The Platonic Defense: Backdoor Defense for Self-Supervised Encoders in the Era of Large Scale Pre-training

Tuo Chen[1,2] **Minjing Dong**[3] **Benlei Cui**[4] **Jie Gui**[1,5*] **Jian Liu**[2*]
[1]Southeast University [2]Ant Group
[3]City University of Hong Kong [4]Alibaba Group
[5]Purple Mountain Laboratories
tchen@seu.edu.cn, minjdong@cityu.edu.hk
cubenlei.cbl@alibaba-inc.com
guijie@seu.edu.cn, rex.lj@antgroup.com

Code: https://github.com/jsrdcht/Platonic-Representation-Defense

Abstract

Self-supervised learning (SSL) has become a dominant paradigm for visual representation learning, but SSL encoders are vulnerable to backdoor attacks. Existing defenses struggle against such attacks in a fully black-box setting because they often require access to labels, attack patterns, or training data. To tackle this issue, we propose a new attack-agnostic, model-agnostic, and modality-agnostic black-box test-time defense paradigm, called *Platonic Representation Defense*. It is inspired by the Platonic Representation Hypothesis, which suggests that large-scale independently trained encoders converge toward compatible projections of the same underlying reality. We formalize this idea as a conditional energy function defined over source representations and a set of reference representations. The energy function is trained for backdoor detection through noise-contrastive estimation and for representation purification through denoising score matching. Theoretically, the energy gap between matched and mismatched samples is lower bounded by the mutual information between source and reference representations. We demonstrate the effectiveness of our method on multiple self-supervised encoders and more than 10 attacks. The method performs both representation detection and purification, achieving substantial gains across attacks.

1 Introduction

Artificial intelligence systems are rapidly moving toward the era of large-scale pretraining. Earlier computer vision systems often rely on task-specific solutions for image classification [20], semantic segmentation [72] and other tasks. Modern large models can handle all these tasks with a single set of weights [52, 49], which are built on general SSL encoders that inherit strong representation ability and basic understanding of the world from pretraining [39, 49]. Despite great success, SSL encoders are vulnerable to backdoor attacks [35, 6, 56, 29, 8], which severely undermine their reliability and safety on downstream tasks. An attacker can manipulate model representations by injecting malicious samples into unlabeled training data or by modifying model weights [6, 29]. The compromised encoder behaves normally on clean inputs but exhibits malicious behavior on inputs that contain a trigger pattern. Since such encoders often serve as the backbone for a wide range of downstream

*Corresponding authors.

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

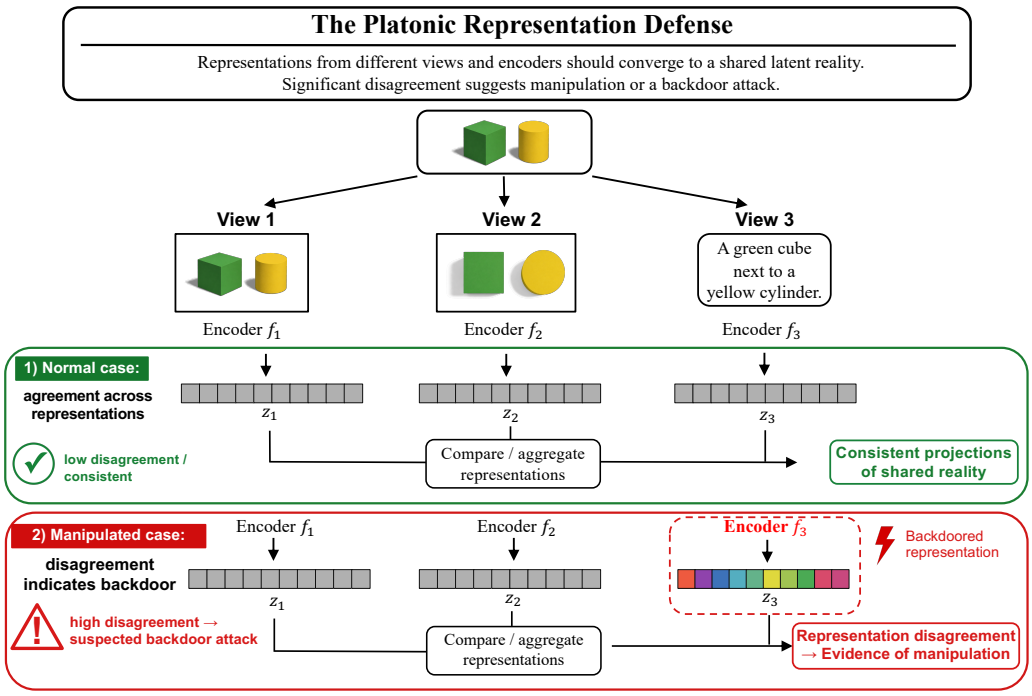

Figure 1: **The Platonic Representation Defense:** Representations (z_1, z_2, z_3) are projections of a shared underlying reality. We conjecture that backdoor attacks will deviate from this pattern.

tasks, developing effective backdoor defenses is essential for artificial intelligence systems in the era of large-scale pretraining [39].

Although many defense strategies have been proposed to mitigate backdoor attacks on SSL encoders, only a few operate in a fully black-box setting. Some studies construct pseudo-labels and reformulate the problem under the framework of supervised backdoor defense [62, 26], which only applies to a closed and finite label space. Other methods restrict the defense scope to specific types of backdoor triggers [22, 62] or multimodal backdoor attacks [3]. Their defense capability depends on access to the model's internal structure or the upstream training process. Such knowledge is difficult to obtain in advance in SSL settings, since SSL models serve a broad range of downstream users who typically cannot access the upstream pretraining data or training procedure.

To make defenses attack-agnostic, model-agnostic, and modality-agnostic, we propose a new framework that we call the *Platonic Representation Defense*. The name is borrowed from the Platonic Representation Hypothesis [27], which hypothesizes that representations produced by different encoders are projections of one common underlying reality and that large-scale pre-training drives these projections toward a shared limit. The compatibility between representations can thus be a good indicator of backdoor attacks if the hypothesis is true. We illustrate this intuition in Figure 1. We formalize this intuition as a conditional energy function $E_\theta\left(z^{\mathsf{s}},\, z^{\mathsf{r}_{1:K}}\right)$, where z^{s} is the output of a source encoder under inspection and $z^{\mathsf{r}_{1:K}}$ are the outputs of K trusted reference encoders on the same input. The energy is low on clean (compatible) samples and rises sharply on backdoored (incompatible) ones, recasting backdoor defense as a problem solved entirely in representation space, without any downstream labels and without access to poisoned samples. We train this energy with Noise Contrastive Estimation (NCE), which sidesteps the intractable partition function while delivering a calibrated detection score. To additionally support purification, we train a denoiser of the source representation conditioned on the reference representations via denoising score matching, by which Tweedie's identity recovers the score field of the same energy. Both objectives share a single architecture and only diverge at the output head and the loss. Theoretically, we show that when E_θ is close to the Bayes optimal, the energy gap between the matched (clean) and the mismatched (backdoor) distributions is lower bounded by the mutual information $I\left(Z^{\mathsf{s}};\, Z^{\mathsf{r}_{1:K}}\right)$ between the source and reference representations. Under the Platonic Representation Hypothesis, independently trained

encoders retain shared information, which gives clean representations a systematically lower energy than mismatched ones.

Our main contributions can be summarized as follows. **(1)** To our knowledge, this is the first work to connect representation convergence with backdoor defense, identifying the convergence of independently trained encoders as a previously unused security signal that is intrinsic to large-scale pre-training. **(2)** We introduce the platonic representation defense, a black-box test-time framework that formalizes multi-encoder compatibility as a conditional energy model and realizes it through two complementary training routes, discriminative NCE for detection and conditional denoising score matching for purification. **(3)** We provide both theoretical and empirical evidence. We validate our method on diverse SSL encoders against more than 10 attacks. Our framework preserves clean sample performance well, which makes it better suited to black-box settings.

2 Preliminaries

We focus on visual representations and leave other modalities to future work. A visual representation is a function $f : \mathcal{X} \to \mathcal{Z}$ that maps an image x to a latent vector $f(x)$; we characterize it by its kernel, i.e., the pairwise similarity structure $\langle f(x_i), f(x_j)\rangle$ it induces. Comparing representations via kernels is standard practice [27, 67, 33], since it does not depend on the dimensionality or coordinate system of the feature space. We mainly quantify the alignment between two representations with Centered Kernel Nearest-Neighbor Alignment (CKNNA) [27].

Representation convergence. A growing body of work shows that, as model capacity and training data scale up, independently trained models converge toward increasingly similar kernels. The *Plato Representation Hypothesis* [27, 32] formalizes this observation: the kernel-alignment between two models grows monotonically with their capability, irrespective of architecture, training objective, or even input modality, suggesting that sufficiently capable models recover a shared statistical structure underlying visual data. Earlier studies of cross-architecture and layer-wise similarity [33, 53, 47] report the same trend at smaller scales, and recent evidence shows that this convergence is semantic and extends across languages, modalities, and architectures [1], with strong visual encoders even beginning to mirror low-level characteristics of human vision [5].

