical Properties of NTEs

We prove two output-norm bounds at the first-layer operator $W$: retention for the authorized model (Section 4.1) and degradation for any unauthorized model (Section 4.2). A network-wide propagation result (Proposition 1) then lifts both through the full network via layer-wise composition, instantiating Eq. (1) and supplying the norm-level separation underlying Eq. (2). We defer full proofs to Appendix D, where the cross-model separation is derived from a quantifiable misalignment condition with explicit, finite-sample constants.

**Standing Assumptions and Notation**. Let $W = W^\star$ (Definition 1) have SVD $USV^\top$ with $s_1 \le \cdots \le s_n$; partition $V = [V^{(s)}, V^{(l)}]$ at $\tau$, so $V^{(s)} \in \mathbb{R}^{n\times k}$ spans $\mathrm{Ins}_\tau(W)$. The recoding of Section 3.2 sets $z \sim \mathcal{N}(0, \sigma^2 I_k)$, $\delta = V^{(s)}z$, $\tilde{x} = x + \delta$. For $f^\star = g_L \circ \cdots \circ g_1$ and unauthorized $f' = g_L' \circ \cdots \circ g_1'$, each post-first $g_i$ is $L_i^\star$-Lipschitz and $g_i'$ is $\nu_i'^\star$-co-Lipschitz on the trajectory pair $(x, \tilde{x})$; write $L^\star_{\mathrm{prod}} := \prod_{i=2}^L L_i^\star$ and $\nu^\star_{\mathrm{prod}} := \prod_{i=2}^L \nu_i'^\star$. Assumption 1 is invoked where stated.

### 4.1 Authorized-Utility Retention

We first bound the first-layer deviation induced by the recoding $\tilde{x} = x + \delta$.

**Theorem 1** (Bounding Authorized Utility)**.** *Under the standing assumptions, for any $t > 0$, with probability at least $1 - 2k\sigma^4/t^2$,*

$$\|W\tilde{x} - Wx\|_2 < \tau\sqrt{k\sigma^2 + t}. \tag{4}$$

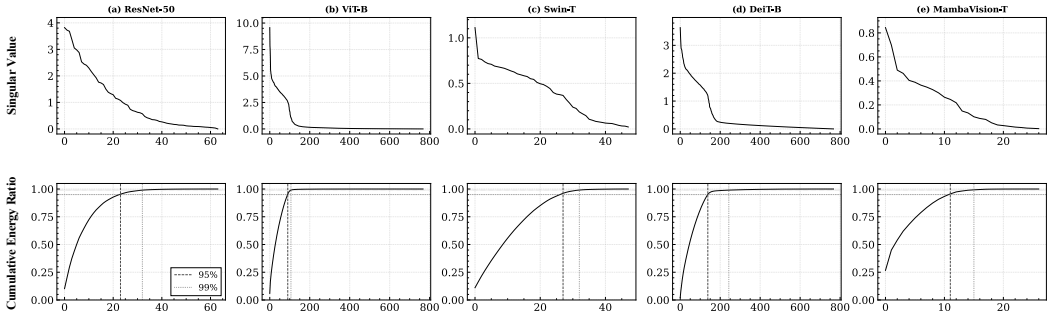


Figure 2: First-layer singular values (top) and cumulative energy share (bottom) across five models.

*Proof.* Orthonormality of $V$ gives $(V^{(s)})^\top W^\top W V^{(s)} = (S^{(s)})^2$, with $S^{(s)} = \mathrm{diag}(s_1, \ldots, s_k)$ (zeros included). Since $s_q \le \tau$ for $q \le k$, $\|W\delta\|_2 = \|S^{(s)} z\|_2 \le \tau \|z\|_2$. A Chebyshev tail bound on $\|z\|_2^2$ (mean $k\sigma^2$, variance $2k\sigma^4$) yields $\|z\|_2 < \sqrt{k\sigma^2 + t}$ at the stated probability, giving Eq. (4); the Chebyshev computation that supplies the probability constant is in Appendix D.1. □

The threshold $\tau$ and noise scale $\sigma$ act as direct knobs on the first-layer deviation. Spectral flatness (Assumption 1) makes $\mathrm{Ins}_\tau(W)$ high-dimensional, so the bound is non-vacuous in practice; Figure 2 and Appendix B.1 verify this empirically across the backbones we examine. Proposition 1 below lifts the bound to the full network and to expected utility.

### 4.2 Unauthorized-Utility Degradation

The cross-model bound follows from a measurable misalignment between the authorized and unauthorized spectral bases. We assume the *misalignment condition* $\|(V'^{(l)})^\top V^{(s)}\|_2 \ge \alpha$ for some $\alpha > 0$, verified across all backbones. Figure 1 (Section 3) gives the corresponding 2D picture, where $\alpha$ reduces to $\sin\theta$ for $\theta = \angle(v_2, v_2')$ and generalizes it to higher dimensions.

**Theorem 2** (Bounding Unauthorized Utility)**.** *Let $W' \in \mathbb{R}^{m\times n}$ be the first-layer operator of an unauthorized model with SVD $W' = U' S' V'^\top$ and singular values $s'_1 \le \cdots \le s'_n$. Partition $V' = [V'^{(s)}, V'^{(l)}]$ at index $k$, with $V'^{(l)} \in \mathbb{R}^{n\times(n-k)}$, and set $s'_{\min} := s'_{k+1}$. Under the standing assumptions and the misalignment condition, for any $t > 0$ and $c \in (0, \sqrt{\pi/2}]$, with probability at least $1 - 2k\sigma^4/t^2 - c\sqrt{2/\pi}$,*

$$\|W'\tilde{x} - W'x\|_2 - \|W\tilde{x} - Wx\|_2 \;\ge\; c\alpha\, s'_{\min}\sigma - \tau\sqrt{k\sigma^2 + t}. \tag{5}$$

*Proof.* Set $A := (V'^{(l)})^\top V^{(s)}$. Orthogonality of $U'$ and the lower bound $s'_{k+i} \ge s'_{\min}$ give $\|W'\delta\|_2 \ge s'_{\min}\|Az\|_2$. Letting $p_1$ be the top right singular vector of $A$, $\|Az\|_2 \ge \|A\|_2 |p_1^\top z|$ with $p_1^\top z \sim \mathcal{N}(0, \sigma^2)$; combining $\|A\|_2 \ge \alpha$ with Gaussian anti-concentration gives $\|Az\|_2 \ge c\alpha\sigma$ with probability at least $1 - c\sqrt{2/\pi}$. A union bound with Theorem 1 yields Eq. (5); Gaussian anti-concentration on $p_1^\top z$ and joint-failure accounting against retention appear in Appendix D.2. □

The same recoding $\delta$ is suppressed by $W$'s small singular values and amplified by $W'$'s large ones whenever the two bases disagree on $\mathrm{Ins}_\tau(W)$. The size of the gap is governed by the misalignment $\alpha$ and the smallest sensitive gain $s'_{\min}$ of $W'$, both measured across architectures.

**Network-Wide Propagation**. We lift the first-layer bounds to the full network via layer-wise composition, then derive the expected-utility form of Eqs. (1), (2).

**Proposition 1** (Network-wide propagation)**.** *Under the standing assumptions and the misalignment condition, with probability at least $1 - 2k\sigma^4/t^2 - c\sqrt{2/\pi}$ (for $t > 0$, $c \in (0, \sqrt{\pi/2}]$): (a) Lipschitz composition lifts Theorem 1 to a network-wide authorized-retention bound scaled by $L^\star_{\mathrm{prod}}$, and (b) co-Lipschitz composition lifts Theorem 2 to a network-wide cross-model separation scaled by $\nu^\star_{\mathrm{prod}}$. Explicit forms (Eqs. (6) and (7)) are deferred to Appendix D.*

*Proof.* Part (a) follows by Lipschitz composition through $g_2, \ldots, g_L$ from Theorem 1; part (b) by co-Lipschitz composition through $g'_2, \ldots, g'_L$ applied to Theorem 2, joined by a union bound on the same draw $z$. Appendices D.3, D.4 give the layer-wise induction and joint-failure union bound. □

The constants $L_i^\star$ and $\nu_i'^\star$ are *local* to the trajectory pair $(x, \tilde{x})$. Appendix B.2 (Table 7) confirms that the authorized deviations avoid the usual product explosion. Comparing the two parts of Proposition 1 yields the design rule $\nu^\star_{\text{prod}}\, c\alpha\, s'_{\min}\, \sigma > L^\star_{\text{prod}}\, \tau\sqrt{k\sigma^2 + t}$. Connecting back to Eqs. (1) and (2), we specialize $\mu$ to top-1 accuracy: let $d(f^\star, x) := [f^\star(x)]_{\hat{y}} - \max_{y' \neq \hat{y}} [f^\star(x)]_{y'}$ be the logit margin ($\hat{y} = \arg\max_y [f^\star(x)]_y$). If the clean-data margin satisfies $\Pr_{x \sim P_{\mathcal{D}}}[d(f^\star, x) > \sqrt{2}\, L^\star_{\text{prod}}\, \tau\sqrt{k\sigma^2 + t}] \geq 1 - \eta$, then Proposition 1(a) and a Cauchy–Schwarz step ($\|e_{\hat{y}} - e_{y'}\|_2 = \sqrt{2}$) together preserve the top-1 prediction under recoding; the failure-event mass then gives $M(f^\star;\, \mathcal{X}_{\mathcal{D}}) - M(f^\star;\, \widetilde{\mathcal{X}}_{\mathcal{D}}) \leq \eta + 2k\sigma^4/t^2$, instantiating Eq. (1) with $\rho = \eta + 2k\sigma^4/t^2$. Proposition 1(b) yields the norm-level separation underlying Eq. (2), realized as a prediction-level gap across architectures in Table 1.

