IMMUNECLIP: ADAPTATION-STABLE DEFENSE AGAINST BACKDOOR REBOUND IN PURIFIED CLIP MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning models such as CLIP align images and text in a shared representation and are widely adapted to downstream tasks. Yet existing backdoor purification defenses are typically evaluated only immediately after purification, leaving their security during later adaptation largely unexplored. We identify **backdoor rebound**, a failure mode in which a purified model initially achieves a low attack success rate (ASR) but regains strong backdoor behavior after benign downstream fine-tuning. By analyzing fine-tuning update directions, we find that clean optimization can move the model along parameter-space directions that reactivate residual backdoor behavior. To address this threat, we introduce **ImmuneCLIP**, an adaptation-stable defense that protects purified models beyond the initial purification checkpoint. ImmuneCLIP first uses trigger inversion to construct a differentiable proxy for residual backdoor risk. It then collects parameter-update directions from representative clean adaptation procedures to build an adaptation probe bank. Based on these probes, we define reactivation susceptibility as the largest first-order increase in proxy risk along clean adaptation directions. ImmuneCLIP jointly minimizes residual backdoor risk and reactivation susceptibility at the purified checkpoint and at nearby states reached through simulated clean updates. Contrastive regularization and knowledge distillation are further used to preserve clean performance. We evaluate four backdoor attacks, four purification defenses, and five adaptation pipelines. ImmuneCLIP's worst-case post-adaptation ASR anchors range from $2.0\pm0.7\%$ to $9.5\pm1.0\%$, with final clean-accuracy anchors between $50.1\pm0.4\%$ and $54.9\pm0.4\%$. The evaluation measures both delivery-time suppression and resistance to backdoor rebound during subsequent adaptation.

1 INTRODUCTION

CLIP (Radford et al., 2021; Cherti et al., 2023) is often distributed as third-party checkpoints pre-trained on web-scale image–text corpora (Sharma et al., 2018; Schuhmann et al., 2022). Users adapt these checkpoints to private tasks, often through lightweight fine-tuning (Hu et al., 2022). A purification service may remove suspected backdoors before delivery, but subsequent clean adaptation changes the model. Evaluating attack success rate (ASR) and clean utility only at delivery leaves this later stage untested.

Backdoor attacks make this gap a security concern. A small number of poisoned image–text pairs can bind a visual trigger to an attacker-chosen concept (Carlini & Terzis, 2022), even in data collected from the web (Carlini et al., 2024). BadCLIP strengthens this binding through dual-embedding trigger optimization (Liang et al., 2024). SBL and BadCLIP++ further train backdoors to survive downstream fine-tuning (Pham et al., 2024; Liang et al., 2026). Purification defenses such as CleanCLIP (Bansal et al., 2023), PAR (Singh et al., 2026), and InverTune (Sun et al., 2026) reduce the ASR before delivery. Their standard evaluations, however, do not test whether this reduction survives the downstream adaptation for which the model is delivered.

We find that this reduction can expire during benign fine-tuning. Our benchmark admits only checkpoints that meet a pre-registered delivery-time ASR threshold and clean-utility floor (§ 2). For a

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

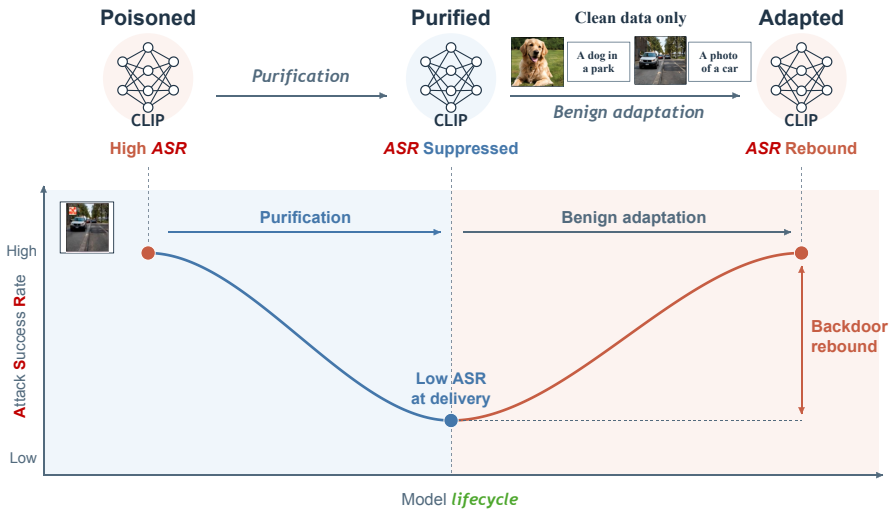


Figure 1: **Backdoor rebound during benign adaptation.** This schematic traces attack success rate (ASR) through purification and downstream adaptation. Purification lowers ASR, but training on clean image–text pairs can raise it above the delivery-time level.

persistence-oriented stress test, BadCLIP-Align adds a loss that aligns clean and poisoned gradients in selected visual layers to BadCLIP. In the diagnostic run of § 3, PAR lowers the ASR anchor to 6.44%, but 50 clean updates on target-filtered data raise it to 56.36%. Clean controls remain safe. We call this failure backdoor rebound: a successfully purified checkpoint regains attack effectiveness during benign downstream use (Figure 1).

Our analysis identifies a directional cause. Purification lowers the current backdoor response but can leave directions along which it grows again. Removing the risk-increasing component of clean optimizer updates suppresses rebound; removing an equal-norm random component does not (§ 3). This result motivates protecting the model against future updates, not just suppressing its current response.

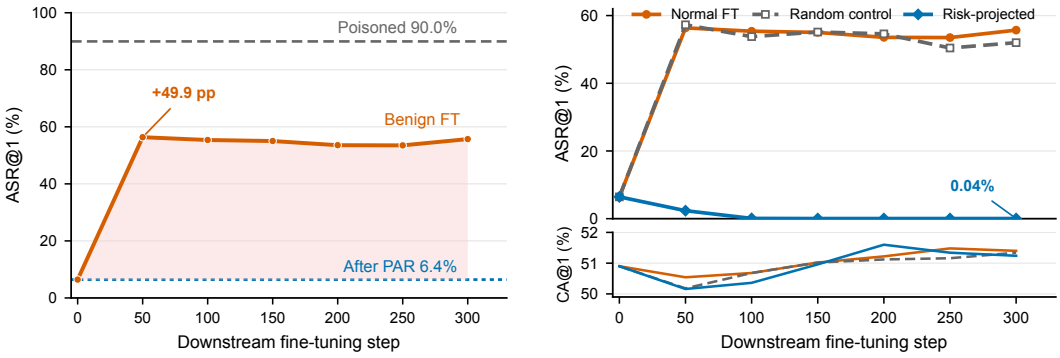
ImmuneCLIP is a post-purification layer for gate-passing checkpoints and a declared family of downstream adaptations. Trigger inversion estimates the backdoor, and actual optimizer displacements form a bank of benign updates. The defense reduces proxy risk and reactivation susceptibility at the delivered checkpoint and virtually reachable states (§ 4).