Energy-based models and EDM parameterization. An energy-based model (EBM) [34] parameterizes a probability density through a scalar energy function $E_\theta : \mathcal{Z} \to \mathbb{R}$ as

$$p_\theta(z) \;=\; \frac{1}{Z(\theta)} \exp\big(-E_\theta(z)\big), \qquad Z(\theta) \;=\; \int \exp\big(-E_\theta(z)\big)\, \mathrm{d}z, \tag{1}$$

trading the normalization constraint of standard likelihood models for a free-form energy landscape. The partition function $Z(\theta)$ is intractable, so we bypass it with noise-contrastive estimation (NCE) [17] and score matching [28] which fits score $-\nabla_z E_\theta(z)$ without evaluating $Z(\theta)$. We adopt the practical EDM [31] variant of score matching used by modern diffusion models [59, 42]. For a clean data point $z \sim p_{\mathrm{data}}$, a Gaussian perturbation $\epsilon \sim \mathcal{N}(0, I)$, and a noise level $\sigma > 0$, define

$$\tilde{z} \;=\; z + \sigma\, \epsilon, \qquad p_\sigma(\tilde{z}) \;=\; \int \mathcal{N}(\tilde{z}; z, \sigma^2 I)\, p_{\mathrm{data}}(z)\, \mathrm{d}z. \tag{2}$$

Rather than regressing on $\nabla_{\tilde{z}} \log p_\sigma$ directly, score matching learns a *denoiser* $\hat{D}_\theta(\tilde{z}, \sigma) \in \mathcal{Z}$ that predicts the clean z from $\tilde{z}$, trained by the EDM-weighted regression

$$\mathcal{L}_{\mathrm{DSM}}(\theta) \;=\; \mathbb{E}_{\sigma,\, z,\, \epsilon}\Big[\, \lambda(\sigma)\, \big\|\hat{D}_\theta\big(z + \sigma\epsilon,\, \sigma\big) - z\big\|_2^2 \,\Big]. \tag{3}$$

Its optimum is the Bayes-optimal denoiser $\hat{D}^\star(\tilde{z}, \sigma) = \mathbb{E}[z \mid \tilde{z}, \sigma]$, and Tweedie's identity ties it back to the score:

$$\nabla_{\tilde{z}} \log p_\sigma(\tilde{z}) \;=\; \big(\hat{D}^\star(\tilde{z}, \sigma) - \tilde{z}\big)/\sigma^2, \tag{4}$$

so any denoiser trained with Eq. (3) directly induces a score estimator $s_\theta(\tilde{z}, \sigma) = (\hat{D}_\theta(\tilde{z}, \sigma) - \tilde{z})/\sigma^2$ that recovers $-\nabla_{\tilde{z}} E_\theta$ on the Gaussian-smoothed EBM. Following the EDM parameterization [31], we do not have the network predict z from $\tilde{z}$ directly; instead we expose a residual subnetwork F_θ and apply fixed σ-dependent preconditioning,

$$\hat{D}_\theta(\tilde{z}, \sigma) \;=\; c_{\mathrm{skip}}(\sigma)\, \tilde{z} \;+\; c_{\mathrm{out}}(\sigma)\, F_\theta\big(c_{\mathrm{in}}(\sigma)\, \tilde{z},\; c_{\mathrm{noise}}(\sigma)\big), \tag{5}$$

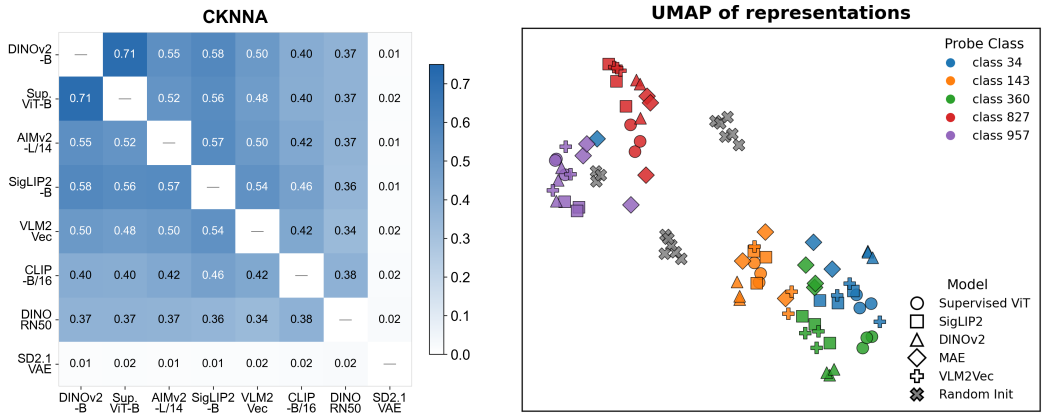

Figure 2: **Cross-model representation alignment on ImageNet-1K**. **LEFT:** Pairwise CKNNA scores with 10 neighbors among 8 diverse independently trained models. All discriminative models exhibit substantial mutual alignment despite differing architectures and training objectives. SD-VAE [54] serves as comparision. **RIGHT:** UMAP of the jointly embedded relational representations after per-model centering and row-ℓ_2 normalization. Samples from different models cluster primarily by semantic class rather than by model identity, providing a geometric view of the convergence.

with coefficients chosen so that both the input and the regression target of F_θ have unit variance at every σ [31]:

$$c_{\text{skip}}(\sigma) = \frac{\sigma_{\text{data}}^2}{\sigma^2+\sigma_{\text{data}}^2}, \quad c_{\text{out}}(\sigma) = \frac{\sigma\,\sigma_{\text{data}}}{\sqrt{\sigma^2+\sigma_{\text{data}}^2}}, \quad c_{\text{in}}(\sigma) = \frac{1}{\sqrt{\sigma^2+\sigma_{\text{data}}^2}}, \quad \lambda(\sigma) = \frac{\sigma^2+\sigma_{\text{data}}^2}{\sigma^2\,\sigma_{\text{data}}^2}, \quad (6)$$

and $c_{\text{noise}}(\sigma) = \frac{1}{4}\log\sigma$, supplied to F_θ as a sinusoidal time embedding [64] so that the noise scale enters the network as an additive token-level signal. Here σ_{data} is the empirical standard deviation of z after a per-dimension standardization step that we apply to every encoder output before training.

3 The Platonic Representation Defense

In this section, we evaluate the aforementioned phenomenon of representation convergence and propose an energy-based modeling framework to model the compatibility of representations.

Threat model. The attacker controls upstream pre-training and releases a backdoored encoder $\tilde{f}$ that behaves normally on clean inputs ($\tilde{f}(x) \approx f(x)$) but, on any input carrying a predefined trigger, maps the image into the embedding cluster of an attacker-chosen target class [29, 6, 56, 38]. We consider diverse backdoor families in existing literature including additive image-space patches [56, 6], optimization backdoor [60], and frequency-domain backdoor [35]. The defender accesses neither the upstream pipeline nor the training data. We further restrict the defender to the representation space of the audited encoder: it only feeds inputs to $\tilde{f}$ and observes the output embeddings $\tilde{f}(x)$, with no access to its weights, gradients, intermediate activations, or architectural details. This is the sense in which our defense is black-box. The defender can access publicly available encoders such as DINOv2 [50] and CLIP [52]. The goal is to build test-time detectors and purifiers for poisoned representations without degrading clean-input utility. We additionally evaluate on adversarial attacks [14, 44] to further demonstrate the effectiveness of our method.

Empirical observation. We investigate four families of models that span different training paradigms and modalities: **(i)** *Visual SSL encoders* (e.g., DINOv2 [50]): self-supervised models that produce semantically rich representations without label supervision; **(ii)** *Vision–language encoders* (e.g., CLIP [52]): contrastive models that align visual and textual modalities in a shared embedding space; **(iii)** *Multi-modal Large Language Models (MLLMs)* (e.g., Qwen2-VL [66]): Large language models with visual encoders; **(iv)** *Generative Encoders* (e.g., AIMv2 [12]): Generative encoders trained with autoregressive pre-training [12] or MAE-style pre-training [21]. We extracted a global representation for Qwen2-VL following VLM2Vec [30]. As shown in Figure 2, substantial alignment emerges among all models. Appendix Figure 6 shows that randomly initialized counterparts

of the same architectures exhibit near-zero alignment (<0.01 CKNNA), confirming that alignment arises from learning other than architectural inductive bias. This convergence has a security implication that, to our knowledge, has not been exploited. *If two independently trained models agree on what an image "means," then disagreement is evidence of manipulation.*

We operationalize this intuition by modeling the joint compatibility of a source representation and a fixed ensemble of K clean reference representations as an energy function. Given a source encoder f^{s} (potentially compromised) and K independently trained reference encoders $\{f^{\mathsf{r}_k}\}_{k=1}^{K}$ drawn from the families above, we extract $z^{\mathsf{s}} = f^{\mathsf{s}}(x) \in \mathbb{R}^{d_s}$ and $z^{\mathsf{r}_k} = f^{\mathsf{r}_k}(x) \in \mathbb{R}^{d_{r_k}}$, and introduce a scalar *platonic energy*

$$E_\theta\big(z^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}\big) \in \mathbb{R} \tag{7}$$

that assigns each tuple a real-valued compatibility score. Following the standard energy-based modeling convention [34], E_θ induces an unnormalized conditional density over the source latent,

$$p_\theta\big(z^{\mathsf{s}} \,\big|\, z^{\mathsf{r}_{1:K}}\big) \;\propto\; \exp\big(-E_\theta(z^{\mathsf{s}}, z^{\mathsf{r}_{1:K}})\big), \tag{8}$$

whose low-energy region captures tuples in which the source and reference representations agree. The reference representations act as clean, uncorruptible *conditions*. Once trained, E_θ provides two complementary primitives for defense. **Detection:** for a query image x with observed source latent $z^{\mathsf{s}} = \tilde{f}^{\mathsf{s}}(x)$ and clean reference features $z^{\mathsf{r}_{1:K}}$, the energy $E_\theta(z^{\mathsf{s}}, z^{\mathsf{r}_{1:K}})$ serves directly as a scalar anomaly score. **Purification:** we can follow $-\nabla_{z^{\mathsf{s}}} E_\theta$ to pull an observed latent toward this low-energy region, yielding a purified estimate $\hat{z}^{\mathsf{s}}$ that is consistent with the references. In Section 4, we elaborate on learning an E_θ that supports these two uses.