## 5 Experiments

NTEs preserve authorized utility while rendering unauthorized models unusable. Section 5.1 establishes cross-model non-transferability against *GA* and *TA*; Section 5.2 positions NTEs against three authorization-related mechanisms; Section 5.3 stress-tests across modalities and against the adaptive adversary *AA* via preprocessing, reconstruction, and learned adaptation.

**Setup**. We evaluate five ImageNet pretrained backbones spanning classic and recent designs: *ResNet-50* [21], *ViT-base-patch16-224* [10], *SwinV2-tiny-patch4-window8-256* [36], *DeiT-base-patch16-224* [45], and *MambaVision-T-1K* [20]. Experiments use CIFAR-10 [28] and ImageNet-1K [41]; CIFAR-10 results follow 10 epochs of fine-tuning from the ImageNet-1K checkpoints. "Baseline" rows in the tables denote clean accuracy under our evaluation pipeline. We also evaluate the generative vision-language models (autoregressive text decoders) *Qwen2.5-VL-3B-Instruct* [2] and *InternVL3-1B* [53] on MMBench [35], which spans perception and reasoning abilities (coarse and fine-grained perception; attribute, relation, and logic reasoning). We parameterize perturbation strength by a PSNR setting [25] (Appendix E.3); lower PSNR is stronger perturbation. For most backbones, unauthorized accuracy collapses around 20–25 dB while authorized accuracy is largely preserved (Figure 3). We therefore use a conservative 20 dB setting unless otherwise specified. Recoding samples $z$ with i.i.d. Gaussian entries and projects onto the $\tau$-insensitive directions; $\tau$ is set per backbone for Table 1, and Figure 3 shows separation even with a shared $\tau$.

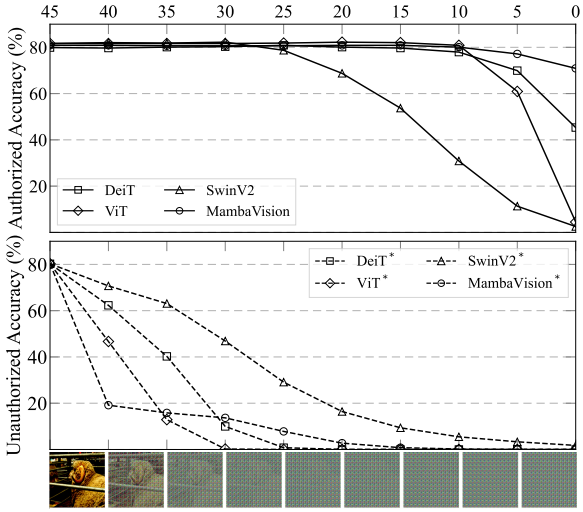


Figure 3: Authorized vs. unauthorized accuracy on target-recoded inputs across perturbation strength.

### 5.1 Cross-Model Non-Transferability

For each authorized target $f^\star$, we report authorized accuracy on $\tilde{x} = \mathcal{T}(x)$ and unauthorized accuracies on the same $\tilde{x}$. The sweep in Figure 3 uses ImageNet, with ResNet-50 as the unauthorized comparator. Table 1 reports a $5 \times 5$ cross-architecture matrix on CIFAR-10 and ImageNet. At 20 dB PSNR, NTEs keep authorized performance close to clean (*e.g.,* 98.8% to 98.7% on CIFAR-10; 80.3% to 80.2% on ImageNet), while unauthorized accuracy drops sharply, with off-diagonals at 5.5–20.6% on CIFAR-10 and 0.0–12.9% on ImageNet. SwinV2 shows the largest authorized drop on ImageNet (80.9%→71.7%); Figure 3 shows substantial separation between authorized and unauthorized accuracy, with clear room for further parameter optimization. Holding the architecture

fixed but changing the weights gives accuracies of 8.2–21.0% on CIFAR-10 and 0.0–1.2% on ImageNet (Appendix C.1, Table 8).

Table 1: Cross-model accuracy at 20 dB PSNR. Rows: authorized target $f^\star$; columns: evaluated model. Diagonals are authorized; off-diagonals unauthorized.

| | CIFAR-10 | | | | | ImageNet | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ResNet-50 | ViT-B | SwinV2-T | DeiT-B | MambaVision-T | ResNet-50 | ViT-B | SwinV2-T | DeiT-B | MambaVision-T |
| Baseline | 98.2% | 98.8% | 96.1% | 96.1% | 96.9% | 80.3% | 81.6% | 80.9% | 79.9% | 82.4% |
| ResNet-50 | **97.7%** | 12.5% | 13.6% | 9.4% | 10.0% | **80.2%** | 0.0% | 0.1% | 0.1% | 0.1% |
| ViT-B | 10.5% | **98.7%** | 9.3% | 12.1% | 9.7% | 0.0% | **81.3%** | 0.0% | 0.0% | 0.0% |
| SwinV2-T | 15.6% | 11.7% | **88.4%** | 20.6% | 18.8% | 9.1% | 4.3% | **71.7%** | 7.4% | 12.9% |
| DeiT-B | 9.4% | 9.8% | 9.8% | **96.1%** | 5.5% | 0.0% | 0.0% | 0.0% | **79.3%** | 1.0% |
| MambaVision-T | 17.6% | 7.0% | 13.7% | 11.3% | **94.5%** | 5.8% | 0.0% | 1.5% | 0.7% | **81.0%** |

### 5.2 Positioning Among Authorization-Related Methods

NTEs provide inference-time model-specific authorization. We position them against three mechanisms with different threat models, not like-for-like baselines: Differential Privacy (DP) [11] bounds training-time leakage; Fully Homomorphic Encryption (FHE) provides confidentiality via encrypted inference; ALGOSPEC [32], the closest comparator, gates unauthorized inference via low-degree polynomial surrogates. Table 2 reports ResNet-50 and ViT-B on CIFAR-10 and ImageNet.

Table 2: Positioning against authorization-related methods.

| | CIFAR-10 | | | | | ImageNet | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Plain | DP | FHE[3] | ALGOSPEC | NTE (Ours) | Plain | DP | FHE[3] | ALGOSPEC | NTE (Ours) |
| ResNet-50 | 98.2% | 59.8%[1] | 87.8%[3] | 6.4% | 97.7% | 80.3% | 63.1%[1] | – | 0.1% | 80.2% |
| ViT-B | 98.8% | –[1] | – | 10.0% | 98.7% | 81.6% | –[1] | – | 0.0% | 81.3% |
| Protection | ✗ | ✗[2] | ✓ | ✗[2] | ✓ | ✗ | ✗[2] | ✓ | ✗[2] | ✓ |

[1] The evaluated DP pipeline struggles with batch norm and does not support multi-head attention in Transformers.
[2] The authorized model performance is significantly impacted.
[3] FHE inference at this scale is computationally prohibitive; reported figures are from the cited public ResNet-20 benchmark.

For DP training, we follow Li et al. [31] using IBM `DiffPrivLib` [23]. Authorized accuracy drops sharply, by more than 30% on CIFAR-10 and around 20% on ImageNet for ResNet-50, due to interactions with batch normalization. The pipeline also does not support transformer variants, limiting applicability. For FHE, we adopt a CKKS encrypted-inference configuration following TenSEAL [3]; running it at our scale is computationally infeasible, with single-image latency exceeding 30 minutes. We therefore cite published ResNet-20 CIFAR-10 results [38], which preserve clean authorized accuracy but incur prohibitive cost. For ALGOSPEC, polynomial approximation of modern deep networks accumulates error with depth and width, driving authorized accuracy toward chance on both datasets. NTEs, by contrast, are a lightweight input-side recoding tied to the target model: under matched conditions they preserve authorized accuracy and drive mean unauthorized accuracy to chance on both datasets, with negligible overhead.

### 5.3 Practicality of NTEs

**Generative Vision-Language Models**. We extend the construction to generative vision-language models, with `InternVL3` as authorized and `Qwen2.5-VL` as unauthorized, evaluated on MMBench [35] across capability dimensions AR, CP, FP-C, FP-S, LR, and RR. Table 3 shows that authorized performance remains close to the clean baseline across all dimensions while unauthorized accuracy collapses (overall 78.8% to 18.3%). Figures 5, 6a, and 6b show that the unauthorized model perceives NTEs as random noise.