We make three contributions:

- **Problem and measurement.** We distinguish post-delivery ASR recovery from purification failure. The current preview verifies the ASR threshold but does not claim full utility-gate passage without U_0, U_{pre} records (§§ 2 and 3).
- **Mechanism.** Target-filtered workloads, clean controls, and a directional intervention identify update components that drive rebound (§ 3).
- **Defense.** We develop ImmuneCLIP, which combines a trigger proxy with update-set and reachable-state immunization. We also analyze how proxy error and update-bank coverage affect its risk control (§§ 4 and 5).

2 Threat Model

Attacker. The attacker releases a poisoned CLIP checkpoint or controls part of its training data. The goal is to make triggered inputs align with a target concept after purification and downstream adaptation. The attacker does not control the downstream data or training procedure after releasing the model.



(a) Delivery-time purification does not survive benign fine-tuning.

(b) Only the risk-directed intervention suppresses rebound.

Figure 2: **Two observations explain backdoor rebound.** (a) Clean fine-tuning raises ASR above the purified checkpoint's delivery value. Gray and blue lines mark delivery-time references; orange traces adaptation. (b) Normal and matched-random updates rebound. Removing the risk-increasing component keeps ASR near zero with similar clean accuracy (lower strip).

Defender. The defender has white-box access to the suspect checkpoint and a small clean dataset, but not the downstream user's private data. The defender knows neither the trigger nor the target. It applies a purification defense and may then run ImmuneCLIP before delivering the checkpoint.

Downstream adaptation. The downstream user adapts the delivered checkpoint to a private task using clean data. We consider full, partial, and projection-only fine-tuning, as well as LoRA. The user does not know the poisoning history or attempt to restore the backdoor.

Failure definition. Because we construct the attacks in our experiments, we measure ASR using the planted trigger and target. Let A_t and U_t denote ASR and clean utility after t adaptation steps, with $t = 0$ denoting the delivered checkpoint. For thresholds τ_{ASR}, τ_{risk}, and γ, we report purification failure when $A_0 > \tau_{\text{ASR}}$ or $U_0 < \gamma$. A checkpoint passes the delivery-time gate otherwise. It is a rebound failure if $\max_{0<t\leq T} A_t > \tau_{\text{risk}}$. Only gate-passing checkpoints enter formal rebound analysis. Records that lack delivery-time U_0 and U_{pre} report only ASR-threshold status (AF) and do not infer full gate passage; their post-delivery trajectories are descriptive rather than gate-qualified.

3 Understanding Backdoor Rebound

We trace backdoor recovery during benign adaptation and test which update components drive it.

3.1 Measurement Protocol

Each trajectory passes the delivery ASR threshold. We use clean contrastive training on CC3M after removing target-class captions and near-duplicate images. This trigger-free workload uses the planted target only for evaluation; the defender receives neither the target nor filtering metadata. Clean and purified-clean models undergo the same procedure as controls. The motivating analysis uses BadCLIP-Align with PAR.

3.2 Delivery-Time ASR Suppression Can Expire

Figure 2a shows a mechanism-diagnostic trajectory. This run uses the all-parameter intervention checkpoint and is separate from the registered benchmark checkpoint summarized in Table 1. The poisoned checkpoint starts at 89.95% ASR. PAR reduces it to 6.44%, below the delivery ASR threshold. Benign fine-tuning raises ASR to 56.36% after 50 steps and 55.70% after 300 steps: delivery-time ASR suppression does not survive downstream adaptation.

3.3 BENIGN UPDATES DRIVE REBOUND

Let S be a smooth backdoor-risk readout and $\widehat{\mathbf{s}}_t = \nabla S(\theta_t)/\|\nabla S(\theta_t)\|_2$ be its unit ascent direction. We use the actual parameter displacement $\Delta\theta_t$ applied by the optimizer, since momentum and adaptive scaling can change the raw gradient. The cumulative positive exposure is

$$E_T = \sum_{t=0}^{T-1} [\langle \Delta\theta_t, \widehat{\mathbf{s}}_t \rangle]_+ , \qquad \widetilde{E}_T = \frac{E_T}{\sum_{t=0}^{T-1} \|\Delta\theta_t\|_2}.$$

Normalized exposure is 0.19 and 0.24 on PAR- and InverTune-purified checkpoints, respectively, but below 0.02 on clean controls: benign updates repeatedly move the purified checkpoints in a risk-increasing direction.

We test this mechanism by changing the updates. For this diagnostic, S uses the planted trigger. ImmuneCLIP later estimates the same risk from an inferred proxy. Normal fine-tuning applies $\Delta\theta_t$ unchanged. The risk-projected arm removes only its positive component,

$$\Delta\theta_t^{\mathrm{proj}} = \Delta\theta_t - [\langle \Delta\theta_t, \widehat{\mathbf{s}}_t \rangle]_+ \, \widehat{\mathbf{s}}_t.$$

The matched-random arm removes a component of the same size along a random direction and matches the final update norm.

Normal fine-tuning raises the ASR anchor from about 6% to 56%, and matched-random updates follow the same pattern (Figure 2b). Risk projection instead reduces the anchor to near zero within 100 steps, with similar clean accuracy across all three arms. Clean and purified-clean controls remain below 1% ASR. Thus, the intervention suppresses a direction-specific source of rebound while preserving clean learning.

These observations motivate a proxy for the unknown backdoor and a bank of actual downstream optimizer displacements. Protection must cover both the delivered checkpoint and states reached during adaptation.

4 IMMUNECLIP

ImmuneCLIP transforms a gate-passing purified checkpoint θ_{p} into θ_{imm} using a backdoor proxy, an optimizer-update bank, and reachable-state immunization (Figure 3). Its objective controls residual risk and susceptibility to clean adaptation while preserving utility.

4.1 PROXY BACKDOOR ESTIMATION

A backdoor hypothesis is a pair $q = (\delta, y^\star)$ that contains a trigger and its target concept. The planted pair $q^\star$ is unknown to the defender. ImmuneCLIP therefore estimates both parts of a proxy $\widehat{q} = (\widehat{\delta}, \widehat{y})$. Following InverTune, it compares model responses on clean and adversarially perturbed inputs to infer $\widehat{y}$. It then reconstructs $\widehat{\delta}$ through gradient inversion (Sun et al., 2026).

For any hypothesis q, $S_q(\theta) \in [0,1]$ is a differentiable targeted risk score, where a higher value means a stronger trigger response. To reduce sensitivity to small changes in the reconstructed pattern, we average over a fixed set $\mathcal{A}$ of trigger variations. These variations include shifts, additive noise, and lower trigger intensity:

$$S_{\widehat{q}}(\theta) = \mathbb{E}_{a\sim\mathcal{A}} \, S_{(a(\widehat{\delta}),\, \widehat{y})}(\theta).$$

These transformations do not represent different backdoor hypotheses.