Detection guarantee from representation convergence. Before turning to training, we show that an E_θ that is close to the Bayes-optimal platonic energy inherits a guarantee *directly* from the representation convergence. Let P_{SR} denote the distribution of matched tuples $(Z^{\mathsf{s}}, Z^{\mathsf{r}_{1:K}})$ extracted from the same image, and let $P_\perp := P_S \otimes P_R$ denote the distribution of cross-sample mismatched tuples in which Z^{s} is drawn independently of the references. Define the oracle energy

$$E^\star\big(z^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}\big) \;=\; -\log p\big(z^{\mathsf{s}} \,\big|\, z^{\mathsf{r}_{1:K}}\big) + \psi\big(z^{\mathsf{r}_{1:K}}\big), \tag{9}$$

which is the Bayes-optimal target of Eq. (8) up to a reference-only shift ψ [34]. Then we characterize the detection behavior of $E^\star$ and any E_θ that approximates it in mean.

Proposition 1 (Oracle energy gap)**.** *Assume $P_S \ll P_{S|R=r}$ for P_R-a.e. r and the KL terms below are finite. Then the oracle energy $E^\star$ satisfies*

$$\mathbb{E}_{P_\perp}\big[E^\star\big] \;-\; \mathbb{E}_{P_{SR}}\big[E^\star\big] \;=\; I\big(Z^{\mathsf{s}}; Z^{\mathsf{r}_{1:K}}\big) \;+\; \mathbb{E}_{P_R}\Big[D_{\mathrm{KL}}\big(P_S \,\|\, P_{S|R}\big)\Big] \;\geq\; I\big(Z^{\mathsf{s}}; Z^{\mathsf{r}_{1:K}}\big). \tag{10}$$

Moreover, if $|\mathbb{E}_P[E_\theta - E^\star]| \leq \epsilon_P$ for $P \in \{P_{SR}, P_\perp\}$, then $\mathbb{E}_{P_\perp}[E_\theta] - \mathbb{E}_{P_{SR}}[E_\theta] \geq I(Z^{\mathsf{s}}; Z^{\mathsf{r}_{1:K}}) - \epsilon_{SR} - \epsilon_\perp$.

Proposition 2 (High-probability rejection of mismatched sources)**.** *Let $\gamma := \mathbb{E}_{P_\perp}[E_\theta] - \mathbb{E}_{P_{SR}}[E_\theta]$ be the effective energy gap and assume $\gamma > 0$. Suppose further that $E_\theta - \mathbb{E}_P[E_\theta]$ is κ^2-sub-Gaussian under both P_{SR} and $P_\perp$ (which is automatic when E_θ is bounded within $[a, b]$, with $\kappa^2 = (b-a)^2/4$). Setting the threshold $\tau = \mathbb{E}_{P_{SR}}[E_\theta] + \gamma/2$,*

$$\Pr_{P_\perp}\big[E_\theta \leq \tau\big] \;\leq\; \exp\Big(-\tfrac{\gamma^2}{8\kappa^2}\Big), \qquad \Pr_{P_{SR}}\big[E_\theta \geq \tau\big] \;\leq\; \exp\Big(-\tfrac{\gamma^2}{8\kappa^2}\Big), \tag{11}$$

and consequently, treating $P_\perp$ as the positive class with E_θ scored so that higher values are more anomalous and evaluating AUROC on independent draws from $P_\perp$ and P_{SR}, $\mathrm{AUROC}(E_\theta) \geq 1 - 2\exp(-\gamma^2/8\kappa^2)$.

Proofs can be found in Appendix B. Proposition 1 identifies the mutual information $I(Z^{\mathsf{s}}; Z^{\mathsf{r}_{1:K}})$ as the energy gap which is exactly the quantity whose strict positivity is the empirical content of the Plato Representation Hypothesis. Proposition 2 upgrades the expectation-level gap to a per-sample guarantee via a standard Chernoff bound for sub-Gaussian variables [65].

4 Training the Platonic Energy

We pursue two training routes for the platonic energy of Eq. (7), sharing the same architecture but differing in readout and loss: a *discriminative* route that fits E_θ as a scalar compatibility score with NCE for backdoor detection, and a *score-matching* route that fits a conditional denoiser $\hat{D}_\theta$ which yields a score field for feature purification.

Algorithm 1 Discriminative training.

Require: source z^{s}, references $\{z^{\mathsf{r}_k}\}_{k=1}^{K}$, negative count M, position encoding $e_{\mathsf{s}}, \dots, e_{\mathsf{r}_k}$

1: $h_0^{\mathsf{s}} \leftarrow g^{\mathsf{s}}(z^{\mathsf{s}}) + e_{\mathsf{s}}$ ▷ project to d_{tok}, add space embed
2: **for** $k = 1, \dots, K$ **do**
3: $\quad h_0^{\mathsf{r}_k} \leftarrow g^{\mathsf{r}_k}(z^{\mathsf{r}_k}) + e_{\mathsf{r}_k}$
4: **end for**
5: $[\tilde{h}_\theta^{\mathsf{s}},\, \tilde{h}_\theta^{\mathsf{r}_{1:K}}] \leftarrow \mathrm{Transformer}([h_0^{\mathsf{s}},\, h_0^{\mathsf{r}_{1:K}}])$
6: $-E_\theta \leftarrow \alpha \cdot \frac{1}{K}\sum_{k=1}^{K} \cos(\tilde{h}_\theta^{\mathsf{s}}, \tilde{h}_\theta^{\mathsf{r}_k}) + \beta$ ▷ Eq. (12)
7: **Training:** draw M mismatched tuples from $P_\perp$, recompute $-E_\theta^{(m)}$ via lines 1–5
8: $\mathcal{L}_{\mathrm{NCE}} \leftarrow -\log \mathrm{sig}(-E_\theta) - \sum_{m=1}^{M} \log(1 - \mathrm{sig}(-E_\theta^{(m)}))$ ▷ Eq. (13)

Ensure: scalar energy $E_\theta \in \mathbb{R}$

Algorithm 2 cDSM training.

Require: standardized source z^{s}, references $\{z^{\mathsf{r}_k}\}_{k=1}^{K}$, noise level σ, position encoding $e_{\mathsf{s}}, \dots, e_{\mathsf{r}_k}$, time embedding module η

1: $\epsilon \sim \mathcal{N}(0, I_{d_s}), \quad \tilde{z}^{\mathsf{s}} \leftarrow z^{\mathsf{s}} + \sigma\,\epsilon$ ▷ Eq. (2)
2: Compute $c_{\mathrm{in}}(\sigma)$, $c_{\mathrm{out}}(\sigma)$, $c_{\mathrm{skip}}(\sigma)$, $c_{\mathrm{noise}}(\sigma)$ ▷ Eq. (6)
3: $h_0^{\mathsf{s}} \leftarrow g^{\mathsf{s}}\big(c_{\mathrm{in}}(\sigma)\,\tilde{z}^{\mathsf{s}}\big) + \eta\big(c_{\mathrm{noise}}(\sigma)\big) + e_{\mathsf{s}}$
4: **for** $k = 1, \dots, K$ **do**
5: $\quad h_0^{\mathsf{r}_k} \leftarrow g^{\mathsf{r}_k}(z^{\mathsf{r}_k}) + e_{\mathsf{r}_k}$
6: **end for**
7: $[\tilde{h}_\theta^{\mathsf{s}},\, \cdot] \leftarrow \mathrm{Transformer}([h_0^{\mathsf{s}},\, h_0^{\mathsf{r}_{1:K}}])$
8: $F_\theta \leftarrow \pi\big(\tilde{h}_\theta^{\mathsf{s}}\big) \in \mathbb{R}^{d_s}$ ▷ residual readout
9: $\hat{D}_\theta \leftarrow c_{\mathrm{skip}}(\sigma)\,\tilde{z}^{\mathsf{s}} + c_{\mathrm{out}}(\sigma)\,F_\theta$ ▷ EDM precond
10: **Training:** $\mathcal{L}_{\mathrm{cDSM}} \leftarrow \lambda(\sigma)\,\big\|\hat{D}_\theta - z^{\mathsf{s}}\big\|_2^2$ ▷ Eq. (14)

Ensure: denoised $\hat{D}_\theta \in \mathbb{R}^{d_s}$; score $s_\theta = (\hat{D}_\theta - \tilde{z}^{\mathsf{s}})/\sigma^2$

Shared feature-space backbone architecture. The network for both routes shares the same first two stages, after which the readout is route-dependent. **(i)** Per-space projection heads $g^{\mathsf{s}} : \mathbb{R}^{d_s} \to \mathbb{R}^{d_{\mathrm{tok}}}$ and $g^{\mathsf{r}_k} : \mathbb{R}^{d_{r_k}} \to \mathbb{R}^{d_{\mathrm{tok}}}$ $(k = 1, \dots, K)$ map heterogeneous-dimensional encoder outputs to a shared token dimension d_{tok} via three-layer LayerNorm-terminated MLPs. **(ii)** A pre-LayerNorm Transformer [64] processes the $(K{+}1)$-token sequence, letting the source token attend to every reference token. The discriminative route reads E_θ off the post-transformer source/reference tokens via the cosine head of Eq. (12). The score-matching route instead realizes $\hat{D}_\theta(\tilde{z}^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}, \sigma)$ with an additional MLP readout $\pi : \mathbb{R}^{d_{\mathrm{tok}}} \to \mathbb{R}^{d_s}$, which maps the post-transformer source token back to the encoder dimension, and a time embedding module $\eta : \mathbb{R} \to \mathbb{R}^{d_{\mathrm{tok}}}$ to inject noise-level information. We refer readers to EDM [31] for more details on the denoiser parameterization, and to Section C for a more detailed architectural description. Figure 7 visualizes the backbone architecture. Algorithms 1 and 2 concretely describe the training process of two routes.