Table 3: Model-specific Authorization on MMBench for VLMs.

| | InternVL3-1B (authorized) | | | | | | | Qwen2.5-VL-3B (unauthorized) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | AR | CP | FP-C | FP-S | LR | RR | Overall | AR | CP | FP-C | FP-S | LR | RR |
| Baseline | 72.7% | 77.4% | 82.4% | 58.7% | 79.4% | 50.3% | 66.8% | 78.8% | 81.3% | 83.1% | 69.2% | 84.9% | 66.5% | 75.4% |
| NTE (Ours) | 72.6% | 77.8% | 82.0% | 58.3% | 78.9% | 50.9% | 67.3% | 18.3% | 29.2% | 15.6% | 17.0% | 12.8% | 17.3% | 22.3% |

**Robustness to Preprocessing**. Common distortions leave the gap intact: on ViT-B/CIFAR-10, authorized NTE-accuracy stays within 1–3 pp of clean under random 10%–40% crops and at 68.4% under JPEG re-encoding to quality 75, with unauthorized models at random-guess in both settings; the same pattern holds across the VLM pipeline (Table 3, Appendix E.4).

**Adaptive Attacks**. We evaluate two attackers against NTEs. An SR-ResNet [29] reconstruction attacker trained on recoded-to-recoded (Noise2Noise) or recoded-to-clean (Noise2Clean) pairs improves PSNR by at most 0.7 dB (Appendix E.5). The stronger adversary *AA* tries to *learn around* NTEs by retraining an unauthorized model on encoded inputs. We grant *AA* maximally favorable capabilities: clean labeled pairs $(x, y)$ from the same task distribution, and an arbitrarily large *matched* encoded corpus produced by replicating the NTE pipeline and pairing each $\tilde{x}$ with the original label. This removes the usual obstacles to adaptation, namely distribution shift, label scarcity, and limited encoded data; the attacker trains directly on the inference-time format under full supervision.

Table 4 summarizes the results. All five attacker backbones evaluated on NTEs from the authorized ViT-B collapse to random-guess (row "Attacker"), including a same-architecture ViT-B attacker: matching the architecture is not enough without the exact authorized weights. We then permit end-to-end adaptation by applying the same NTE procedure to every training example and fine-tuning each backbone with full-parameter updates for 3 epochs at batch size 64 and learning rate $5 \times 10^{-4}$. Even under this distribution-matched, fully supervised adaptation (row "NTE-Attacker"), performance on NTEs remains at random-guess, showing that specificity holds against our strongest adversary.

Table 4: Learned adaptation attack against NTEs.

| Attacker Model | ResNet-50 | ViT-B | SwinV2-T | DeiT-B | MambaVision-T |
|---|---|---|---|---|---|
| Clean | 98.2% | 98.8% | 96.1% | 96.1% | 96.9% |
| Attacker | 11.1% | 10.9% | 10.5% | 13.8% | 10.9% |
| NTE-Attacker | 10.3% | 10.7% | 10.3% | 10.9% | 11.2% |

### 5.4 Runtime Analysis

NTE generation is lightweight enough for real-time inference and large-scale deployment. The only nontrivial step is a one-time SVD of the first-layer per-patch map $W \in \mathbb{R}^{m \times n}$ ($n$=147 for ResNet-50, $n$=768 for ViT-B/DeiT-B) at cost $O(\min\{m^2 n, mn^2\})$, amortized over all releases; recoding then samples $z \in \mathbb{R}^k$, computes $\delta = V^{(s)} z$ at cost $O(nk)$, and sets $\tilde{x} = x + \delta$. On a single RTX A6000 (FP32, batch 1, host–device transfer included), encoding is sub-millisecond ($\approx 0.11$ ms per $224 \times 224 \times 3$ image; Table 5), negligible against a forward pass; FLOPs and per-backbone breakdown in Appendix C.3.

Table 5: NTE construction time.

| Model | SVD (ms)[1] |
|---|---|
| ResNet-50 | 1.8 |
| ViT-B | 144.8 |
| SwinV2-T | 0.7 |
| DeiT-B | 148.3 |
| MambaVision-T | 0.2 |
| Encoding / image[2] | 0.11 ms |

[1] Compute **once** per model.
[2] For a $224 \times 224 \times 3$ image.

## 6 Related Work

**Watermarking and Content Provenance**. Watermarking embeds imperceptible signals to assert ownership or enable tracing [8, 54], and provenance standards such as C2PA Content Credentials attach cryptographically signed, tamper-evident origin and edit metadata [7]. Yet watermarks can be removed or forged under realistic adversaries [5] and, even when intact, only describe content rather than gate its use: a watermarked input remains consumable by any model.

**Learning-Time Defenses**. A complementary family intervenes during training: unlearnable [34, 48] and ungeneralizable [52] examples add bilevel-crafted perturbations that block standard learners from generalizing, with the latter preserving learnability for a designated learner. Non-transferable training alters objectives or parameterizations to suppress cross-model or cross-domain transfer [24, 49]. All presume training-pipeline influence and offer no leverage when models use released data at inference.

**Confidentiality Methods**. Differential privacy bounds any record's influence on the trained model [12], while fully homomorphic encryption keeps inputs encrypted under computation [16, 18, 27, 30] at latency and memory costs precluding routine media or MLaaS workloads [30, 38]. Neither directly provides model-specific authorization for released inputs.

## 7 Discussion and Conclusion

We introduce NTEs, a training-free approach to model-specific data authorization that exploits the authorized model's insensitive directions. Our spectral analysis bounds output deviations and links cross-model separation to subspace misalignment. Experiments across five vision backbones and two vision-language models show that authorized utility is preserved while unauthorized models are disrupted, even under preprocessing, reconstruction, and adaptation attacks. After a one-time spectral computation, encoding takes less than a millisecond per image.

Future enforcement could estimate these directions through model queries alone, without direct access to model weights. Yet adversaries with similar query budgets might do the same. Our lightweight construction nevertheless allows data owners to vary the selected directions, perturbation strength, and random codes across releases. Beyond authorization, the same simple principle could find use in many other domains. One example is model-specific compression, where video or 3D data could be compressed by discarding information that the authorized model does not need.

## 8 Ethical Considerations and Reproducibility

### 8.1 Ethical Considerations

All experiments use only publicly available benchmark datasets (*e.g.,* CIFAR-10, ImageNet-1K, and MMBench) and standard pretrained checkpoints. We do not collect, infer, or expose personally identifiable information (PII), nor link data to real individuals. This work does not constitute human-subjects research and does not require IRB approval.

**Stakeholders**. The main stakeholders are *data owners* seeking control over downstream reuse, *data processors* (model owners or service providers) considering deployment, and the *research and policy community* studying data governance.

**Intended Benefits and Potential Misuse**. NTEs are intended to restrict unauthorized secondary use of released inputs. They could also be used to impose opaque access restrictions or impede legitimate interoperability and auditing. Deployment should therefore specify who is authorized and provide clear disclosure and oversight.

**Harms, Fairness, and Limitations**. Our evaluation focuses on utility preservation and degradation across models and tasks, not demographic fairness. We introduce no new sensitive attributes and do not train on protected labels, but real-world deployment could interact with biases in the underlying model or dataset. Sending recoded inputs to the wrong model can cause silent failures. Deployments therefore need access control, model attestation, and clear interfaces.

### 8.2 Reproducibility and Artifacts

1. **Public datasets.** We use only public datasets: CIFAR-10 and ImageNet-1K for image classification, MMBench for vision-language evaluation, and GLUE for language understanding. Each is obtained from its official source under its standard license.
2. **Source code.** Code is publicly available at https://github.com/model-specific/non-transferable-examples.
3. **Models and checkpoints.** We use publicly available pretrained checkpoints for vision backbones (ResNet-50, ViT-B, SwinV2-T, DeiT-B, MambaVision-T), vision-language models (InternVL3-1B and Qwen2.5-VL-3B), and language models (BERT and RoBERTa). We do not redistribute third-party datasets or checkpoints.

## Acknowledgments and Disclosure of Funding

We thank all anonymous reviewers for their constructive feedback. We also thank Dr. Naipeng Dong and Dr. Dongge Liu for their insightful discussions. Zihan Wang is in part supported by the Google PhD Fellowship.

## References


[1] Associated Press. AI startup Anthropic agrees to pay $1.5bn to settle book piracy lawsuit. The Guardian. https://www.theguardian.com/technology/2025/sep/05/anthropic-settlement-ai-book-lawsuit, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[3] Ayoub Benaissa, Bilal Retiat, Bogdan Cebere, and Alaa Eddine Belfedhal. TenSEAL: A library for encrypted tensor operations using homomorphic encryption, 2021. URL https://arxiv.org/abs/2104.03152.