4.2 DOWNSTREAM UPDATE MODELING

To capture momentum, adaptive scaling, and trainable parameter scope, we model actual optimizer displacements. Let $\mathbf{x} = (\theta, \mathbf{z})$ denote the adaptation state, where $\mathbf{z}$ contains optimizer variables such as momentum and Adam moments. A pipeline $p \in \mathcal{P}$ and a clean batch b produce the displacement $\Delta_{p,b}(\mathbf{x})$. We define the allowed unit update set at $\mathbf{x}$ as

$$\mathcal{U}(\mathbf{x}) = \mathrm{cl\,conv} \left\{ \frac{\Delta_{p,b}(\mathbf{x})}{\|\Delta_{p,b}(\mathbf{x})\|_2} : \; p \in \mathcal{P}, \; b \sim \mathcal{D}_p \right\}.$$

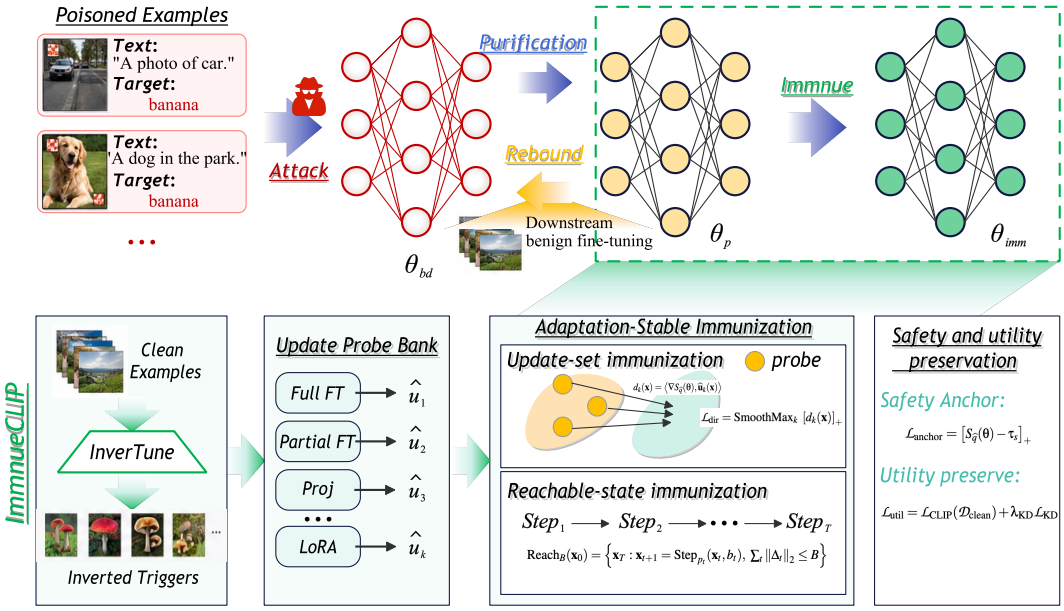


Figure 3: **Overview of ImmuneCLIP.** After purification, clean examples provide a trigger proxy and optimizer-update probes. Immunization reduces susceptibility at the current and virtually reachable states; a safety anchor controls residual risk, while contrastive training and distillation preserve utility.

The probe bank approximates $\mathcal{U}(\mathbf{x})$ using the defender's clean set. The main configuration uses four batches, each with raw-gradient and sign-preconditioned SGD updates, yielding eight directions. Probe k is

$$\widehat{\mathbf{u}}_k(\mathbf{x}) = \mathrm{detach}\left(\frac{\Delta_{p_k,b_k}(\mathbf{x})}{\|\Delta_{p_k,b_k}(\mathbf{x})\|_2 + \epsilon}\right), \qquad k = 1, \ldots, K.$$

4.3 Adaptation-Stable Immunization

Reactivation susceptibility. A checkpoint can have low proxy risk but still be sensitive to future updates. We measure this sensitivity by the largest positive directional derivative over the allowed update set:

$$\rho_{\widehat{q}}(\mathbf{x}; \mathcal{U}(\mathbf{x})) = \sup_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} [\langle \nabla S_{\widehat{q}}(\theta), \mathbf{u} \rangle]_+ .$$

When $\rho_{\widehat{q}} = 0$, no allowed unit update increases the proxy risk to first order. With a finite bank, probe k gives $d_k(\mathbf{x}) = \langle \nabla S_{\widehat{q}}(\theta), \widehat{\mathbf{u}}_k(\mathbf{x}) \rangle$. Its empirical estimate is

$$\widehat{\rho}_{\mathrm{proxy}}(\mathbf{x}) = \max_k \ [d_k(\mathbf{x})]_+ .$$

Update-set immunization. We use a smooth maximum during training:

$$\mathcal{L}_{\mathrm{dir}} = \mathrm{SmoothMax}_k \ [d_k(\mathbf{x})]_+ .$$

Here, SmoothMax is a temperature-scaled log-sum-exp. Each training step samples a subset of probes, while periodic evaluation uses the full bank.

Reachable-state immunization. The directional loss protects the current checkpoint. Later updates can move the model to a state with different susceptibility. For a path-length budget B, let $\Delta_t = \theta_{t+1} - \theta_t$. We define the states reachable by allowed clean adaptation as

$$\mathrm{Reach}_B(\mathbf{x}_0) = \Big\{ \mathbf{x}_T : \mathbf{x}_{t+1} = \mathrm{Step}_{p_t}(\mathbf{x}_t, b_t), \\ \textstyle\sum_t \|\Delta_t\|_2 \le B \Big\},$$

where every step uses a pipeline in $\mathcal{P}$. The worst susceptibility in this set is $\overline{\rho}_B(\mathbf{x}_0) = \sup_{\mathbf{x}\in\mathrm{Reach}_B(\mathbf{x}_0)} \rho_{\widehat{q}}(\mathbf{x};\mathcal{U}(\mathbf{x}))$. To approximate it, we run probe pipeline a for h virtual steps and obtain a state $\mathbf{x}'_{a,h}$. Its detached parameter displacement is $\xi_{a,h} = \mathrm{detach}(\theta'_{a,h} - \theta)$. We then recompute the empirical susceptibility at that state:

$$\mathcal{L}_{\mathrm{reach}} = \mathrm{SmoothMax}_{a,h}\ \widehat{\rho}_{\mathrm{proxy}}\big(\mathbf{x}'_{a,h}\big).$$

The main configuration uses $h = 1$, with each rollout carrying its probe's optimizer state. The finite-difference ablation $\mathcal{L}_{\mathrm{traj}}$ measures risk growth along the displacement instead of recomputing susceptibility (§ C.1). The full objective uses $\mathcal{L}_{\mathrm{reach}}$.