Discriminative training: a sigmoid-NCE method. We first introduce a discriminative training route that uses sigmoid-NCE to learn the matched-versus-mismatched density ratio underlying the effective gap γ in Proposition 2. We parameterize the discriminative training following SigLIP [69]:

$$-E_\theta\big(z^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}\big) \;:=\; \alpha \cdot \tfrac{1}{K}\sum_{k=1}^{K} \cos\big(\tilde{h}_\theta^{\mathsf{s}},\, \tilde{h}_\theta^{\mathsf{r}_k}\big) \;+\; \beta, \tag{12}$$

where $\tilde{h}_\theta^{\mathsf{s}}, \tilde{h}_\theta^{\mathsf{r}_k} \in \mathbb{R}^{d_{\mathrm{tok}}}$ are the post-transformer source and reference token embeddings, and $\alpha = e^{\log \alpha} > 0$, $\beta \in \mathbb{R}$ are two learned scalars (with β initialised to $-\log M$ so that the initial log-odds match a uniform prior over the $M{+}1$ candidates). For each clean tuple $(z^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}) \sim P_{SR}$ extracted from a single image (the positive) we draw M mismatched negatives. The training loss is the standard binary noise-contrastive objective [17, 48]

$$\mathcal{L}_{\mathrm{NCE}}(\theta) \;=\; -\,\mathbb{E}_{P_{SR}}\big[\log \mathrm{sig}\big(-E_\theta\big)\big] \;-\; M\,\mathbb{E}_{P_\perp}\big[\log\big(1 - \mathrm{sig}(-E_\theta)\big)\big], \qquad \mathrm{sig}(t) := \tfrac{1}{1+e^{-t}}, \tag{13}$$

whose Bayes-optimal logit equals $\log\big(p(z^{\mathsf{s}} \mid z^{\mathsf{r}_{1:K}})/(M\,p(z^{\mathsf{s}}))\big)$, so E_θ converges in expectation to $E^\star$ of Eq. (9) up to a sum of a z^{s}-only and a z^{r}-only shift that both cancel in the energy gap of Eq. (10) [17]. The discriminative variant produces only a scalar for compatibility judgement, but it cannot induce the score field that the feature purification requires, motivating the DSM-based instantiation that follows.

Conditional denoising score matching (cDSM) objective. We train the network with the conditional form of Eq. (3),

$$\mathcal{L}_{\mathrm{cDSM}}(\theta) \;=\; \mathbb{E}_{\sigma,\,(z^{\mathsf{s}},\,z^{\mathsf{r}_{1:K}}),\,\epsilon}\Big[\,\lambda(\sigma)\,\big\|\hat{D}_\theta\big(z^{\mathsf{s}} + \sigma\epsilon,\; z^{\mathsf{r}_{1:K}},\; \sigma\big) - z^{\mathsf{s}}\big\|_2^2\Big], \tag{14}$$

with $\sigma \sim \mathrm{LogUniform}[\sigma_{\min}, \sigma_{\max}]$, $\epsilon \sim \mathcal{N}(0, I_{d_s})$, and $\lambda(\sigma)$ as in Eq. (6). At its optimum, Eq. (14) recovers the Bayes-optimal *conditional* denoiser $\hat{D}^\star(\tilde{z}^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}, \sigma) = \mathbb{E}[z^{\mathsf{s}} \mid \tilde{z}^{\mathsf{s}}, z^{\mathsf{r}_{1:K}}, \sigma]$ of

Table 1: The (CA %, PA %, AUC, ASR %) of defense methods against different backdoor attacks on ImageNet-1K. The adversarial attacks are implemented in an untargeted manner. † denotes our implementation.

Attack	No Defense			Decomp [22] (ICML 2025)				DeDe [24] (CVPR 2025)				ZIP[57] (NeurIPS 2023)			Beatrix [43] (NDSS 2023)	DetectCLIP [25] (ICLR 2025)	Platonic Defense (Ours)			
	CA ↑	PA ↑	ASR ↓	CA ↑	PA ↑	AUC ↑	ASR ↓	CA† ↑	PA† ↑	AUC ↑	ASR† ↓	CA ↑	PA ↑	ASR ↓	AUC ↑	AUC ↑	CA ↑	PA ↑	AUC ↑	ASR ↓
Backdoor attacks on unimodal encoders																				
SSL-Backdoor [56]	76.49	0.03	99.96	76.75	0.12	**1.00**	99.97	31.1	0.2	0.65	37.1	66.73	0.07	100.0	0.96	0.55	**77.47**	**74.13**	0.99	**0.10**
CorruptEncoder [6]	76.25	0.08	99.90	76.35	0.16	**1.00**	99.89	30.2	0.3	0.69	46.3	66.33	0.07	99.93	0.97	0.31	**77.31**	**73.94**	**1.00**	**0.30**
CTRL [35]	76.59	0.03	99.96	76.93	0.14	**1.00**	99.96	30.1	0.2	0.93	38.5	66.40	57.20	**0.07**	**1.00**	0.49	**77.42**	**64.65**	0.99	1.98
BLTO [60]	76.66	3.76	94.44	76.87	3.28	**0.96**	95.36	29.94	1.34	0.77	34.57	64.40	**63.93**	**0.40**	0.89	0.60	**77.34**	55.80	0.84	3.31
NA [8]	74.68	0.10	100.0	74.98	0.11	**1.00**	100.00	31.0	0.2	0.76	**0.6**	65.73	0.07	100.0	0.95	0.00	**77.41**	**49.55**	0.95	18.15
BadEncoder [29]	75.78	0.10	100.0	75.83	0.11	**1.00**	100.00	28.4	0.2	0.78	28.3	66.47	0.07	100.0	**1.00**	0.05	**78.05**	**74.44**	0.84	**0.16**
DRUPE [61]	72.93	0.10	100.0	73.11	0.11	0.99	100.00	26.2	0.1	0.86	51.4	63.73	0.07	100.0	**1.00**	0.14	**77.32**	**72.91**	0.99	**0.03**
Average	75.63	0.60	99.18	75.83	0.58	**0.99**	99.31	29.56	0.36	0.78	33.82	65.68	17.35	71.49	0.97	0.31	**77.47**	**66.49**	0.94	**3.43**
Backdoor attacks on image-text encoders																				
CLIP-Backdoor [6]	53.25	0.24	99.51	54.60	8.97	0.95	65.33	14.8	1.4	0.69	22.1	43.47	1.07	98.06	0.76	**1.00**	**63.99**	**57.41**	0.95	**0.00**
BadCLIP [38]	51.01	0.55	98.08	52.08	0.40	0.48	96.24	16.4	1.0	0.68	10.7	40.47	19.80	5.54	0.41	**1.00**	**64.07**	**52.19**	0.94	**0.00**
Average	52.13	0.40	98.80	53.34	4.69	0.72	80.79	15.60	1.20	0.69	16.40	41.97	10.44	51.80	0.59	**1.00**	**64.03**	**54.80**	0.95	**0.00**
Adversarial Attacks																				
FGSM [14]	78.30	2.20	97.80	78.40	2.20	0.47	97.80	48.6	2.7	0.53	97.3	67.80	43.80	56.20	0.50	**0.75**	**78.6**	**58.4**	0.63	**41.60**
PGD [44]	78.30	0.00	100.00	78.40	0.00	0.45	100.00	48.6	0.00	0.54	100.00	68.00	57.50	42.50	0.30	0.89	**78.6**	**61.8**	**0.92**	**38.20**
MI-FGSM [10]	78.30	0.00	100.00	78.40	0.00	0.45	100.00	48.6	0.00	0.54	100.00	67.80	47.80	52.20	0.31	0.87	**78.6**	**53.8**	**0.88**	**46.20**
CW-L2 [7]	78.30	0.00	100.00	78.50	2.00	0.49	98.00	43.5	18.0	0.49	82.0	68.33	67.67	32.33	0.50	0.47	**78.6**	**75.0**	**0.53**	**25.00**
AutoAttack [9]	78.30	0.00	100.00	77.40	0.00	0.46	100.00	44.6	2.6	0.55	97.4	68.00	60.33	39.67	0.34	**0.96**	**78.6**	**69.4**	0.83	**30.60**
Average	78.30	0.44	99.56	78.22	0.84	0.46	99.16	46.78	4.66	0.53	95.34	67.99	55.42	44.58	0.39	**0.79**	**78.60**	**63.68**	0.76	**36.32**

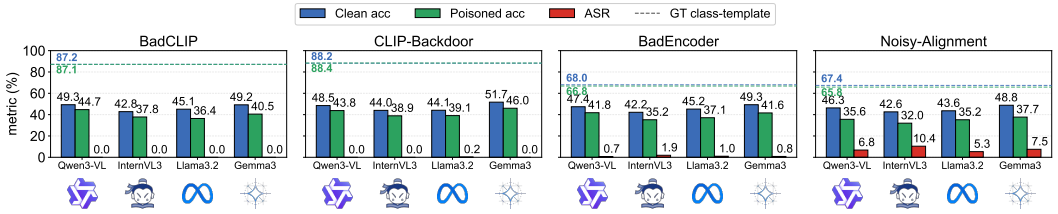

Figure 3: Evaluations on the text condition.

the *clean* joint distribution. By Tweedie's identity (Eq. (4)), we can derive the conditional score $(\hat{D}_\theta - \tilde{z}^{\mathsf{s}})/\sigma^2$ that approximates $\nabla_{\tilde{z}^{\mathsf{s}}} \log p_\sigma(\tilde{z}^{\mathsf{s}} \mid z^{\mathsf{r}_{1:K}})$, i.e., the gradient of the implicit platonic energy with respect to the source latent. For each mini-batch we draw per-sample Bernoulli masks with drop probability 0.1 to implement classifier-free-guidance [23] (CFG).

5 Experiments

5.1 Setup

For defense evaluation, we consider 3 types of attacks: backdoor attacks to unimodal encoders, backdoor attacks to image-text encoders, and adversarial attacks. For unimodal visual encoders, we poison DINO ViT-B/16 on ImageNet-1K and cover 7 representative SSL backdoor attacks: SSL-Backdoor [56], CorruptEncoder [6], CTRL [35], BLTO [60], Noisy-Alignment [8], BadEncoder [29], and DRUPE [61]. These attacks span patch-based, optimization-based, frequency-domain, and distribution-preserving poisoning mechanisms. For image–text encoders, we poison CLIP ViT-B/16 on a 60K subset of CC3M and evaluate CLIP-Backdoor [6] and BadCLIP [38]. We additionally test FGSM [14], PGD [44], MI-FGSM [10], CW-L2 [7], and AutoAttack [9] to assess whether the same representation-compatibility principle extends beyond backdoor attacks.