[4] Board of Governors of the Federal Reserve System. Supervisory letter SR 26-2: Revised Guidance on Model Risk Management. Supervisory letter, Federal Reserve System, 2026. URL https://www.federalreserve.gov/supervisionreg/srletters/sr2602.htm. Attachment jointly issued with the OCC and FDIC.

[5] Tu Bui, Shruti Agarwal, and John Collomosse. TrustMark: Robust watermarking and watermark removal for arbitrary resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18629–18639, 2025.

[6] Yi Cai and Gerhard Wunder. On gradient-like explanation under a black-box setting: When black-box explanations become as good as white-box. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5360–5382. PMLR, 2024. URL https://proceedings.mlr.press/v235/cai24h.html.

[7] Coalition for Content Provenance and Authenticity (C2PA). Content credentials: C2PA technical specification, September 2024. URL https://spec.c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html. Version 2.1 (September 2024).

[8] Ingemar J. Cox, Matthew L. Miller, and Jeffrey A. Bloom. *Digital Watermarking*. Morgan Kaufmann, 2002. ISBN 978-1-55860-714-9.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[11] Cynthia Dwork. Differential privacy. In *Encyclopedia of Cryptography, Security and Privacy*, pages 649–652. Springer Nature Switzerland, 2025.

[12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042.

[13] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. OJ L, 2024/1689, 12.7.2024, in force 1 August 2024.

[14] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L 119:1–88, 2016. URL http://data.europa.eu/eli/reg/2016/679/oj. Article 15(1)(h).

[15] Four Flynn, Helen King, and Anca Dragan. Strengthening our Frontier Safety Framework. https://deepmind.google/discover/blog/strengthening-our-frontier-safety-framework/, 2025.

[16] Lars Folkerts, Charles Gouert, and Nektarios Georgios Tsoutsos. REDsec: Running Encrypted Discretized Neural Networks in Seconds. In *Network and Distributed System Security Symposium (NDSS)*, pages 1–17, 2023.

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The 11th International Conference on Learning Representations (ICLR)*, 2023.

[18] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 169–178, 2009. ISBN 9781605585062. doi: 10.1145/1536414.1536440.

[19] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[20] Ali Hatamizadeh and Jan Kautz. MambaVision: A hybrid Mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 25261–25270, 2025.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[22] Melissa Heikkilä. This artist is dominating AI-generated art, and he's not happy about it. MIT Technology Review, 2022. URL https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/.

[23] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], 2019.

[24] Ziming Hong, Yongli Xiang, and Tongliang Liu. Toward robust non-transferable learning: A survey and benchmark. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 10455–10463, 2025. doi: 10.24963/ijcai.2025/1161.

[25] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *The 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.

[26] Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. $M^4I$: Multi-modal models membership inference. In *Advances in Neural Information Processing Systems*, volume 35, pages 1867–1882. Curran Associates, Inc., 2022. doi: 10.52202/068431-0136. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0c79d6ed1788653643a1ac67b6ea32a7-Paper-Conference.pdf.

[27] Jae-Yun Kim, Saerom Park, Joohee Lee, and Jung Hee Cheon. Privacy-preserving embedding via look-up table evaluation with fully homomorphic encryption. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 24437–24457, 2024.

[28] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[29] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[30] Joon-Woo Lee, Hyungchul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, and Jong-Seon No. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *IEEE Access*, 10:30039–30054, 2022. doi: 10.1109/ACCESS.2022.3159694.

[31] Qiushi Li, Yan Zhang, Ju Ren, Qi Li, and Yaoxue Zhang. You can use but cannot recognize: Preserving visual privacy in deep neural networks. In *The Network and Distributed System Security Symposium (NDSS)*, 2024.

[32] Shuofeng Liu, Zihan Wang, Minhui Xue, Long Wang, Yuanchao Zhang, and Guangdong Bai. Being transparent is merely the beginning: Enforcing purpose limitation with polynomial approximation. In *The 33rd USENIX Security Symposium (USENIX Security)*, pages 6507–6524, 2024. ISBN 978-1-939133-44-1.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[34] Yixin Liu, Kaidi Xu, Xun Chen, and Lichao Sun. Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 38(4):3783–3791, 2024. doi: 10.1609/aaai.v38i4.28169.

[35] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, pages 216–233, 2024.

[36] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.

[37] Zhongkui Ma, Jiaying Li, and Guangdong Bai. ReLU hull approximation. *Proceedings of the ACM on Programming Languages*, 8(POPL):2260–2287, 2024. doi: 10.1145/3632917.

[38] Souhail Meftah, Benjamin Hong Meng Tan, Chan Fook Mun, Khin Mi Mi Aung, Bharadwaj Veeravalli, and Vijay Chandrasekhar. DOReN: Toward efficient deep convolutional neural networks with fully homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 16:3740–3752, 2021. doi: 10.1109/TIFS.2021.3090959.

[39] noyb – European Center for Digital Rights. Criminal complaint against facial recognition company Clearview AI, October 2025. URL https://noyb.eu/en/criminal-complaint-against-facial-recognition-company-clearview-ai.

[40] Mauro Ribeiro, Katarina Grolinger, and Miriam AM Capretz. MLaaS: Machine learning as a service. In *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 896–902. IEEE, 2015.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[43] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. ISSN 1941-0026. doi: 10.1109/tevc.2019.2890858.

[44] The White House. America's AI Action Plan. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/, July 2025. Released July 23, 2025 as mandated by EO 14179.

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021.

[46] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[47] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

[48] Derui Wang, Minhui Xue, Bo Li, Seyit Camtepe, and Liming Zhu. Provably unlearnable data examples. In *The Network and Distributed System Security Symposium (NDSS)*, 2025.

[49] Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. In *International Conference on Learning Representations (ICLR)*, 2022.

[50] Zihan Wang, Zhongkui Ma, Xinguo Feng, Ruoxi Sun, Hu Wang, Minhui Xue, and Guangdong Bai. CoreLocker: Neuron-level usage control. In *IEEE Symposium on Security and Privacy (SP)*, pages 2497–2514, 2024. doi: 10.1109/SP54263.2024.00233.

[51] Nathan Wycoff, Mickaël Binois, and Stefan M. Wild. Sequential learning of active subspaces. *Journal of Computational and Graphical Statistics*, 30(4):1224–1237, 2021. doi: 10.1080/10618600.2021.1874962. URL https://doi.org/10.1080/10618600.2021.1874962.

[52] Jingwen Ye and Xinchao Wang. Ungeneralizable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11944–11953, 2024.

[53] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

[54] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*, pages 682–697, 2018. ISBN 978-3-030-01267-0.

# A Supplementary Preliminaries

## A.1 Eigendecomposition and PCA

Principal component analysis (PCA) projects data onto directions of maximal variance, obtained by eigendecomposing the data covariance. We recall the eigendecomposition of a real symmetric matrix.

**Definition 3** (Eigendecomposition)**.** *The eigendecomposition of a real symmetric matrix $C \in \mathbb{R}^{n\times n}$ is a factorization of the form $C = V\Lambda V^\top$, where $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns are the eigenvectors of $C$, and $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $C$. The eigenvalues are the scalars $\lambda_i$ such that $Cv_i = \lambda_i v_i$, where $v_i$ is the $i$-th eigenvector of $C$.*

By default, eigenvalues are arranged in ascending order along the diagonal of $\Lambda$. Geometrically, $V^\top$ rotates the coordinate system, $\Lambda$ scales the new axes by the eigenvalues, and $V$ rotates back. For a centered random vector $x \in \mathbb{R}^n$ with $\mathbb{E}[x] = 0$:

**Definition 4** (Principal Component Analysis (PCA))**.** *The covariance matrix is $C = \mathbb{E}[xx^\top] \in \mathbb{R}^{n\times n}$, where $\mathbb{E}$ denotes the expectation operator. The eigendecomposition of the covariance matrix is given by $C = V_{pca}\Lambda V_{pca}^\top$, where $V_{pca}$ is the matrix of eigenvectors (column vectors) and $\Lambda$ is the diagonal matrix of eigenvalues. The principal components are the columns of the matrix $V_{pca}$, and the projection of the data onto the principal components is given by $V_{pca}^\top x$.*

PCA commonly retains the top $k$ components for dimensionality reduction. The bottom components identify directions in which the input varies least.

### A.2 Singular Value Decomposition

We use the singular value decomposition (SVD) to analyze the spectral structure of weight matrices, following Section 2.1 and omitting biases for simplicity.

**Lemma 1** (Singular Value Decomposition (SVD))**.** *For any matrix $W \in \mathbb{R}^{m\times n}$, there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$, and a diagonal matrix $S \in \mathbb{R}^{m\times n}$ such that*

$$W = USV^\top.$$

*The diagonal entries of $S$ are the singular values of $W$, and the columns of $U$ and $V$ are the left and right singular vectors, respectively.*

Singular values are non-negative, and we index them in ascending order along $S$ throughout this appendix. Larger singular values correspond to higher-gain directions. Small singular values define the low-sensitivity directions used in our construction.