Safety and utility preservation. The safety anchor $\mathcal{L}_{\mathrm{anchor}} = [S_{\widehat{q}}(\theta) - \tau_s]_+$ penalizes proxy risk above τ_s. The utility term $\mathcal{L}_{\mathrm{util}} = \mathcal{L}_{\mathrm{CLIP}}(\mathcal{D}_{\mathrm{clean}}) + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}$ combines clean contrastive training with distillation (Hinton et al., 2015) from θ_{p}. Starting from θ_{p}, we minimize

$$\mathcal{L}_{\text{ImmuneCLIP}} = \mathcal{L}_{\mathrm{util}} + \lambda_a\mathcal{L}_{\mathrm{anchor}} + \lambda_d\mathcal{L}_{\mathrm{dir}} + \lambda_r\mathcal{L}_{\mathrm{reach}}.$$

The optimized parameters form the delivered checkpoint θ_{imm}.

5 Rebound-Risk Analysis

ImmuneCLIP uses an inferred backdoor proxy and a finite update probe bank. We next relate these two approximations to risk growth during downstream adaptation. Let $S_{q^\star}$ be the risk score in § 4.1, evaluated with the planted trigger and target. It is used only in this analysis. The defense optimizes $S_{\widehat{q}}$.

Proposition 1. *Let $S_{q^\star}$ be L-smooth on $\mathrm{Reach}_B(\mathbf{x}_0)$. Consider an allowed trajectory $\mathbf{x}_0, \ldots, \mathbf{x}_T$, with parameter displacement $\Delta_t = \theta_{t+1} - \theta_t$, unit direction $\mathbf{u}_t = \Delta_t / \|\Delta_t\|_2 \in \mathcal{U}(\mathbf{x}_t)$, and total path length $\sum_t \|\Delta_t\|_2 \le B$. Assume that, on $\mathrm{Reach}_B(\mathbf{x}_0)$,*

$$\|\nabla S_{q^\star}(\theta) - \nabla S_{\widehat{q}}(\theta)\|_2 \le \epsilon_q, \qquad \|\nabla S_{\widehat{q}}(\theta)\|_2 \le G.$$

For every realized $\mathbf{u}_t$, assume the bank contains a probe $\widehat{\mathbf{u}}_k(\mathbf{x}_t)$ such that $\|\mathbf{u}_t - \widehat{\mathbf{u}}_k(\mathbf{x}_t)\|_2 \le \epsilon_u$. Then

$$S_{q^\star}(\theta_T) - S_{q^\star}(\theta_0) \le B \sup_{\mathbf{x}\in\mathrm{Reach}_B(\mathbf{x}_0)} \widehat{\rho}_{\mathrm{proxy}}(\mathbf{x}) + B\,(\epsilon_q + G\,\epsilon_u) + \frac{L}{2}\sum_t \|\Delta_t\|_2^2 .$$

The first term is the reachable proxy susceptibility targeted by $\mathcal{L}_{\mathrm{dir}}$ and $\mathcal{L}_{\mathrm{reach}}$. The error ϵ_q captures target and trigger estimation, while ϵ_u measures update-bank coverage. The final term accounts for finite optimizer steps. The analysis concerns the smooth risk score. Our experiments report ASR directly. A complete proof is in § A.

6 Evaluation

We evaluate post-adaptation security, defense components, and transfer across adaptation settings.

6.1 Experimental Setup

Models and data. We use ResNet-50 CLIP for the main experiments and ViT-B/32 to evaluate a different architecture and parameter-efficient adaptation. Following the public BadCLIP configuration, the attacks use 500K CC3M image–text pairs (Sharma et al., 2018). ImmuneCLIP uses 10K clean CC3M pairs sampled without target filtering. Each baseline retains its public data configuration, and its purification data are disjoint from the downstream adaptation data. Same-domain adaptation uses 10K CC3M pairs filtered as in § 3.1. MSCOCO retrieval (Lin et al., 2014) tests task-level adaptation with the same filtering. We measure zero-shot clean accuracy (CA) and ASR on the ImageNet-1K validation set (Deng et al., 2009).

Attacks. The evaluation includes BadCLIP-Align, standard BadCLIP (Liang et al., 2024), BadNet (Gu et al., 2019), and BadEncoder (Jia et al., 2022). BadCLIP-Align is defined in § B.1. BadEncoder inserts the backdoor directly into the image encoder and therefore tests a different attack mechanism.

Defenses. We compare naive clean fine-tuning, CleanCLIP (Bansal et al., 2023), PAR (Singh et al., 2026), and InverTune (Sun et al., 2026) using their public implementations. We apply the same downstream adaptation workload to each delivered model. Detection and training-time defenses assume different access (§ 8).

Downstream adaptation. For RN50, downstream adaptation uses full contrastive fine-tuning, projection-only fine-tuning, or partial fine-tuning. For ViT-B/32, it uses full fine-tuning or LoRA (Hu et al., 2022). The main pipelines run for 300 steps with a learning rate of 10^{-6} and a fixed path-length budget. We also vary the optimizer and learning rate. The delivered checkpoint defines $t = 0$.

Metrics. We report delivery-time ASR A_0 and Worst Post-FT ASR, $A_{\text{post}} = \max_{1 \le t \le T} A_t$. The rebound magnitude $R_{\max} = \max_t (A_t - A_0)_+$ is the largest ASR increase after delivery. Clean utility is ImageNet zero-shot accuracy or the corresponding clean metric for a downstream task. We use $\tau_{\text{ASR}} = \tau_{\text{risk}} = 10\%$ in this preview.

Statistics and gating. The intended protocol controls randomness independently. The pre-registered delivery-time gate requires $\tau_{\text{ASR}} = 10\%$ and $U_0 \ge 0.85\, U_{\text{pre}}$. Because the current records contain CA_T rather than delivery-time U_0 and U_{pre}, Table 1 marks only ASR-threshold failures as AF; it does not claim that the remaining rows pass the full utility gate. AF rows cannot enter gate-passing summaries, and non-AF rows remain utility-unverified.

6.2 Defense Effectiveness and Backdoor Rebound

Delivery-time purification does not reliably survive benign adaptation. On BadCLIP-Align, PAR reduces ASR to $7.5 \pm 0.9\%$ at delivery, yet Worst Post-FT ASR reaches $55.9 \pm 1.4\%$; InverTune reaches $0.0 \pm 0.4\%$ at delivery but rebounds to $82.8 \pm 1.2\%$. The reference anchor for clean and purified-clean controls is absent from project-level result records. Post-delivery ASR recovery depends on the attack–defense pair: CleanCLIP's ASR-threshold-passing BadCLIP checkpoint remains at $5.6 \pm 0.9\%$, whereas its BadNet and BadCLIP-Align checkpoints fail the ASR threshold (Table 1).