We compare against recent test-time defenses, including Decomp [22], DeDe [24], ZIP [57], Beatrix [43], and DetectCLIP [25]. We report clean accuracy (CA), poisoned accuracy (PA), area under the ROC curve (AUC), and attack success rate (ASR). For multimodal encoders, we use zero-shot accuracy [52] for evaluation. Otherwise, we use linear probing accuracy. Unless otherwise specified, Platonic Defense uses $K{=}2$ clean reference encoders: the visual branch of SigLIP2-Base/16 [63] and DINOv2-Base/16. We use the Heun ODE solver [55] (20 steps) for cDSM inference. The conditional energy and DSM purifier share the same backbone architecture with 12 layers and hidden dimension 1536, trained only on a subset of ImageNet-1K unavailable during poisoning to avoid overfitting. More details are provided in Appendix C.

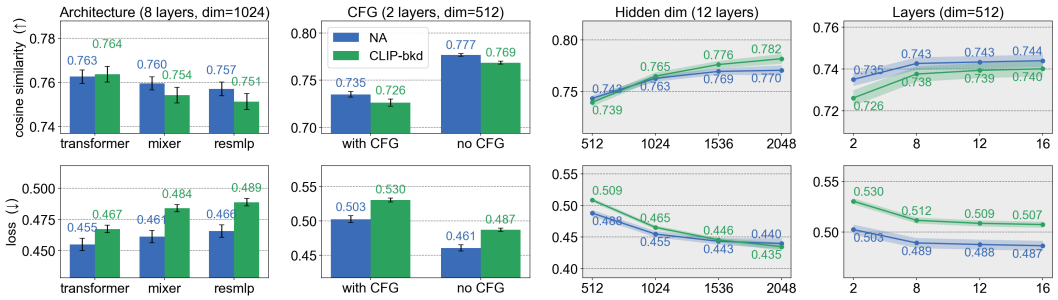

Figure 4: Evaluations on the architecture.

Table 2: Defense against enhanced attack. CR means recovered cosine similarity. ✓ denotes a reference that receives the backdoored image.

DINOv2	SigLIP2	AIMv2	VLM2Vec	CA ↑	PA ↑	ASR ↓	CR ↑
Defense upper bound							
				62.5	**62.5**	**0.0**	**0.683**
One backdoored reference							
✓				59.0	58.9	**0.0**	**0.654**
	✓			32.7	32.4	25.7	0.460
		✓		59.6	59.6	**0.0**	0.594
			✓	**60.0**	**59.9**	**0.0**	0.629
Two backdoored references							
✓	✓			4.3	4.1	70.3	0.374
✓		✓		49.5	49.1	**0.0**	0.532
✓			✓	**52.7**	**52.8**	0.5	**0.581**
	✓	✓		8.0	7.3	46.2	0.314
	✓		✓	6.1	5.7	78.2	0.365
		✓	✓	52.6	52.7	0.5	0.515
Three backdoored references							
✓	✓	✓		0.3	0.3	84.3	0.234
✓	✓		✓	0.3	0.3	96.0	0.272
✓		✓	✓	**34.9**	**34.3**	**3.9**	**0.451**
	✓	✓	✓	0.4	0.4	67.9	0.217
Worst case							
✓	✓	✓	✓	0.0	0.0	100.0	0.138

5.2 Main Results

Table 1 shows that platonic defense is the only method that consistently preserves clean utility while suppressing attacks across unimodal backdoors, image–text backdoors, and adversarial perturbations. On 7 unimodal SSL backdoors, it improves the average poisoned accuracy from 0.60% without defense to 66.49% and reduces ASR from 99.18% to 3.43%, while slightly increasing clean accuracy to 77.47%. Competing defenses either preserve accuracy but leave ASR near 100% (Decomp) or reduce ASR by sacrificing most clean accuracy (DeDe). The same pattern holds for image–text backdoors, where our method reaches 64.03% clean accuracy, 54.80% poisoned accuracy, and 0.00% ASR on average, and extends to untargeted adversarial attacks with the best average poisoned accuracy (63.68%) and lowest ASR (36.32%). Figure 3 further shows that the text-conditioned variant remains effective. We train our method on CC3M with the text branch of SigLIP2-Base/16 and CLIP ViT-B/16. Since ImageNet-1K does not provide textual descriptions, we generate image captions using two strategies: (i) class-template labels following CLIP [52]; and (ii) off-the-shelf VLMs, including Llama3.2-11B-Vision [15], Gemma3-27B [13], Qwen3-VL-4B-Instruct [2], and InternVL3-2B [73]. For all VLMs, we use the prompt: "Describe this image in one short sentence."

5.3 Ablation Studies

Evaluation on enhanced attack. We consider a more challenging scenario in Table 2, where we assume the attacker is aware of our defense and has access to the reference encoders. With all references clean, the defense reaches 62.5% CA/PA and 0.0% ASR. Corrupting one reference is usually tolerated: three of four choices keep ASR at 0.0% with PA around 59–60%. With two corrupted references, the defense remains effective when the corrupted set excludes SigLIP2 (ASR 0.0–0.5%, PA up to 52.8%). Three or four corrupted references largely exceed the intended threat model.

Inference overhead. Figure 5 probes the cost–accuracy trade-off of the purification solver. Stochastic Langevin sampling [58] is ineffective even with 800 network-function evaluations (NFEs). Deterministic ODE solvers are substantially more favorable. The 10-step Heun solver attains the strongest PA in this sweep (60.3% across the two attacks) with only 36 NFEs. More aggressive settings reduce the cost further, down to 2–18 NFEs for Euler and DPM++ [41], but their accuracy becomes attack-dependent. 2-step Euler/DPM++ preserves CLIP-Backdoor accuracy

Table 3: Backbone transfer between the discriminative and cDSM training. Gap means platonic energy gap. "B" prefix means baseline.

Source	Variant	B-cDSM Loss	B-cDSM Cos	B-cDSM PA	B-cDSM ASR	Disc. → cDSM Loss	Disc. → cDSM Cos	Disc. → cDSM PA	Disc. → cDSM ASR	B-disc Loss	B-disc Gap	B-disc AUC	cDSM → Disc. Loss	cDSM → Disc. Gap	cDSM → Disc. AUC
CLIP-Backdoor	Warm-start					0.605	0.678	0.496	0.000				0.147	3.18	0.9999
	Frozen probe	0.571	0.701	0.554	0.000	0.773	0.553	0.100	0.000	0.144	3.24	0.9999	0.319	1.35	0.9954
	Joint train					0.591	0.688	0.514	0.000				0.143	3.23	1.0000
Noisy-Alignment	Warm-start					0.619	0.659	0.466	0.116				0.148	3.17	0.9999
	Frozen probe	0.581	0.685	0.458	0.182	0.844	0.489	0.120	0.006	0.143	3.25	0.9988	0.271	1.80	0.9972
	Joint train					0.603	0.671	0.596	0.042				0.142	3.26	1.0000

while dropping Noisy-Alignment to 11.2%. We therefore use 20-step Heun as a conservative default for the main evaluation.

Architecture. Figure 4 ablates the architecture, reporting the denoised cosine similarity and cDSM training loss (eq. 14). The transformer is consistently the strongest backbone among the tested token mixers: MLP-Mixer [16] and ResMLP [37]. Increasing the transformer width has a larger effect than increasing depth: at 12 layers, the cosine improves from $0.743/0.739$ at width 512 to $0.769/0.776$ at width 1536 and saturates near width 2048, whereas increasing depth from 8 to 16 at width 512 changes cosine by less than 0.002. These trends motivate our default 12-layer, 1536-width Transformer, which captures most of the width-scaling gain without paying for the marginal 2048-width improvement. Removing the unconditional branch improves raw denoising metrics, but we keep CFG because it gives a principled way to tune conditioning strength and allows us to drop partial conditions to reduce inference cost when needed. Table 3 tests whether the two objectives can be joint trained. Warm-start initializes one branch from the model trained for the other objective and then fine-tunes it. Frozen probe directly loads the trained model and evaluates it without adapting the backbone. Joint training optimizes $\mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{cDSM}}$ from scratch. The transfer is asymmetric. Moving from cDSM to the discriminative branch works well but the reverse direction is less reliable, especially without fine-tuning. This gap suggests that detection and purification use related but not identical representations. Joint training largely closes the gap: the discriminative branch stays at baseline AUROC, and the cDSM branch approaches standalone cDSM performance with only a small loss on CLIP-Backdoor (PA 0.514 vs. 0.554) while improving the harder Noisy-Alignment case (PA 0.596 vs. 0.458, ASR 0.042 vs. 0.182).

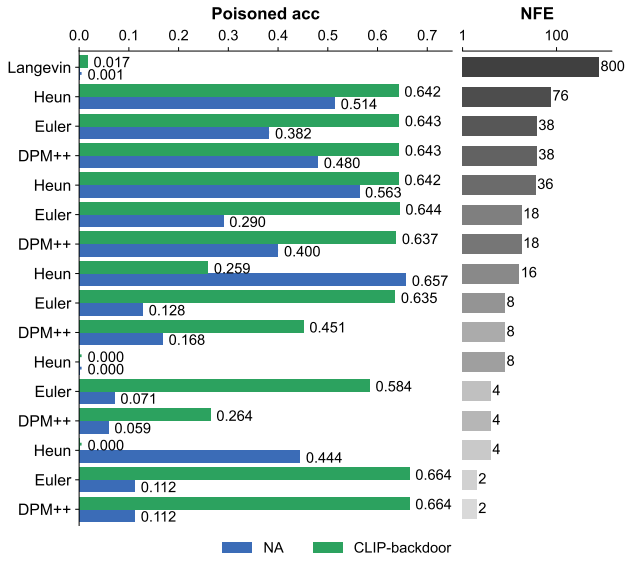

Figure 5: Inference efficiency probing.