### A.3 Convolution

Ignoring the bias, convolution is a linear map expressible as matrix multiplication. We write the ungrouped convolution with unit stride and dilation, no padding, and no batch dimension.

**Definition 5** (Convolution)**.** *Let $X \in \mathbb{R}^{c_1\times h\times w}$, $K \in \mathbb{R}^{c_1\times c_2\times k_h\times k_w}$, and $b \in \mathbb{R}^{c_2}$ denote the input, kernel, and bias, with $c_1$ input and $c_2$ output channels. The output $Y \in \mathbb{R}^{c_2\times h'\times w'}$ has entries*

$$Y_{j,h'_i,w'_i} = \sum_{c=1}^{c_1}\sum_{u=1}^{k_h}\sum_{v=1}^{k_w} K_{c,j,u,v}\cdot X_{c,h'_i+u-1,w'_i+v-1} + b_j,$$

*where $j \in \{1,\dots,c_2\}$ indexes the output channel and $(h'_i, w'_i)$ is the spatial position in the output.*

Unfolding the input into patches reduces convolution to matrix multiplication.

**Lemma 2** (Convolution as Matrix Multiplication)**.** *Unfold $X$ into $X' \in \mathbb{R}^{c_1k_hk_w\times h'w'}$ and reshape $K$ into $K' \in \mathbb{R}^{c_2\times c_1k_hk_w}$. Compute $Y' = K'X' + b$, broadcasting $b$ over spatial positions, and reshape $Y' \in \mathbb{R}^{c_2\times h'w'}$ into $Y \in \mathbb{R}^{c_2\times h'\times w'}$, which equals the convolution output in Definition 5.*

*Proof.* Each local input patch becomes one column of $X'$, and each output-channel kernel becomes one row of $K'$. Multiplying $K'X'$ and adding the bias gives $Y'$, which is reshaped to $Y$. □

**Operator scope used in Section 4.** The spectral construction uses a per-architecture local linear map $W_{\text{loc}}$: the unfolded per-patch kernel $K' \in \mathbb{R}^{c_2\times c_1k_hk_w}$ of Lemma 2 for convolutional fronts ($\mathbb{R}^{64\times 147}$ for ResNet-50's $7\times 7$ stem), the patch-embedding projection for vision-transformer fronts, and the input weight matrix for fully connected fronts. Preprocessing (resize, clipping, channel normalization) is applied before recoding; $\delta$ is sampled in $W_{\text{loc}}$'s standardized coordinates and lifted to the native input via the architecture-specific fold/tile/embed map. In the network-wide bounds, $W$ denotes the corresponding full input-to-feature operator, written $\bar{W}$ below.

For non-overlapping patch embeddings (ViT, DeiT, SwinV2; stride equals patch size), tiling is exact: a single $\delta_{\text{patch}} \in \text{Ins}_\tau(W_{\text{loc}})$ is replicated identically across all non-overlapping patches of the lifted $\delta$, so Theorem 1 applies per token. With $P = h'w'$ and the per-patch parameters $k, \sigma, t$, the tiled instantiation of Eq. (6) is

$$\|f^\star(\tilde{x}) - f^\star(x)\|_2 \le L^\star_{\text{prod}}\sqrt{P}\,\tau\sqrt{k\sigma^2 + t},$$

with the same probability as Theorem 1. The logit-margin threshold in Section 4 is $\sqrt{2}$ times this full-tensor bound. When both models share the same aligned patch layout, both terms on the right-hand side of Eq. (7) acquire the factor $\sqrt{P}$ under this local parameterization. For overlapping convolutional stems (*e.g.,* ResNet-50's 7×7 stride-2 conv and MambaVision's 3×3 stride-2 conv), the full image-to-feature map is the Toeplitz extension $\bar{W} = (I_{h'w'} \otimes K')\, U$ (with $U$ the unfold); local $\tau$-insensitivity bounds the per-patch deviation but does not automatically place every overlapping patch of the lifted $\delta$ in $\mathrm{Ins}_\tau(W_{\mathrm{loc}})$. The retention bound thus lifts exactly for the three non-overlapping stems. For overlapping stems, the retention bound applies exactly to the local form used for authorized evaluation, and ResNet-50 layer-wise deviations stay at most 0.102 at 20 dB (Appendix B.2).

The cross-model $\alpha$ in Theorem 2 is measured *pairwise* on a shared local domain (or, when first-layer domains differ across architectures, after a common pixel-space lift used uniformly in Section 5).

### A.4 Token Embedding

Token embedding represents each vocabulary item by a learned vector [9]. With a one-hot input, it is a linear transformation.

**Definition 6** (Token Embedding)**.** *The token embedding is a linear transformation that maps a discrete token $t \in \{0,1\}^d$, encoded as a one-hot vector over a vocabulary of size $d$, to a continuous vector representation $e \in \mathbb{R}^m$ via a weight matrix $W \in \mathbb{R}^{m\times d}$, where $m$ is the embedding dimension. The token embedding is $e = Wt$, so that $e$ selects the column of $W$ indexed by the active entry of $t$.*

## B Supplementary Assumptions

We refer to the *spectral distribution* of a matrix as its eigenvalues or singular values, as appropriate.

**Standing constants.** Throughout this appendix, we use two layer-wise composition constants of the authorized and unauthorized networks, evaluated on the trajectory of interest:

$$L^\star_{\mathrm{prod}} := \prod_{i=2}^{L} L^\star_i, \qquad \nu^\star_{\mathrm{prod}} := \prod_{i=2}^{L} \nu'^\star_i,$$

the authorized model's Lipschitz product and the unauthorized model's co-Lipschitz product, respectively.

### B.1 Spectral Flatness of the First-Layer Operator

A spectrum is *flat* when a nontrivial fraction of the singular values lie below a small threshold.

**Assumption 1** (Spectral Flatness of the First-Layer Operator)**.** *Given a network trained on a specific task, there exist a threshold $\tau > 0$ and an index $k$ such that the bottom $k$ singular values of the first-layer weight matrix $W$*

$$s_1 \le s_2 \le \cdots \le s_k \le \tau$$

*form a nontrivial fraction of the spectrum: the spectral mass below $\tau$ is bounded below by a constant $\beta > 0$ uniformly across architectures.*

Figure 2 and Table 6 show low-energy tails in the first-layer spectra. The cumulative-energy curves are head-concentrated: ResNet-50 reaches 95.4% of total spectral energy with 24 components and 99.0% with 33, while ViT-B requires 91 and 107 components for the same levels despite a higher-dimensional patch projection. The corresponding spread is wide; minima are several orders of magnitude below maxima ($7.77 \times 10^{-8}$ vs. 3.8203 for ResNet-50; $3.68 \times 10^{-5}$ vs. 9.5686 for ViT-B), and medians sit well below means, consistent with Assumption 1.

### B.2 Layer-wise Deviation under NTEs

Table 7 reports ResNet-50 layer-wise deviation norms at seven checkpoints (L0–L6, one every 40 forward operations counting normalizations and activations) for NTE strengths from 40 to 0 dB PSNR. Deviations generally increase as PSNR decreases. At 0 dB, the largest measured deviation is 1.5062 at L5, while the deviation at L6 is 0.0674.

Table 6: Summary statistics of first-layer singular values.

| | Max | Min | Mean | Median |
|---|---|---|---|---|
| ResNet-50 | 3.8203 | $7.77\times10^{-8}$ | 1.0242 | 0.5904 |
| ViT-B | 9.5686 | $3.68\times10^{-5}$ | 0.5666 | 0.0482 |

Table 7: Layer-wise deviation norms for ResNet-50 under different NTE strengths. L0–L6 index seven checkpoints, one every 40 forward operations (counting normalizations and activations).

| PSNR (dB) | L0 | L1 | L2 | L3 | L4 | L5 | L6 |
|---|---|---|---|---|---|---|---|
| 40 | 0.0032 | 0.0097 | 0.0014 | 0.0062 | 0.0043 | 0.0210 | 0.0008 |
| 35 | 0.0035 | 0.0107 | 0.0017 | 0.0072 | 0.0050 | 0.0247 | 0.0010 |
| 30 | 0.0071 | 0.0200 | 0.0031 | 0.0146 | 0.0102 | 0.0498 | 0.0023 |
| 25 | 0.0134 | 0.0331 | 0.0052 | 0.0272 | 0.0184 | 0.0859 | 0.0040 |
| 20 | 0.0112 | 0.0367 | 0.0068 | 0.0282 | 0.0207 | 0.1016 | 0.0046 |
| 15 | 0.0240 | 0.0680 | 0.0104 | 0.0499 | 0.0349 | 0.1672 | 0.0079 |
| 10 | 0.0513 | 0.1731 | 0.0293 | 0.1108 | 0.0849 | 0.4344 | 0.0192 |
| 5 | 0.0754 | 0.3013 | 0.0501 | 0.1506 | 0.1239 | 0.6520 | 0.0264 |
| 0 | 0.1916 | 0.5643 | 0.0994 | 0.4031 | 0.2991 | 1.5062 | 0.0674 |

# C Supplementary Experiments and Discussion

All experiments use Python 3.12.3 and PyTorch 2.3.0 (CUDA 12.3) on a workstation with an AMD Ryzen Threadripper PRO 5965WX (24 cores), 256 GB RAM, and two NVIDIA RTX A6000 GPUs.