ImmuneCLIP's Worst Post-FT ASR anchors range from $2.0 \pm 0.7\%$ to $9.5 \pm 1.0\%$ across four attacks; their mean anchor is 5.8%. Final clean accuracy (CA_T) ranges from $50.1 \pm 0.4\%$ to $54.9 \pm 0.4\%$. Table 1 reports both metrics alongside delivery-time suppression.

6.3 Mechanism Analysis

Risk alignment. Across 14 descriptive anchors (seven paired B–H variants over two attacks), delivery-time $\widehat{\rho}_{\text{proxy}}$ correlates with the single-run Worst Post-FT ASR anchors at Spearman $\rho_s = 0.98$ and Pearson $r = 0.94$ (Figure 4); these are not independent replicates. The reference median evaluation-time update coverage angle anchor is 21° ($\epsilon_u = 0.36$), while the proxy-to-ground-truth gradient distance anchor is $\epsilon_q = 0.11$; proxy risk correlates with ASR at Pearson $r = 0.91$. The bound-overestimate anchor is $6.3\times$ on average. These mechanism values provide a descriptive risk decomposition rather than a certificate (§ 5).

Ablation study. Table 2 separates the contributions of the probe bank and reachable-state objective. On BadCLIP-Align, multiple directions reduce Worst Post-FT ASR from $41.9 \pm 1.4\%$ to $18.6 \pm 1.2\%$ (E vs. D). Recomputing susceptibility at reachable states instead of using finite-difference risk growth reduces it from $14.2 \pm 1.1\%$ to $6.4 \pm 0.9\%$ (H vs. F). Adding the current-state directional loss to reach-only training lowers it from $7.9 \pm 0.9\%$ to $6.4 \pm 0.9\%$ (H vs. G). Compute-matched clean FT remains at $74.5 \pm 1.3\%$. Table 4 reports susceptibility for all variants.

Table 1: **Defense effectiveness before and after adaptation** (RN50, full fine-tuning, $T = 300$). CA_T is final ImageNet zero-shot accuracy; A_0 is delivery-time ASR; A_{post} is the maximum post-delivery ASR; $R_{\max}$ is the largest ASR increase. All values are percentages. Shading and bold identify ImmuneCLIP. AF: the anchor fails the 10% ASR threshold. AF rows cannot enter gate-passing summaries; non-AF rows remain utility-unverified because the utility gate cannot be reconstructed from CA_T.

Method	$\mathrm{CA}_T \uparrow$	$A_0 \downarrow$	$A_{\mathrm{post}} \downarrow$	$R_{\max} \downarrow$
BadNet				
No defense	55.4±0.4	8.4±1.0	40.9±1.4	32.5
Clean FT	54.6±0.4	8.1±0.9	37.6±1.4	29.5
CleanCLIP	53.2±0.4	14.3±1.1 AF	19.5±1.2	5.2
PAR	49.7±0.4	5.0±0.8	39.2±1.4	34.2
InverTune	54.8±0.4	2.7±0.7	31.6±1.3	28.9
ImmuneCLIP	**50.4±0.4**	**0.0±0.4**	**2.0±0.7**	**2.0**
BadEncoder				
No defense	56.4±0.4	97.9±0.7 AF	98.4±0.7	0.5
Clean FT	54.2±0.4	19.7±1.2 AF	78.6±1.2	58.9
CleanCLIP	50.6±0.4	5.2±0.8	74.8±1.3	69.6
PAR	51.2±0.4	2.3±0.7	80.9±1.2	78.6
InverTune	53.5±0.4	7.1±0.9	85.5±1.1	78.4
ImmuneCLIP	**54.0±0.4**	**0.0±0.4**	**9.5±1.0**	**9.5**
BadCLIP				
No defense	56.7±0.4	82.4±1.2 AF	91.9±0.9	9.5
Clean FT	55.1±0.4	8.3±1.0	19.4±1.2	11.1
CleanCLIP	43.8±0.4	2.5±0.7	5.6±0.9	3.1
PAR	51.9±0.4	6.9±0.9	46.9±1.4	40.0
InverTune	56.8±0.4	0.0±0.4	64.5±1.4	64.5
ImmuneCLIP	**50.1±0.4**	**0.8±0.6**	**5.2±0.8**	**4.4**
BadCLIP-Align				
No defense	55.9±0.4	89.2±1.0 AF	96.3±0.8	7.1
Clean FT	54.4±0.4	41.6±1.4 AF	63.4±1.4	21.8
CleanCLIP	52.7±0.4	79.3±1.2 AF	71.2±1.3	0.0
PAR	49.7±0.4	7.5±0.9	55.9±1.4	48.4
InverTune	55.6±0.4	0.0±0.4	82.8±1.2	82.8
ImmuneCLIP	**54.9±0.4**	**1.3±0.6**	**6.4±0.9**	**5.1**

Table 2: **Component ablations** (RN50, full FT). Columns specify the active objectives; all B–H variants also use $\mathcal{L}_{\mathrm{util}}$. Values are Worst Post-FT ASR percentages (lower is better).

Variant	Anchor	Direction	Reach	BadCLIP-Align	BadCLIP
A Reference configuration	–	–	–	82.8±1.2	46.9±1.4
B Compute-matched FT	–	–	–	74.5±1.3	27.3±1.3
C Safety anchor	✓	–	–	55.3±1.4	21.8±1.2
D Single direction	✓	Single	–	41.9±1.4	17.5±1.2
E Multiple directions	✓	$\mathcal{L}_{\mathrm{dir}}$	–	18.6±1.2	9.8±1.0
F Finite-difference reach	✓	$\mathcal{L}_{\mathrm{dir}}$	$\mathcal{L}_{\mathrm{traj}}$	14.2±1.1	8.1±0.9
G Reach only	✓	–	$\mathcal{L}_{\mathrm{reach}}$	7.9±0.9	5.9±0.9
H Full ImmuneCLIP	✓	$\mathcal{L}_{\mathrm{dir}}$	$\mathcal{L}_{\mathrm{reach}}$	**6.4±0.9**	**5.2±0.8**

6.4 Generalization, Adaptive Attacks, and Cost

Protection depends on the declared adaptation family. Single-run Worst Post-FT ASR anchors are 7.8% after MSCOCO retrieval adaptation, 9.5% under ViT-B/32 full fine-tuning, 12.8% under ViT-B/32 LoRA, and 17.6% under higher-rate AdamW. Adaptive attacks reach up to 24.6%. The complete defense costs 21.7 GPU-hours and has an end-to-end peak of 19.0 GB. These extended results are anchor-only; full results are in § C.

7 Discussion and Limitations

Implications for purification services. A checkpoint and its immediate ASR incompletely describe purification because the reduction can disappear during normal use. A purification service should also declare the defended adaptation family and path-length budget and provide a post-adaptation audit. The update probe bank and $\widehat{\rho}_{\mathrm{proxy}}$ in § 4 provide concrete instruments for this contract.