6 Related Work

6.1 Backdoor Attacks on Visual Encoders

A backdoor implanted in a pre-trained visual encoder, via poisoned pre-training data or modified weights, is inherited by all downstream tasks built on it. For self-supervised learning (SSL), Saha *et al.* [56] showed that adding a patch trigger to a small fraction of unlabeled target-category images suffices; later poisoning attacks strengthen the trigger–target association under SSL augmentations via optimized image layouts [71, 8], frequency-domain triggers [35], or bi-level trigger optimiza-

tion [60]. BadEncoder [29] instead fine-tunes a clean encoder to embed triggered images near target-class references; DRUPE [61] makes such poisoned embeddings in-distribution and less concentrated to evade defenses, and INACTIVE [70] optimizes imperceptible triggers disentangled from SSL augmentations. For image–text encoders, Carlini and Terzis [6] first backdoored CLIP-style models by poisoning only 0.01% of web-scale image–caption pairs, and BadCLIP [38] aligns the trigger with both textual and visual target embeddings to resist detection and clean fine-tuning.

6.2 Backdoor Detection and Mitigation on Visual Encoders

Defending visual encoders is harder than defending supervised classifiers: labels and downstream tasks are unavailable, and Li *et al.* [36] showed that backdoor and benign representations are deeply entangled in contrastive learning, so defenses designed for supervised models transfer poorly. DECREE [11] detects backdoored encoders by inverting a minimal trigger that maps arbitrary inputs to highly similar embeddings. To repair encoders, MIMIC [18] distills a randomly initialized student from the backdoored encoder, using mutual information to locate benign knowledge, and Mudjacking [40] patches the encoder given a single reported misclassified input; for CLIP, CleanCLIP [3] fine-tunes with additional unimodal self-supervised objectives, while RoCLIP [68] pre-trains by periodically matching each image to its most similar caption in a random pool. Test-time defenses instead detect triggered inputs via decoder reconstruction discrepancies (DeDe [24]), infected attention heads of CLIP (Decomp [22]), or the sparse local neighborhoods of poisoned samples (DetectCLIP [25]); ZIP [57] and Beatrix [43], designed for supervised classifiers, purify inputs with a pre-trained diffusion model and detect them via Gram-matrix statistics of activations, respectively. Unlike these methods, our defense relies only on the agreement between the inspected encoder and reference encoders.

6.3 Similarity on Representations

Representational similarity analysis asks whether independently trained networks learn similar internal representations. Centered Kernel Alignment (CKA) [33] enables such comparisons across widths, depths, and architectures: wide and deep networks learn similar representations outside a model-specific block structure [47], whereas Vision Transformers have more uniform representations across layers than CNNs [53]. Model stitching further shows that strong networks trained very differently, e.g., with supervised and self-supervised objectives, can be connected by a simple trainable layer without loss in performance [4]. At larger scale, the Platonic Representation Hypothesis [27] posits that more capable models converge toward a shared representation regardless of architecture, objective, or modality. Consistent with this view, unaligned vision and language encoders are semantically similar enough to be matched without training [46], and Acevedo *et al.* [1] used the information imbalance to find semantic convergence across languages, modalities, and architectures, with cross-modal predictability peaking in the final layers of vision encoders, while Koepke *et al.* [32] cautioned that image and text representations share only coarse rather than fine-grained structure. The convergence is also not uniform across depth: Mahaut and Baroni [45] showed that layers at similar relative depths across vision models are the most similar, yet classifiers discard low-level image statistics in their final layers and transformers change representations more smoothly than CNNs. Beyond model-to-model comparisons, Cai *et al.* [5] tested 45 vision models on low-level properties of human vision such as contrast detection, masking, and constancy, and found that models such as DINOv2 and CLIP partially share them.

7 Conclusion

We introduced a black-box test-time method for defending SSL encoders against backdoor attacks. The central idea is to use the compatibility between representations from independently trained models as a security signal. We formalized it as a conditional energy model and trained it with noise-contrastive estimation and conditional denoising score matching. Our analysis connects the energy gap to the mutual information between source and reference representations, giving a theoretical guarantee for the defense. Empirically, the method improves robustness across multiple SSL encoders and a broad set of attacks, while preserving clean utility. Extending the same principle to language, audio, and fully multimodal latent spaces is an important direction for future work.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62676086; partially supported by SEU Kunpeng&Ascend Center of Cultivation; and supported in part by the CIPS-SMP-Zhipu Large Model Fund. We thank the Big Data Computing Center of Southeast University for providing facility support for the numerical calculations.

References

[1] Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images. *arXiv preprint arXiv:2505.17101*, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–123, 2023.

[4] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021.

[5] Yancheng Cai, Fei Yin, Dounia Hammou, and Rafal Mantiuk. Do computer vision foundation models learn the low-level characteristics of the human visual system? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20039–20048, 2025.

[6] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2021.

[7] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks . In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, Los Alamitos, CA, USA, May 2017. IEEE Computer Society. doi: 10.1109/SP.2017.49. URL `https://doi.ieeecomputersociety.org/10.1109/SP.2017.49`.

[8] Tuo Chen, Jie Gui, Minjing Dong, Ju Jia, Lanting Fang, and Jian Liu. Backdooring self-supervised contrastive learning by noisy alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3684–3693, 2025.

[9] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[11] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16352–16362, 2023.

[12] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, and Sai Aitharaju. Multimodal autoregressive pre-training of large vision encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[13] Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

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

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[16] Ming Gui, Johannes Schusterbauer, Timy Phan, Felix Krause, Joshua M. Susskind, Miguel Ángel Bautista, and Björn Ommer. Adapting self-supervised representations as a latent space for efficient generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[17] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2), 2012.

[18] Tingxu Han, Weisong Sun, Ziqi Ding, Chunrong Fang, Hanwei Qian, Jiaxun Li, Zhenyu Chen, and Xiangyu Zhang. Mutual information guided backdoor mitigation for pre-trained encoders. *IEEE Transactions on Information Forensics and Security*, 20:3414–3428, 2025. ISSN 1556-6021.

[19] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

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

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[22] Shuo He, Zhifang Zhang, Feng Liu, Roy Ka-Wei Lee, Bo An, and Lei Feng. A closer look at backdoor attacks on clip. In *Forty-Second International Conference on Machine Learning*, 2025.

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

[24] Sizai Hou, Songze Li, and Duanyi Yao. Dede: Detecting backdoor samples for ssl encoders via decoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20675–20684, 2025.

[25] Hanxun Huang, Sarah Monazam Erfani, Yige Li, Xingjun Ma, and James Bailey. Detecting backdoor samples in contrastive language image pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025.

[26] Yuxian Huang, Geng Yang, Dong Yuan, and Shui Yu. Dbssl: A scheme to detect backdoor attacks in self-supervised learning models. *IEEE Transactions on Dependable and Secure Computing*, pages 1–12, 2025. ISSN 1941-0018.

[27] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

[28] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[29] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059, May 2022.

[30] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TE0KOzWYAF.

[31] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[32] A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into plato's cave: Examining cross-modal representational convergence at scale, April 2026.

[33] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

[34] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[35] Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4367–4378, 2023.

[36] Changjiang Li, Ren Pang, Bochuan Cao, Zhaohan Xi, Jinghui Chen, Shouling Ji, and Ting Wang. On the difficulty of defending contrastive learning against backdoor attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2901–2918, 2024.

[37] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[38] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24645–24654, 2024.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, December 2023.

[40] Hongbin Liu, Michael K Reiter, and Neil Zhenqiang Gong. Mudjacking: Patching backdoor vulnerabilities in foundation models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2919–2936, 2024.

[41] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025.

[42] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. URL `https://arxiv.org/abs/2310.04378`.

[43] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The "beatrix" resurrections: Robust backdoor detection via gram matrices. In *Proceedings 2023 Network and Distributed System Security Symposium*, 2023. ISBN 978-1-891562-83-9.

[44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

[45] Matéo Mahaut and Marco Baroni. Similarity of processing steps in vision model representations. *arXiv preprint arXiv:2601.21621*, 2026.

[46] Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14334–14343, 2024.

[47] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KJNcAkY8tY4.

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[49] OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[50] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[51] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, July 2021.

[53] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Gl8FHfMVTZu.

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[55] André Ronveaux and Felix Medland Arscott. *Heun's differential equations*. Clarendon Press, 1995.

[56] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13337–13346, 2022.

[57] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[59] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32211–32252, July 2023.

[60] Weiyu Sun, Xinyu Zhang, Hao LU, Ying-Cong Chen, Ting Wang, Jinghui Chen, and Lu Lin. Backdoor contrastive learning via bi-level trigger optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oxjeePpgSP.

[61] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 2029–2047. IEEE, 2024.

[62] Ajinkya Tejankar, Maziar Sanjabi, Qifan Wang, Sinong Wang, Hamed Firooz, Hamed Pirsiavash, and Liang Tan. Defending against patch-based backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12239–12249, 2023.

[63] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

[65] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

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

[67] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9929–9939, November 2020.

[68] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10678–10691, 2023.

[69] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[70] Hanrong Zhang, Zhenting Wang, Boheng Li, Fulin Lin, Tingxu Han, Mingyu Jin, Chenlu Zhan, Mengnan Du, Hongwei Wang, and Shiqing Ma. Invisible backdoor attack against self-supervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25790–25801, 2025.