## C.1 Weight-level Specificity Matrix

Table 8 evaluates models with weights different from the authorized model: diagonal entries share its architecture, and shaded off-diagonals use different architectures. Diagonal accuracies are 8.2–21.0% on CIFAR-10 and 0.0–1.2% on ImageNet. Off-diagonal accuracies are also low.

Table 8: Model-specific non-transferability. All entries evaluate NTEs on weights different from the authorized model: diagonals share architecture, shaded off-diagonals are cross-architecture. Vertical labels give the training set of the authorized (row) model.

| | | CIFAR-10 | | | | | | ImageNet | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ResNet-50 | ViT-B | SwinV2-T | DeiT-B | MambaVision-T | | ResNet-50 | ViT-B | SwinV2-T | DeiT-B | MambaVision-T |
| ResNet-50 | ImageNet | 13.3% | 9.4% | 8.2% | 9.4% | 7.8% | CIFAR-10 | 1.2% | 0.0% | 4.4% | 0.0% | 0.0% |
| ViT-B | ImageNet | 10.1% | 9.6% | 10.1% | 11.5% | 9.8% | CIFAR-10 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| SwinV2-T | ImageNet | 10.5% | 10.2% | 21.0% | 11.3% | 14.5% | CIFAR-10 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| DeiT-B | ImageNet | 12.5% | 9.3% | 7.0% | 14.8% | 10.9% | CIFAR-10 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| MambaVision-T | ImageNet | 14.5% | 5.1% | 12.5% | 9.8% | 8.2% | CIFAR-10 | 0.0% | 0.0% | 1.9% | 0.0% | 0.0% |

**Experiments on Language Models**. Table 9 reports GLUE [47] results for BERT-base and RoBERTa-base [33]. Authorized scores remain within 2 points of their clean baselines. Unauthorized scores fall on every task, with RoBERTa retaining a score of 63.5 on STS-B.

Table 9: GLUE performance of authorized and unauthorized language models.

| | BERT-base | | | | | | | RoBERTa-base | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CoLA | MNLI | QNLI | QQP | RTE | SST-2 | STS-B | CoLA | MNLI | QNLI | QQP | RTE | SST-2 | STS-B |
| Baseline | 54.2 | 83.4 | 90.5 | 90.1 | 60.3 | 91.6 | 87.1 | 53.8 | 87.7 | 92.8 | 90.9 | 66.1 | 94.5 | 87.5 |
| NTE (Ours) | 54.5 | 82.9 | 89.5 | 89.6 | 60.3 | 89.6 | 86.9 | 55.5 | 87.5 | 92.4 | 90.8 | 65.3 | 94.6 | 87.2 |
| Unauthorized | 33.4 | 31.8 | 50.5 | 36.8 | 47.3 | 50.9 | 60.2 | 37.1 | 35.4 | 49.5 | 63.2 | 52.7 | 49.1 | 63.5 |

## C.2 Limitations

Our setting grants the defender white-box access to $f^\star$ to estimate an insensitive subspace used for recoding. An attacker that estimates the recoding subspace can use an input-side projection to reduce the perturbation without direct access to $f^\star$. This may restore some unauthorized utility, but does not guarantee full recovery across models.

An attacker with training-time control may also recover utility by regularizing its model to be insensitive along the estimated subspace. Other practical limits include attenuation under aggressive acquisition pipelines or domain shift that changes early-layer geometry. Our analysis starts from the first-layer operator; extensions to deeper Jacobians and temporally coupled modalities remain open.

### C.3 Runtime Breakdown

Table 5 reports per-backbone wall-clock measurements (10 runs). The one-time SVD takes 0.2 ms (MambaVision-T) to 148.3 ms (DeiT-B) and is reused across releases. Per-image NTE encoding takes about 0.11 ms for a $224 \times 224 \times 3$ input across all backbones.

**Computational cost (ResNet-50).** For the ResNet-50 stem, the local kernel matrix is $W_{\text{loc}} \in \mathbb{R}^{64\times 147}$. Computing a full right singular basis costs $\mathcal{O}(64 \cdot 147^2)$ operations. For a $224 \times 224$ RGB image, the projection and input addition cost $\mathcal{O}(nk + 3 \cdot 224^2)$ operations, where $n = 147$ and $k$ is the number of selected directions. For ViT-B and DeiT-B, $n = 768$; measured encoding time remains below one millisecond per image (Table 5).

### C.4 Statistical Uncertainty

For the fixed evaluation settings in Table 1, Hoeffding's inequality [46] and a union bound give simultaneous 95% lower confidence bounds of 64.0 percentage points on CIFAR-10 ($n = 10{,}000$) and 57.0 on ImageNet ($n = 50{,}000$) for all 40 authorized–unauthorized accuracy gaps. These bounds account for independent per-image recoding and rounding of the reported accuracies.

## D Deferred Proofs from Section 4

This section gives the full proofs of Theorems 1 and 2 and Proposition 1, under the standing assumptions in Section 4.

### D.1 Full Proof of Theorem 1: Chebyshev Tail on $\|z\|_2^2$

The deterministic step in the main text gives $\|W\delta\|_2 \le \tau \|z\|_2$. We complete the proof by bounding $\|z\|_2$ in probability. For $z \sim \mathcal{N}(0, \sigma^2 I_k)$, $\mathbb{E}[\|z\|_2^2] = k\sigma^2$ and $\mathrm{Var}(\|z\|_2^2) = 2k\sigma^4$, so Chebyshev's inequality gives $\mathbb{P}[\|z\|_2^2 \ge k\sigma^2 + t] \le 2k\sigma^4/t^2$. Therefore, with probability at least $1 - 2k\sigma^4/t^2$,

$$\|W\tilde{x} - Wx\|_2 = \|W\delta\|_2 \le \tau \|z\|_2 < \tau \sqrt{k\sigma^2 + t},$$

which concludes the proof.

### D.2 Full Proof of Theorem 2: Gaussian Anti-Concentration on $p_1^\top z$

The main-text proof establishes

$$\|W'\delta\|_2 \ge s'_{\min} \|(V'^{(l)})^\top V^{(s)} z\|_2.$$

We complete the proof by lower-bounding $\|(V'^{(l)})^\top V^{(s)} z\|_2$ in probability.

Let $A := (V'^{(l)})^\top V^{(s)} \in \mathbb{R}^{(n-k)\times k}$, and let $u_1, p_1$ be the top left/right singular vectors of $A$, so that $u_1^\top A = \|A\|_2 p_1^\top$. Then

$$u_1^\top Az = \|A\|_2 p_1^\top z, \qquad p_1^\top z \sim \mathcal{N}(0, \sigma^2),$$

since $\|p_1\|_2 = 1$ and $z \sim \mathcal{N}(0, \sigma^2 I_k)$. The density of $\mathcal{N}(0, \sigma^2)$ is bounded above by $1/(\sigma\sqrt{2\pi})$, so for any $c \in (0, \sqrt{\pi/2}]$,

$$\Pr\big[|p_1^\top z| < c\sigma\big] \;\le\; 2c\sigma \cdot \frac{1}{\sigma\sqrt{2\pi}} \;=\; c\sqrt{\tfrac{2}{\pi}},$$

the standard Gaussian anti-concentration bound. Hence with probability at least $1 - c\sqrt{2/\pi}$,

$$\|Az\|_2 \;\ge\; |u_1^\top Az| \;=\; \|A\|_2 |p_1^\top z| \;\ge\; c\sigma \|A\|_2,$$

where the first inequality is obtained by projecting onto the unit vector $u_1$. By the misalignment condition (standing assumption), $\|A\|_2 \geq \alpha$, so $\|Az\|_2 \geq c\alpha\sigma$, and therefore

$$\|W'\delta\|_2 \geq s'_{\min}\, c\alpha\sigma.$$

Taking the difference with the $W$ bound from Theorem 1 (which holds with probability at least $1 - 2k\sigma^4/t^2$) and union-bounding the two failure events yields Eq. (5).

### D.3 Full Proof of Proposition 1(a): Layer-Wise Induction

Following the standing decomposition $f^\star = g_L \circ \cdots \circ g_2 \circ g_1$ with $g_1 = W$ the first linear operator, let $h_i = g_i \circ \cdots \circ g_1(x)$ and $\tilde{h}_i = g_i \circ \cdots \circ g_1(\tilde{x})$. By induction on $i$, we show $\|\tilde{h}_i - h_i\|_2 \leq (\prod_{j=2}^{i} L_j^\star) \cdot \tau\sqrt{k\sigma^2 + t}$.