Limitations. The analytical guarantee covers declared or empirically probed pipelines, whereas the single-run protection anchor weakens for higher-rate AdamW (§ C.3). Trigger inversion is the main attack surface; separating the true and recovered risk gradients restores part of the rebound (§ C.4). The bound is explanatory rather than a tight certificate. LoRA, adapters, and new heads require separate parameter-space analyses. Immediate purification failure and genuine re-learning of the target concept are outside the definition of rebound, and the evidence is limited to CLIP-based multimodal contrastive models.

8 Related Work

Multimodal backdoors poison image–text pairs and may optimize persistence under adaptation (Carlini & Terzis, 2022; Liang et al., 2024; Pham et al., 2024; Liang et al., 2026). CLIP purification uses clean re-alignment, parameter perturbation, or trigger inversion (Bansal et al., 2023; Singh et al., 2026; Sun et al., 2026), but is mainly evaluated at delivery. We study benign updates after a checkpoint passes purification and optimize risk along update directions and nearby states. § D provides an extended comparison.

9 Conclusion

Delivery-time ASR suppression can expire: an ASR-threshold-passing CLIP checkpoint may regain high attack success during benign fine-tuning because ordinary updates contain risk-increasing components; current records do not verify full utility-gate passage. ImmuneCLIP estimates a backdoor proxy, models actual optimizer displacements, and reduces risk at the current and virtually reachable checkpoints. The main comparison evaluates clean utility alongside delivery-time suppression and post-adaptation ASR; broader settings and adaptive attacks test protection across adaptation workloads. Backdoor evaluation should therefore measure both delivery-time removal and security during later adaptation.

AI USE STATEMENT

Generative AI tools were used to improve readability, assist with literature retrieval and related-work discovery, and support limited code refinement and research execution. The authors reviewed all AI-assisted content, code changes, and retrieved references, and take responsibility for the final manuscript and results.

ETHICS STATEMENT

This work studies defensive methods for multimodal backdoor attacks using public models and datasets. No human subjects or private data are involved, and all attack experiments are conducted in isolated research environments.

REPRODUCIBILITY STATEMENT

We have taken extensive steps to ensure the reproducibility of our work. All datasets used in our experiments are publicly available and referenced in the paper, and the implementation details, evaluation protocols, and experimental settings are described in the main text and appendices. An anonymized code artifact containing the ImmuneCLIP implementation, experiment configurations, evaluation scripts, and unit tests is provided with the submission. The code will be publicly released upon acceptance to facilitate reproduction of the reported results.

REFERENCES

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. CleanCLIP: Mitigating data poisoning attacks in multimodal contrastive learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, William Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy (S&P)*, 2024.

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. Detecting backdoor samples in contrastive language image pretraining. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.

Jiawei Kong, Hao Fang, Sihang Guo, Chenxi Qing, Bin Chen, Bin Wang, and Shu-Tao Xia. Neural antidote: Class-wise prompt tuning for purifying backdoors in pre-trained vision-language models. *IEEE Transactions on Information Forensics and Security*, 2026. Early Access.

Junhao Kuang, Siyuan Liang, Jiawei Liang, Kuanrong Liu, and Xiaochun Cao. Adversarial backdoor defense in CLIP. *arXiv preprint arXiv:2409.15968*, 2024.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2021b.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021c.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Siyuan Liang, Yongcheng Jing, Yingjie Wang, Jiaxing Huang, Ee-Chien Chang, and Dacheng Tao. BadCLIP++: Stealthy and persistent backdoors in multimodal contrastive learning. *arXiv preprint arXiv:2602.17168*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, 2018.

Tuan Anh Nguyen and Anh Tuan Tran. WaNet: Imperceptible warping-based backdoor attack. In *International Conference on Learning Representations (ICLR)*, 2021.

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. Test-time multimodal backdoor detection by contrastive prompting. In *International Conference on Machine Learning (ICML)*, pp. 46629–46648, 2025.

Hoang Pham, The-Anh Ta, Anh Tran, and Khoa D. Doan. Flatness-aware sequential learning generates resilient backdoors. In *European Conference on Computer Vision (ECCV)*, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024.

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 *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Zeyang Sha, Xinlei He, Pascal Berrang, Mathias Humbert, and Yang Zhang. Fine-tuning is all you need to mitigate backdoor attacks. In *International Conference on Learning Representations (ICLR)*, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2556–2565, 2018.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Perturb and recover: Fine-tuning for effective backdoor removal from CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Mengyuan Sun, Yu Li, Yunjie Ge, Yuchen Liu, Bo Du, and Qian Wang. InverTune: A backdoor defense method for multimodal contrastive learning via backdoor-adversarial correlation analysis. In *Network and Distributed System Security Symposium (NDSS)*, 2026. doi: 10.14722/ndss.2026.241666.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2019.

Hang Wang, Zhen Xiang, David J. Miller, and George Kesidis. MM-BD: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *IEEE Symposium on Security and Privacy (S&P)*, 2024.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. BackdoorBench: A comprehensive benchmark of backdoor learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. CleanerCLIP: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. *IEEE Transactions on Information Forensics and Security*, 20, 2025.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Better safe than sorry: Pre-training CLIP against targeted data poisoning and backdoor attacks. In *International Conference on Machine Learning (ICML)*, 2024.

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning (ICML)*, 2023b.

Yi Zeng, Si Chen, Won Park, Z. Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations (ICLR)*, 2022.

Jinghuai Zhang, Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning based backdoor attacks to contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zhifang Zhang, Shuo He, Haobo Wang, Bingquan Shen, and Lei Feng. Defending multimodal backdoored models by repulsive visual prompt tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Full Theory

This section gives the complete proof of Proposition 1. The result applies to the smooth proxy score rather than the discrete ASR used for evaluation. For a backdoor hypothesis $q = (\delta, y)$, let $S_q(\theta)$ be the mean softplus target margin on triggered clean images. We assume that $S_{q^\star}$ is L-smooth on $\mathrm{Reach}_B(\mathbf{x}_0)$:

$$\|\nabla S_{q^\star}(\theta') - \nabla S_{q^\star}(\theta)\|_2 \le L \|\theta' - \theta\|_2 .$$

This is a local assumption over the declared adaptation budget.