[71] Jinghuai Zhang, Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning based backdoor attacks to contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[72] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[73] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

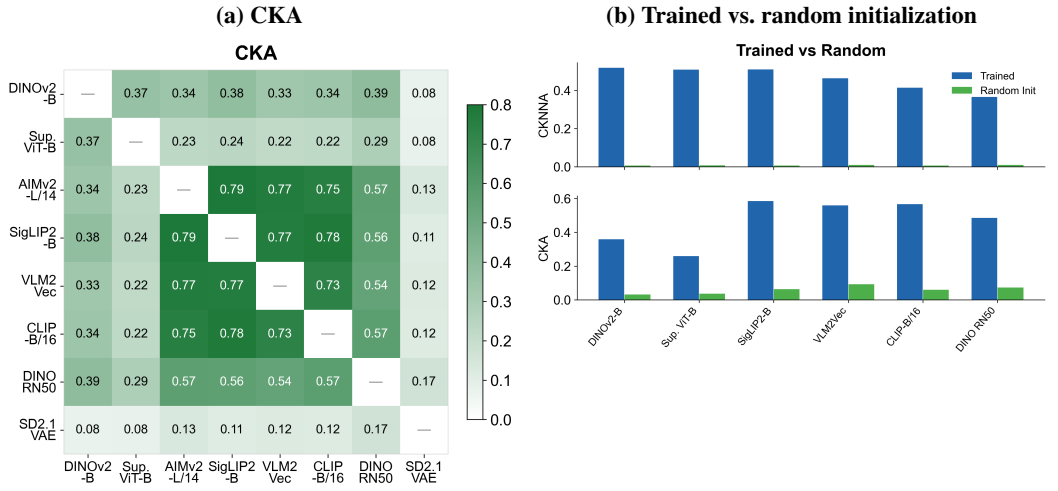

Figure 6: **Additional alignment visualizations complementing Figure 2.** **(a)** Pairwise CKA scores. **(b)** Trained models vs. their random-initialized counterparts on both metrics: trained models achieve 30–100$\times$ higher alignment, confirming that the effect arises from learning rather than architectural inductive bias.

A Limitations

Although the Platonic defense is formulated in a modality-agnostic way, our empirical study focuses primarily on visual representation spaces. We therefore leave a systematic evaluation of multimodal conditioning, such as conditioning vision features on language, audio, or other heterogeneous reference modalities, to future work. Such settings may introduce a stronger modality mismatch than the visual-to-visual cases studied in this paper: different modalities can encode semantic content at different granularities, with different nuisance factors and latent geometries. As a result, the conditional energy or score model may need substantially more paired or weakly aligned data to learn a reliable cross-modal compatibility signal. This requirement could increase both data collection cost and training cost, especially when multiple reference modalities are used simultaneously. Understanding how to reduce this data and compute overhead is an important direction for scaling the method beyond the visual setting.

B Proofs for the Detection Guarantee

Throughout this section, Z^{s} denotes the source latent and Z^{r} denotes the joint reference latent $Z^{\mathrm{r}_{1:K}}$. Let P_{SR} be their matched joint distribution, and let $P_{\perp} = P_S \otimes P_R$ be the independent cross-sample product. We assume that the matched joint distribution admits a regular conditional density $p(s \mid r)$, and that the source marginal admits a density $p_S(s)$, so that the conditional cross-entropies and mutual information below are well defined. We write $I(S;R) := I(Z^{\mathrm{s}}; Z^{\mathrm{r}})$ for brevity.

To avoid degenerate cases in which entropy differences involve undefined $\infty - \infty$ terms, we make the following standing integrability assumption:

$$\begin{aligned} \mathbb{E}_{P_{SR}}\big[\,|\log p(S \mid R)|\,\big] < \infty, \qquad \mathbb{E}_{P_{\perp}}\big[\,|\log p(S \mid R)|\,\big] < \infty, \\ I(S;R) < \infty, \quad \mathbb{E}_{P_R}\Big[D_{\mathrm{KL}}\big(P_S \,\|\, P_{S|R}\big)\Big] < \infty. \end{aligned} \tag{15}$$

These regularity conditions are mild for the continuous, normalized latent representations considered in Section 5; in practice they can be enforced by the standard small-noise or smoothing view of learned features.

B.1 Proof of Proposition 1

Let $E^{\star}(s,r) = -\log p(s \mid r) + \psi(r)$ be the oracle energy of Eq. (9), with $\psi \in L^1(P_R)$, i.e., $\mathbb{E}_{P_R}[\psi(R)] < \infty$. Because the R-marginals of P_{SR} and $P_{\perp}$ coincide, the reference-only term $\psi(R)$

contributes the same expectation to both populations and cancels from the gap:

$$\mathbb{E}_{P_\perp}\big[\psi(R)\big] = \mathbb{E}_{P_R}\big[\psi(R)\big] = \mathbb{E}_{P_{SR}}\big[\psi(R)\big]. \tag{16}$$

By definition, $\mathbb{E}_{P_{SR}}\big[-\log p(S \mid R)\big] = H(S \mid R)$. Under $P_\perp = P_S \otimes P_R$, we treat S and R as independent and apply Fubini's theorem:

$$\mathbb{E}_{P_\perp}\big[-\log p(S \mid R)\big] \;=\; \mathbb{E}_{P_R}\Big[\,\mathbb{E}_{S\sim P_S}\big[-\log p(S \mid R)\big]\Big]. \tag{17}$$

The inner expectation is the cross-entropy of P_S relative to $P_{S|R=r}$. The cross-entropy decomposes as follows:

$$\mathbb{E}_{S\sim P_S}\big[-\log p(S \mid r)\big] \;=\; H(S) \;+\; D_{\mathrm{KL}}\big(P_S \,\big\|\, P_{S|R=r}\big), \tag{18}$$

where $H(S)$ is the differential entropy of S and the KL divergence is non-negative (and, by (15), has finite P_R-expectation). Substituting Eq. (18) into Eq. (17) and subtracting $\mathbb{E}_{P_{SR}}\big[-\log p(S \mid R)\big] = H(S \mid R)$ yields

$$\begin{aligned}\mathbb{E}_{P_\perp}\big[E^\star\big] - \mathbb{E}_{P_{SR}}\big[E^\star\big] \;&=\; \big(H(S) - H(S \mid R)\big) + \mathbb{E}_{P_R}\Big[D_{\mathrm{KL}}\big(P_S \,\|\, P_{S|R}\big)\Big] \\ &=\; I(S;R) + \mathbb{E}_{P_R}\Big[D_{\mathrm{KL}}\big(P_S \,\|\, P_{S|R}\big)\Big] \geq I(S;R).\end{aligned} \tag{19}$$

B.2 Proof of Proposition 2

Let $\mu_\perp := \mathbb{E}_{P_\perp}[E_\theta]$ and $\mu_{SR} := \mathbb{E}_{P_{SR}}[E_\theta]$, so that $\gamma = \mu_\perp - \mu_{SR} > 0$. By the sub-Gaussian assumption [65], the standard one-sided Chernoff bound yields, for any $t > 0$,

$$\Pr_{P_{SR}}\big[E_\theta - \mu_{SR} \geq t\big] \leq \exp\!\big(-t^2/2\kappa^2\big), \qquad \Pr_{P_\perp}\big[\mu_\perp - E_\theta \geq t\big] \leq \exp\!\big(-t^2/2\kappa^2\big). \tag{20}$$

Setting the detection threshold $\tau := \mu_{SR} + \gamma/2 = \mu_\perp - \gamma/2$ and applying Eq. (20) with $t = \gamma/2$,

$$\Pr_{P_{SR}}\big[E_\theta \geq \tau\big] \;=\; \Pr_{P_{SR}}\big[E_\theta - \mu_{SR} \geq \gamma/2\big] \;\leq\; \exp\!\big(-\gamma^2/8\kappa^2\big), \tag{21}$$

$$\Pr_{P_\perp}\big[E_\theta \leq \tau\big] \;=\; \Pr_{P_\perp}\big[\mu_\perp - E_\theta \geq \gamma/2\big] \;\leq\; \exp\!\big(-\gamma^2/8\kappa^2\big). \tag{22}$$

For the AUROC, draw two independent tuples, a mismatched one $(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) \sim P_\perp$ and a matched one $(Z^{\mathsf{s}}, Z^{\mathsf{r}}) \sim P_{SR}$. By the probability interpretation of the AUROC [19], we have

$$\begin{aligned}\mathrm{AUROC}(E_\theta) \;&=\; \Pr\big[E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) > E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}})\big] \;+\; \tfrac{1}{2}\Pr\big[E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) = E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}})\big] \\ &\geq\; \Pr\big[E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) > E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}})\big].\end{aligned} \tag{23}$$

Continuity of E_θ under either marginal (always the case for DSM-trained models with Gaussian noise) merely makes the inequality an equality.

We take a looser *union-bound* version. Since

$$\big\{E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) > \tau\big\} \cap \big\{E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}}) < \tau\big\} \;\subseteq\; \big\{E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) > E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}})\big\}, \tag{24}$$

so by the union bound on the complement,

$$\Pr\big[E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) > E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}})\big] \;\geq\; 1 - \Pr_{P_\perp}[E_\theta \leq \tau] - \Pr_{P_{SR}}[E_\theta \geq \tau] \;\geq\; 1 - 2\exp\!\big(-\gamma^2/8\kappa^2\big), \tag{25}$$

which yields the main-text bound.

A tighter rate is available by working directly with the energy difference: $E_\theta(Z^{\mathsf{s}}_\perp, Z^{\mathsf{r}}_\perp) - E_\theta(Z^{\mathsf{s}}, Z^{\mathsf{r}})$ is sub-Gaussian with parameter $2\kappa^2$ and mean γ by independence, so it is non-positive with probability at most $\exp(-\gamma^2/4\kappa^2)$. We keep the looser two-tail version in the main text because it also upper-bounds the clean- and mismatched-side false-alarm rates at any single threshold τ. □

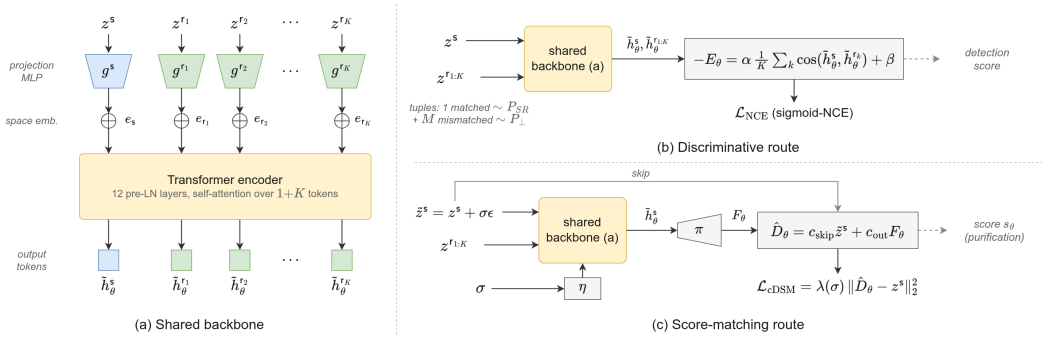

Figure 7: The shared architecture of the platonic energy training for the discriminative and score-matching routes.