For the base case $i = 1$, Theorem 1 gives

$$\|\tilde{h}_1 - h_1\|_2 = \|W^\star\delta\|_2 < \tau\sqrt{k\sigma^2 + t}.$$

Now assume the claim holds for $i - 1$. Since $g_i$ is $L_i^\star$-Lipschitz on the trajectory pair,

$$\|\tilde{h}_i - h_i\|_2 \leq L_i^\star \|\tilde{h}_{i-1} - h_{i-1}\|_2 \leq \left(\prod_{j=2}^{i} L_j^\star\right) \tau\sqrt{k\sigma^2 + t}.$$

At layer $L$, $\|\tilde{h}_L - h_L\|_2 = \|f^\star(\tilde{x}) - f^\star(x)\|_2$, so by induction

$$\|f^\star(\tilde{x}) - f^\star(x)\|_2 \;\leq\; L^\star_{\text{prod}}\, \tau\sqrt{k\sigma^2 + t}, \tag{6}$$

which is Proposition 1(a).

### D.4 Full Proof of Proposition 1(b): Layer-Wise Induction and Joint-Failure Union Bound

By Theorem 2, $\|W'\delta\|_2 \geq c\alpha\, s'_{\min}\sigma$ with probability at least $1 - c\sqrt{2/\pi}$. Under the standing co-Lipschitz assumption (each $g'_i$ is $\nu_i'^\star$-co-Lipschitz on the trajectory pair), composition through $g'_2, \ldots, g'_L$ with constant $\nu^\star_{\text{prod}}$ propagates the lower bound, yielding

$$\|f'(\tilde{x}) - f'(x)\|_2 \;\geq\; \nu^\star_{\text{prod}}\, \|W'\delta\|_2 \;\geq\; \nu^\star_{\text{prod}}\, c\alpha\, s'_{\min}\sigma.$$

Proposition 1(a) gives $\|f^\star(\tilde{x}) - f^\star(x)\|_2 \leq L^\star_{\text{prod}}\tau\sqrt{k\sigma^2 + t}$ with probability at least $1 - 2k\sigma^4/t^2$. Both events depend on the same draw $z$. A union bound gives a probability of at most $2k\sigma^4/t^2 + c\sqrt{2/\pi}$ that either event fails. Taking the difference on the joint success event yields

$$\|f'(\tilde{x}) - f'(x)\|_2 - \|f^\star(\tilde{x}) - f^\star(x)\|_2 \;\geq\; \nu^\star_{\text{prod}}\, c\alpha\, s'_{\min}\, \sigma - L^\star_{\text{prod}}\, \tau\sqrt{k\sigma^2 + t}. \tag{7}$$

## E Threat Model and Deployment

### E.1 Multi-Client Deployment

A data owner serving multiple authorized client models can construct a separate transformation $\mathcal{T}_{f_i}$ for each model $f_i$ and release $\tilde{x}_i = \mathcal{T}_{f_i}(x)$ to that client. The owner controls future issuance separately for each client. A shared transformation would require directions that remain insensitive for every authorized model. As the group grows, this requirement can shrink the available subspace and change the choice of $\tau$ and $\sigma$. Characterizing the resulting retention/degradation trade-off and constructing shared transformations are left to future work.

### E.2 Justification of White-Box Access

Three regulatory frameworks motivate the "delegated auditing" paradigm for white-box access. The EU AI Act (Annex VII, §4.5) grants notified bodies access to models and relevant parameters where necessary for high-risk conformity assessment, upon a reasoned request once all other reasonable means prove insufficient [13]. Under supervisory guidance (SR 26-2), banks were expected to validate their own use of vendor products [4]. GDPR Art. 15(1)(h) grants data subjects access to "meaningful information about the logic" of automated decision-making [14].

### E.3 On Visual Interpretability of $\tilde{x}$

Our design deliberately makes $\tilde{x}$ uninformative to humans. A human-readable $\tilde{x}$ would carry task-relevant information in a generic, model-agnostic form usable by arbitrary downstream models; NTEs instead retain utility through a channel tied to $f^\star$. In this sense, $\tilde{x}$ is a task-level "ciphertext" that only the authorized model can reliably "decode". The construction targets automated model inference. We define PSNR settings in the model's input space (the normalized inputs to the first-layer operator), using a peak value of 255 for all backbones. Because they differ in subspace dimension and input scale, we calibrate each sample's amplitude against a reference energy. The setting controls perturbation strength, and Figure 4 shows how stronger recoding obscures a sample image. At 20 dB, authorized accuracy is largely preserved by confining the perturbation to the insensitive subspace. When human review is required, the data owner can retain the clean input $x$ within the trusted boundary and release $\tilde{x}$ for downstream model inference.

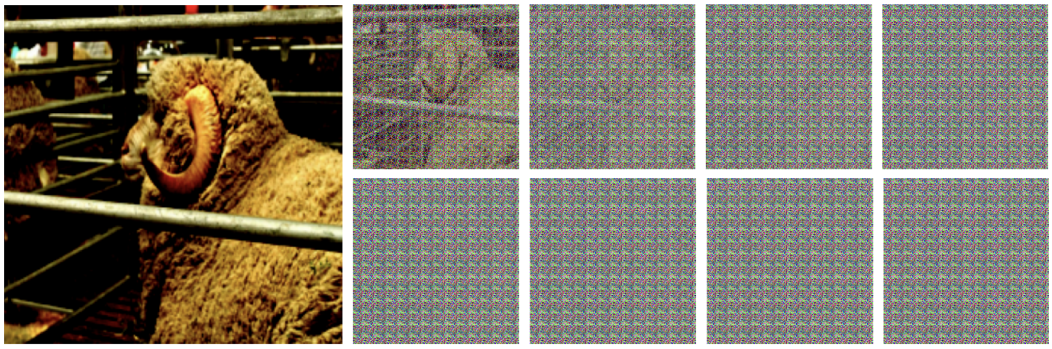

Figure 4: Visual examples under increasing perturbation strength. At 0 dB, authorized ResNet-50 top-1 accuracy on ImageNet decreases by 0.1 percentage points.

### E.4 Vision-Language Evaluation

InternVL3 and Qwen2.5-VL use model-specific image preprocessing. On MMBench, authorized InternVL3-1B accuracy changes from 72.7% on clean inputs to 72.6% on NTEs, while unauthorized Qwen2.5-VL-3B accuracy falls from 78.8% to 18.3%. The same separation appears across the six capability dimensions (AR, CP, FP-C, FP-S, LR, RR) in Table 3; Figure 5 shows an example.

PROMPT *Please describe the image in detail.*

| |  (a) NTE INPUT |  (b) CLEAN INPUT |
|---|---|---|
| **InternVL3** AUTHORIZED | In the image, a person is sitting on a sandy beach, smiling and playing with a dog. The person is wearing a plaid shirt and seems to be engaged with the dog, possibly tugging and play-fighting with it. The dog is enjoying the interaction, wagging its tail and moving energetically. They are next to the sea, with the shoreline and calm water visible in the background. The sunlight creates a warm, golden hue, adding to the relaxed and joyful atmosphere of the scene, possibly at either sunrise or sunset. | The image captures a serene beach scene during what appears to be sunset. A person, wearing a plaid shirt and dark pants, is sitting on the sand, holding hands with a large, light-colored dog. The dog is attentively playing with the person, reaching out its paw with joy and companionship... |
| **Qwen2.5-VL** UNAUTHORIZED | The image appears to be a pixelated or low-resolution representation of a square object. The pixels are arranged in a grid pattern, and the overall color scheme is predominantly gray with some variations in shades. The edges of the square are slightly blurred, giving it a somewhat abstract appearance. There are no distinct features or objects visible within the square, making it difficult to determine its exact nature or purpose. | The image depicts a serene beach scene with a person and a dog. The person is sitting on the sandy beach, facing the ocean. They are wearing a plaid shirt and dark pants, and they have long hair. The dog, which appears to be a Labrador Retriever, is sitting on the sand and is interacting ... |

Figure 5: (a) Responses to a recoded beach image. InternVL3 describes the beach scene, while Qwen2.5-VL describes the image as a pixelated pattern. (b) Responses to the corresponding clean beach image. Both models describe the beach scene.