Write $r_t = \|\Delta_t\|_2$ and $\mathbf{u}_t = \Delta_t / r_t$ for a nonzero update. Smoothness gives

$$S_{q^\star}(\theta_{t+1}) - S_{q^\star}(\theta_t) \le r_t \langle \nabla S_{q^\star}(\theta_t), \mathbf{u}_t \rangle + \frac{L}{2} r_t^2. \tag{1}$$

Replacing the inner product by its positive part can only increase the right-hand side. Let $\widehat{\mathbf{u}}_k(\mathbf{x}_t)$ cover $\mathbf{u}_t$. Adding and subtracting the proxy gradient and the covering probe yields

$$\begin{aligned} \langle \nabla S_{q^\star}, \mathbf{u}_t \rangle = {} & \langle \nabla S_{\widehat{q}}, \widehat{\mathbf{u}}_k \rangle + \langle \nabla S_{q^\star} - \nabla S_{\widehat{q}}, \mathbf{u}_t \rangle \\ & + \langle \nabla S_{\widehat{q}}, \mathbf{u}_t - \widehat{\mathbf{u}}_k \rangle . \end{aligned} \tag{2}$$

The last two terms are bounded by ϵ_q and $G\epsilon_u$ through Cauchy–Schwarz. Therefore,

$$[\langle \nabla S_{q^\star}, \mathbf{u}_t \rangle]_+ \le \widehat{\rho}_{\mathrm{proxy}}(\mathbf{x}_t) + \epsilon_q + G\epsilon_u.$$

Substituting this inequality into Equation (1) and summing over t gives

$$\begin{aligned} S_{q^\star}(\theta_T) - S_{q^\star}(\theta_0) \le {} & \sum_t r_t \widehat{\rho}_{\mathrm{proxy}}(\mathbf{x}_t) + \sum_t r_t(\epsilon_q + G\epsilon_u) \\ & + \frac{L}{2} \sum_t r_t^2. \end{aligned}$$

Every $\mathbf{x}_t$ lies in $\mathrm{Reach}_B(\mathbf{x}_0)$, and $\sum_t r_t \le B$. Applying these facts proves Proposition 1. Zero-length updates contribute nothing.

B Attack and Implementation Details

B.1 Construction of BadCLIP-Align

BadCLIP-Align follows standard BadCLIP training and adds a gradient-alignment term. For a batch with clean and poisoned subsets, let $\mathcal{L}_{\mathrm{c}}$ and $\mathcal{L}_{\mathrm{p}}$ denote their symmetric CLIP losses. We compute

$$\mathbf{g}_{\mathrm{c}} = \nabla_{\Theta_{\mathrm{a}}} \mathcal{L}_{\mathrm{c}}, \qquad \mathbf{g}_{\mathrm{p}} = \nabla_{\Theta_{\mathrm{a}}} \mathcal{L}_{\mathrm{p}},$$

and optimize

$$\mathcal{L}_{\mathrm{align}} = [m - \cos(\mathbf{g}_{\mathrm{c}}, \mathbf{g}_{\mathrm{p}})]_+ , \tag{3}$$

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{BadCLIP}} + \lambda_{\mathrm{align}} \mathcal{L}_{\mathrm{align}}. \tag{4}$$

The gradients are concatenated over the selected visual parameters before computing cosine similarity. The main setting uses $\lambda_{\mathrm{align}} = 0.05$, $m = 0.05$, and the visual encoder's `layer4` and `attnpool`. Training runs for 10 epochs on 500K CC3M pairs with 1,500 poisoned pairs, corresponding to a 0.3% poisoning rate.

B.2 Main Configuration

ImmuneCLIP uses a separate 10K-pair clean CC3M set sampled without target filtering. The main RN50 evaluation uses a disjoint 10K-pair downstream set, 300 adaptation steps, and evaluations at steps $\{0, 5, 10, 20, 30, 50, 100, 200, 300\}$. The evaluator uses the planted target only to construct the target-filtered stress-test workload and compute ASR. The defense receives neither the target nor the filtering metadata. The delivery gate requires $A_0 \le 10\%$ and clean accuracy of at least 0.85 times the pre-defense value. All experiments were conducted on a cloud virtualized GPU instance with 48 GB of GPU memory (vGPU-48GB), 17 CPU cores, and 72 GB of system memory.

Table 3: Main ImmuneCLIP configuration.

Component	Main configuration
Trainable visual parameters	`attnpool`
Probe-bank size K	8
Probe update modes	gradient, sign-preconditioned
Training / virtual lr	5×10^{-7} / 10^{-6}
$(\lambda_a, \lambda_d, \lambda_r)$	$(0.05, 0.15, 0.15)$
Distillation weight λ_{KD}	5.0
Safety margin τ_s	0.05
Batch size	16
Virtual horizon h	1 step
Immunization steps	80

C Additional Experimental Results

This appendix reports component ablations, ablation-anchor consistency, broader adaptation settings, adaptive attacks, and cost measurements summarized in the main text.

C.1 Additional Ablation Results

The full defense comparison appears in Table 1.

Finite-difference alternative. Variant F replaces reachable-state susceptibility with proxy-risk growth along the same detached virtual displacement:

$$\mathcal{L}_{\mathrm{traj}} = \mathrm{SmoothMax}_{a,h} \frac{[S_{\widehat{q}}(\theta + \xi_{a,h}) - S_{\widehat{q}}(\theta)]_+}{\|\xi_{a,h}\|_2 + \epsilon}.$$

Unlike $\mathcal{L}_{\mathrm{reach}}$, this objective does not recompute susceptibility at the reachable state. Both variants use the same probe bank and virtual rollouts.

Table 4: **Ablation variants and susceptibility** (RN50, full FT). Bold values indicate the best anchor, and underlined values indicate the second-best anchor. Objective combinations are shown in Table 2; the key contrasts are E vs. D, H vs. F, and H vs. G.

		BadCLIP-Align		BadCLIP	
	Variant	$A_{\mathrm{post}} \downarrow$	$\widehat{\rho}_{\mathrm{proxy}}$	$A_{\mathrm{post}} \downarrow$	$\widehat{\rho}_{\mathrm{proxy}}$
A	reference configuration	82.8 ± 1.2	0.44 ± 0.02	46.9 ± 1.4	0.31 ± 0.02
B	+ compute-matched clean FT	74.5 ± 1.3	0.41 ± 0.02	27.3 ± 1.3	0.29 ± 0.02
C	+ $\mathcal{L}_{\mathrm{anchor}}$ (suppression only)	55.3 ± 1.4	0.36 ± 0.02	21.8 ± 1.2	0.24 ± 0.02
D	+ single-gradient decorrelation	41.9 ± 1.4	0.28 ± 0.02	17.5 ± 1.2	0.19 ± 0.02
E	+ update-set $\mathcal{L}_{\mathrm{dir}}$ (multi-probe)	18.6 ± 1.2	0.14 ± 0.02	9.8 ± 1.0	0.11 ± 0.01
F	+ directional reachable approx. ($\mathcal{L}_{\mathrm{traj}}$)	14.2 ± 1.1	0.11 ± 0.01	8.1 ± 0.9	0.09 ± 0.01
G	anchor + $\mathcal{L}_{\mathrm{reach}}$ (no $\mathcal{L}_{\mathrm{dir}}$)	$\underline{7.9\pm0.9}$	$\underline{0.06\pm0.01}$	$\underline{5.9\pm0.9}$	$\underline{0.06\pm0.01}$
H	**Full ImmuneCLIP (Ours)**	$\mathbf{6.4\pm0.9}$	$\mathbf{0.05\pm0.01}$	$\mathbf{5.2\pm0.8}$	$\mathbf{0.05\pm0.01}$

C.2 Ablation-Anchor Consistency

The analysis aligns the same 14 B–H checkpoints from Table 4: seven BadCLIP-Align variants and seven BadCLIP variants. Delivery-time susceptibility is plotted directly against the single-run worst post-adaptation ASR anchors. These are 14 descriptive anchors from seven paired variants across two attacks, not 14 independent experimental replicates.