C Experiments Details

Architecture instantiation. Both training routes share the backbone architecture of Figure 7 and differ only in the readout. Each space is projected to a common token width d_{tok}=1536 by a 3-layer MLP, Linear–LayerNorm–GELU–Linear–LayerNorm–GELU–Linear. The discriminative route applies a final LayerNorm to every projected token, while the score-matching route leaves the noised source token unnormalised before EDM preconditioning and applies the final LayerNorm only to reference tokens. A learnable space embedding, initialised from $\mathcal{N}(0, 0.02^2)$, is added to each token, and the resulting $1+K$ tokens are processed by a 12-layer pre-norm Transformer with 12 attention heads, GELU feed-forward blocks of width $4d_{\text{tok}}$, and zero dropout. We implement the transformer backbone with the encoder-only transformer from Pytorch Library [51]. For cDSM, the source token additionally receives a sinusoidal noise embedding with 128 Fourier channels followed by a two-layer SiLU MLP into d_{tok}. The residual denoising head is a Linear–GELU–Linear MLP with hidden width d_{tok} and a zero-initialised final layer, so the initial denoiser follows the EDM skip connection.

Discriminative route. The post-transformer source and reference tokens are scored by a cosine readout averaged over the K references; the logit is $\alpha\,\bar{\cos} + \beta$, where $\alpha = \exp(\log \alpha)$ is constrained positive and β is initialised to $-\log M$ for M negatives per positive (both learnable). Training uses sigmoid-BCE against negatives produced by a mismatch sampler that mixes *full* (all spaces deranged), *source-only*, and *partial* (a randomly chosen subset of references replaced) deranged tuples; before the feature queue is warm we derange within the mini-batch, afterwards we draw from a per-space FIFO queue.

Score-matching route. The source feature is first standardised per dimension using statistics $(\mu, \sigma_{\text{dim}})$ pre-computed on clean training data. We adopt the EDM [31] preconditioning with $c_{\text{skip}} = \sigma_d^2/(\sigma^2+\sigma_d^2)$, $c_{\text{out}} = \sigma\sigma_d/\sqrt{\sigma^2+\sigma_d^2}$, $c_{\text{in}} = 1/\sqrt{\sigma^2+\sigma_d^2}$, $c_{\text{noise}} = \frac{1}{4}\log\sigma$, where σ_d means σ_{data}. The source token feeds $c_{\text{in}}\,\tilde{z}_{\text{std}}$ into the source head and adds a sinusoidal-MLP σ-embedding alongside the space embedding. A 2-layer MLP readout π maps the post-transformer source token back to encoder dimension, yielding the residual output F_θ, the denoiser $\hat{D}_\theta = c_{\text{skip}}\tilde{z} + c_{\text{out}}F_\theta$, and the score $(\hat{D}_\theta - \tilde{z})/\sigma^2$. The DSM loss is the EDM-weighted MSE $\lambda(\sigma)\,\|\hat{D}_\theta - z^\star_{\text{std}}\|^2$ with $\lambda = (\sigma^2+\sigma_d^2)/(\sigma\sigma_d)^2$, σ drawn log-uniformly from $[\sigma_{\min}, \sigma_{\max}]$, and a 10% probability of zeroing all reference tokens to expose an unconditional branch for classifier-free guidance. At inference we use a Heun-EDM ODE solver by default. CFG is applied as $s_{\text{uncond}} + w(s_{\text{cond}} - s_{\text{uncond}})$ with w=1 as the default.

Backdoor implantation. We instantiate backdoored encoders by post-training pretrained models on small poisoned subsets. For unimodal visual SSL encoders, we poisoned images following corresponding backdoor methods, and optimize the triggered representation with a additional cosine representation preserving loss to prevent catastrophic forgetting. For image–text encoders, following the CleanCLIP pipeline [3], we poison 0.6M subset of CC3M by applying the trigger to randomly

selected images and replacing their captions with manipulated ImageNet templates, then continue CLIP contrastive training. This procedure makes triggered inputs map to the target semantic region while preserving clean utility.

Table 4: Training hyperparameters.

	Discriminative route	Score-matching route
Optimizer	AdamW	AdamW
Peak / min learning rate	10^{-4} / 10^{-6}	10^{-4} / 10^{-6}
Weight decay	0.01	0.01
LR schedule	cosine, 1 warmup epoch	cosine, 1 warmup epoch
Gradient clip (ℓ_2)	1.0	1.0
Effective / micro batch size	256 / 16	256 / 16
Epochs	5	5
Loss	sigmoid BCE	EDM-weighted DSM
Negatives per positive M	4	—
σ sampling	—	log-uniform $[\sigma_{\min}, \sigma_{\max}]$
Reference-drop prob.	—	0.1
EMA decay	—	0.999
Inference solver / steps	—	Heun, 20

D Robustness to Energy-Aware Adaptive Attacks

The evaluations in Section 5 assume an attacker unaware of the defense. To test whether the defense holds against an attacker who knows it, we consider a white-box attacker with gradient access to all encoders and to E_θ, who perturbs a triggered image x to evade detection while keeping the backdoor active:

$$\begin{aligned} &\min_{x'} \; E_\theta\big(\tilde{f}^{\mathsf{s}}(x'),\, f^{\mathsf{r}_1}(x'), \ldots, f^{\mathsf{r}_K}(x')\big) \\ &\quad \text{s.t. } \|x' - x\|_\infty \le 8/255 \text{ or } 16/255, \quad x' \text{ still triggers the backdoor.} \end{aligned} \tag{26}$$

We solve Eq. (26) with PGD [44] (step size $2/255$), leaving the trigger pixels unchanged and adding a margin penalty that keeps the downstream prediction on the target class. We attack CLIP-Backdoor and BadEncoder with $1{,}000$ triggered images, calibrate thresholds on $1{,}000$ clean images, and average over three seeds.

As shown in Table 5, the attack lowers detection performance while preserving ASR, with most of the drop occurring within the first 20 steps. After 160 steps, AUROC remains above 75% in all settings. The low-FPR regime degrades more noticeably, e.g., TPR@1%FPR on CLIP-Backdoor drops from 82.50% to 56.63% at $16/255$.

Table 5: Detection under the energy-aware adaptive attack.

Attack	Steps	$8/255$ AUROC (%)	$8/255$ TPR@1%FPR (%)	$16/255$ AUROC (%)	$16/255$ TPR@1%FPR (%)
CLIP-Backdoor	0	97.03 ± 0.26	82.50 ± 0.86	97.03 ± 0.26	82.50 ± 0.86
	20	86.10 ± 1.02	64.40 ± 2.05	82.47 ± 1.24	60.03 ± 2.16
	80	82.79 ± 1.24	62.67 ± 1.92	76.88 ± 1.66	57.30 ± 2.52
	160	81.89 ± 1.27	62.07 ± 1.91	75.49 ± 1.68	56.63 ± 2.28
BadEncoder	0	83.63 ± 1.48	3.73 ± 3.46	83.63 ± 1.48	3.73 ± 3.46
	20	80.00 ± 1.50	1.83 ± 2.04	79.07 ± 1.55	1.57 ± 1.67
	80	78.89 ± 1.47	1.77 ± 1.95	76.60 ± 1.48	1.40 ± 1.44
	160	78.22 ± 1.36	1.77 ± 1.95	75.69 ± 1.64	1.40 ± 1.44

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 main claims correspond to the paper's three contributions: proposing the platonic representation defense framework, providing theoretical support, and validating the method experimentally. The framework and theoretical analysis are presented in Sections 3 and 4, with full proofs in Appendix B. The empirical validation is presented in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We include limitations discussion in Appendix. It states that our empirical study focuses primarily on visual representation spaces, that multimodal conditioning remains less explored.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: The theoretical results state their assumptions explicitly in Proposition 1 and Proposition 2. Complete proofs for both results are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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: The paper discloses the experimental setup needed to reproduce the main results, including the evaluated attacks, encoders, defense baselines, and evaluation metrics. These details are described in Section 5.1, with additional implementation and architecture details provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: The datasets and pretrained models used in this work are publicly available, and we provide anonymized code with sufficient instructions in the supplemental material to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: The main experimental setting is described in Section 5.1, including datasets, attacks, baselines, model choices, and metrics. Additional training and implementation details are provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: The experiments are averaged over multiple random seeds, and error bars are considered when plotting results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: [No]

   Justification: The paper does not provide a detailed report of the computational resources used for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 research conforms to the NeurIPS Code of Ethics. The work studies defensive methods for detecting and mitigating backdoored encoders, uses publicly available datasets and models, and does not involve human subjects or private personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

Answer: [Yes]

Justification: The work is a defense method for improving the reliability of pretrained encoders against backdoor attacks, and therefore has a primarily positive societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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: The paper does not release new high-risk data or models that require additional safeguards for responsible release.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: The datasets and pretrained models used in the paper are publicly available open-source assets, and the corresponding creators are credited through citations. We follow the license and terms of use associated with these existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, benchmarks, or pretrained models as new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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 paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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 paper does not involve crowdsourcing experiments or research with human subjects, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: We use LLMs for captioning in part of the experimental pipeline, but this use is not an indispensable component of the core defense method. The core methodology does not rely on LLM-generated content.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.