PROMPT *Please describe this image in a single concise paragraph.*

| |  (a) NTE INPUT |  (b) CLEAN INPUT |
|---|---|---|
| **InternVL3** AUTHORIZED | This image shows a super busy city street packed with traffic and people, with lots of yellow taxis in the foreground along with other cars all stuck in congestion. The street is lined with tall buildings on both sides, creating a canyon effect, and you can see many traffic lights, street signs, and American flags hanging from poles. There are pedestrians on the sidewalks, especially on the left side of the image, and farther down the street the traffic continues into the distance, getting hazier because of the depth and maybe a bit of atmospheric blur. Overall, it looks like a classic dense downtown scene in a big U.S. city, with heavy traffic, high-rises, and a lot of visual clutter from signs, ads, and lights. | The image depicts a bustling urban street scene, likely in a major city known for its iconic yellow taxis. The street is filled with numerous vehicles, including cars and taxis, indicating heavy traffic. The buildings lining the street are tall and modern, suggesting a downtown area. Street signs and traffic lights are visible, directing the flow of traffic. The sky appears overcast, adding to the busy atmosphere. The overall scene conveys a sense of congestion and the dynamic nature of city life. |
| **GPT-5** UNAUTHORIZED | The image is a densely packed, full-frame pattern made up of extremely small, repeating colorful shapes that create a fine, almost pixel-like texture; from a distance it appears as multicolored visual static with a subtle grid structure, dominated by greens, purples, yellows, and blues distributed uniformly across the entire square. | The image shows a busy city avenue lined with tall buildings, dense traffic, and numerous yellow taxis, with traffic lights, street signs, flags, and storefronts creating a lively, crowded urban scene that fades into a hazy skyline in the distance. |
| **Kimi k1.5** | The image displays a patterned texture resembling a fabric or material with a consistent, tightly woven design. The pattern is composed of small, interlocking shapes that create a complex, geometric lattice. The color palette is muted, featuring a blend of grays, greens, and subtle hints of other colors, giving it an almost camouflage-like appearance. The texture appears flat and uniform throughout, suggesting a synthetic or manufactured origin rather than a natural one. | The image captures a busy New York City street filled with yellow taxis and various vehicles, surrounded by tall buildings adorned with signs and American flags. Traffic lights and street signs, including directions and regulations, are visible, contributing to the organized chaos typical of a bustling urban environment. The overcast sky softens the scene, highlighting the city's dynamic nature. |
| **Sonnet 4.5** | This image appears to be random digital noise or static, displaying a uniform pattern of multicolored pixels (red, green, blue, and gray) distributed across the entire frame with no discernible subject, pattern, or structure. It resembles the visual "snow" or grain you might see on an old television with no signal, or could be the result of a corrupted image file, sensor noise from a camera, or an intentionally generated random noise pattern. | This is a vibrant street photograph of a busy Manhattan avenue in New York City, showcasing the iconic urban canyon effect with tall buildings lining both sides of the street. The scene is packed with yellow taxi cabs and other vehicles in dense gridlock, with multiple traffic lights hanging overhead on wires. The atmospheric haze in the distance adds depth to the composition, while pedestrians and commercial signage contribute to the bustling energy. The photo perfectly captures that quintessential NYC feeling - crowded, vertical, and pulsing with constant movement and organized chaos. |
| **Grok 4** | This image is a stereogram (autostereogram), a 2D pattern of repeating green, purple, brown, and yellow pixel-like dots designed to create a 3D optical illusion when viewed by crossing or diverging the eyes, revealing a hidden image embedded within the structured, tiled texture. | This image captures a bustling urban street scene in a major city, likely New York City, characterized by heavy traffic and iconic yellow taxis. The street is flanked by tall, modern buildings, some of which feature prominent signage and advertisements. Pedestrians can be seen on the sidewalks, and the overall atmosphere is one of a dense, busy metropolitan environment. The lighting suggests it might be daytime, but with an overcast or slightly hazy sky. |

Figure 6: (a) Responses to a recoded street image. InternVL3 describes the street scene, while the other models shown describe patterns or noise. (b) Responses to the corresponding clean street image. All models shown describe the street scene.

### E.5 Reconstruction Attack

Table 10 reports SR-ResNet reconstruction of NTEs at an input PSNR of 10.2–10.7 dB. With Noise2Noise or Noise2Clean training, PSNR increases by only 0.1–0.7 dB. For Gaussian noise, PSNR rises from 15.8 dB to 33.5–35.9 dB.

Table 10: SR-ResNet reconstruction PSNR (dB).

| Method | Input | SR-ResNet[1] | SR-ResNet[2] |
|---|---|---|---|
| Gaussian | 15.8 | 33.5 | 35.9 |
| NTEs | 10.2 | 10.5 | 10.9 |
| NTEs[3] | 10.7 | 10.8 | 11.1 |

[1] Noise2Noise: train on recoded → recoded pairs.
[2] Noise2Clean: train on recoded → clean pairs.
[3] The attacker has white-box access to the model.

### E.6 Learned Adaptation Attack

Table 11 reports CIFAR-10 accuracy of the NTE-Attacker from Table 4 across training lengths and released-data fractions. In this evaluation, the authorized ViT-B achieves 96.2% accuracy, while attacker accuracy ranges from 10.1% to 11.1%, close to the 10% chance level. All attacker parameters are fine-tuned in each run.

Table 11: Top-1 accuracy of the NTE-Attacker on CIFAR-10.

| Training length (ViT-B, full data) | 1 epoch | 3 epochs | 10 epochs | 30 epochs |
|---|---|---|---|---|
| Accuracy | 11.1% | 10.7% | 10.1% | 10.7% |
| Released data (ViT-B, 30 epochs) | 10% | 25% | 50% | 100% |
| Accuracy | 10.1% | 10.7% | 10.3% | 10.7% |

With full data and 30 training epochs, ViT-B and ViT-L attackers achieve 10.7% and 10.3% accuracy, respectively.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state our contributions plainly: a model-specific data-authorization setting, a training-free recoding within the insensitive subspace, two-sided probabilistic guarantees, and an empirical study across seven backbones on ImageNet, CIFAR-10, and MMBench. Each claim is substantiated by the corresponding result in Sections 3, 4, and 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix C.2 states the scope of our guarantees, including white-box access to $f^\star$, residual exposure to method-aware projection and counter-adaptive training, and the early-layer focus of the analysis. The empirical study in Section 5 reports authorization behavior across seven architectures, multiple datasets, and several preprocessing, reconstruction, and adaptive attacks, covering settings outside the strict assumptions of the theory.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 4 states the two main results, Theorems 1 and 2, together with Proposition 1 for network-wide propagation. The standing assumptions are stated in Section 4 and Appendix B, and full proofs are given in Appendix D (Appendices D.1, D.2, D.3, and D.4).

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 specifies the backbones, datasets, perturbation parameterization (PSNR), recoding noise, and the threshold $\tau$. All datasets and checkpoints are public, and Section 8.2 describes the datasets, models, and code availability. Code is available at `https://github.com/model-specific/non-transferable-examples`.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/model-specific/non-transferable-examples`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 reports the datasets (CIFAR-10, ImageNet-1K, MMBench), the evaluated checkpoints, the perturbation strength controlled by PSNR (20 dB by default), the recoding noise distribution, and the threshold $\tau$ for the insensitive subspace; CIFAR-10 results follow 10 epochs of fine-tuning from ImageNet-1K checkpoints. Additional setup, including weight-specificity matrices and runtime breakdowns, is given in Appendices C.1 and C.3.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix C.4 reports simultaneous 95% confidence bounds for the authorized–unauthorized accuracy gaps in Table 1, accounting for independent per-image recoding under the fixed evaluation settings.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C reports the hardware (AMD Ryzen Threadripper PRO 5965WX with 256 GB RAM and two NVIDIA RTX A6000 GPUs) and software stack (Python 3.12.3, PyTorch 2.3.0, CUDA 12.3). Wall-clock measurements per backbone are given in Appendix C.3, with the one-time SVD ranging from 0.2 ms to 148.3 ms and per-image NTE encoding around 0.11 ms.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses only public benchmark datasets and standard pretrained checkpoints, involves no human subjects, and collects no personally identifiable information. Ethical considerations, stakeholders, intended benefits, and potential misuse are discussed in Section 8.1, and we have reviewed the research against the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 8.1 discusses both directions: NTEs give data owners a deployable mechanism for purpose-limited release that supports compliance with frameworks such as the EU AI Act, while the same mechanism, if applied without disclosure or oversight, could be used to enforce restrictive access policies or to impede legitimate interoperability and auditing. We frame NTE as a governance-supporting control to be applied with clear authorization and disclosure.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release code for input recoding; no new datasets or pretrained models are released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (CIFAR-10, ImageNet-1K, MMBench, GLUE) and pretrained checkpoints (ResNet-50, ViT-B, SwinV2-T, DeiT-B, MambaVision-T, Qwen2.5-VL-3B, InternVL3-1B, BERT, RoBERTa) are obtained from their official sources under their standard licenses, with original papers cited at first use. We do not redistribute these assets; dataset and checkpoint sources are summarized in Section 8.2.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The NTE code is publicly available at https://github.com/model-specific/non-transferable-examples. No new datasets or pretrained models are released.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work involves no crowdsourcing and no research with human subjects; all experiments use public benchmark datasets and standard pretrained checkpoints.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects, so IRB approval (or equivalent) is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the core method. Language and vision-language models serve as evaluation targets.