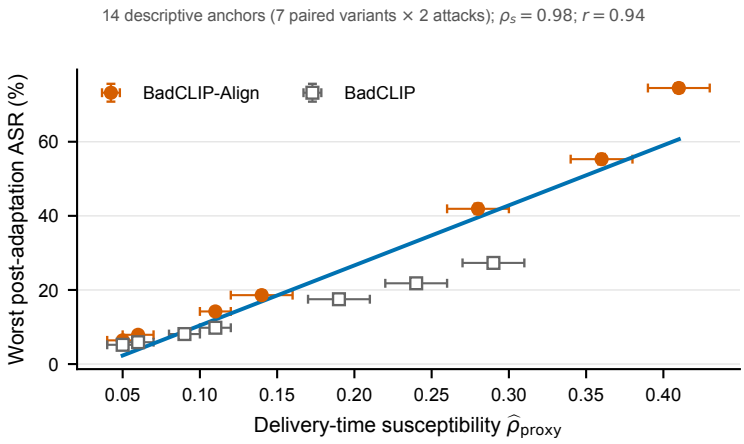


Figure 4: **Ablation-anchor consistency.** The 14 points are exactly variants B–H in Table 4. Across these descriptive anchors, $\rho_s = 0.98$ and $r = 0.94$; no inferential or cross-attack generalization claim is made.

C.3 Tasks and Adaptation Settings

Downstream tasks. After task-level adaptation on MSCOCO, ImmuneCLIP has a Worst Post-FT ASR anchor of 7.8% and R@5 of 70.6% (Figure 5 and Table 5). ImageNet zero-shot evaluation after MSCOCO fine-tuning gives a Worst Post-FT ASR anchor of 6.9% and clean accuracy of 56.4%.

Optimizers and architectures. AdamW at five times the probe learning rate is the hardest adaptation setting, with a Worst Post-FT ASR anchor of 17.6%. This degradation is consistent with lower update coverage and the larger finite-step remainder in Proposition 1. With SGD-momentum, Worst Post-FT ASR reaches 10.3%. On ViT-B/32, full fine-tuning and LoRA reach 9.5% and 12.8%, respectively. These anchor-only results suggest that optimizer scale and parameterization affect protection during adaptation.

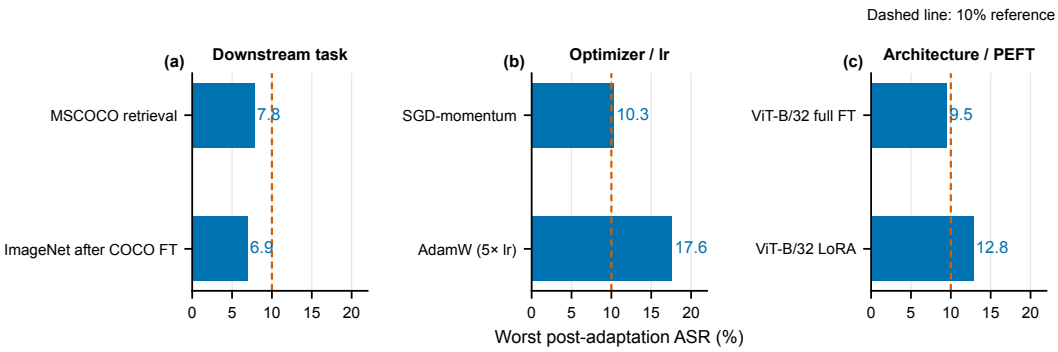


Figure 5: **Tasks and adaptation settings.** The dashed line marks a 10% post-adaptation ASR reference. Bars show ImmuneCLIP's result anchors.

C.4 Adaptive Attacks

Proxy evasion is the strongest evaluated adaptive attack, increasing Worst Post-FT ASR under ImmuneCLIP from the nominal 6.4% to 24.6%. The attacker knows the defense and jointly optimizes delivery-time stealth, persistence, and evasion of the inverted proxy. Proxy evasion moves the true risk gradient away from the gradient recovered by inversion. Layer spreading and bank-aware directions instead target the top-k layer selection and update coverage, reaching 15.2% and 21.9% Worst Post-FT ASR, respectively (Table 6). The anchor-only increase to 24.6% identifies a possible boundary when proxy quality or update coverage deteriorates, but its variability is unknown.

Table 5: **Tasks and adaptation settings.** Each row reports one evaluation setting. *Utility* is retrieval R@5 for MSCOCO and clean accuracy otherwise. Entries report ImmuneCLIP.

Axis	Setting	$A_{\text{post}} \downarrow$	Clean utility
Downstream task	MSCOCO retrieval (ASR@5 / R@5)	7.8	70.6
	ImageNet zero-shot after MSCOCO FT	6.9	56.4
Optimizer / lr	SGD-momentum, lr 10^{-6}	10.3	57.0
	AdamW, lr 5×10^{-6} ($5\times$ probe lr)	17.6	56.5
Architecture / PEFT	ViT-B/32, full FT	9.5	61.8
	ViT-B/32, LoRA	12.8	62.0

Table 6: **Adaptive attacks against ImmuneCLIP** (RN50). Each variant targets one assumption of the defense. Values are Worst Post-FT ASR percentages.

Adaptive attack	$A_{\text{post}} \downarrow$
Proxy evasion	24.6
Layer spreading	15.2
Bank-aware direction	21.9

C.5 Efficiency

The complete defense requires 21.7 GPU-hours and has an end-to-end peak memory of 19.0 GB. The ImmuneCLIP post-layer accounts for 12.6 GPU-hours: 2.2 for proxy inversion, 1.1 for update probe-bank construction, and 9.3 for immunization. The main time cost therefore comes from immunization rather than proxy or probe-bank construction.

Table 7: **Cost breakdown** (RN50, single GPU). Stage rows report local workspace measurements, whereas aggregate rows report absolute end-to-end peak memory; these are intentionally different quantities. Values are run/profile anchors.

Stage	GPU-hours	Stage workspace (GB)	End-to-end peak (GB)
Proxy inversion	2.2	4.5	–
Probe-bank build	1.1	3.6	–
Immunization	9.3	9.0	–
ImmuneCLIP post-layer	**12.6**	–	**19.0**
Complete defense	**21.7**	–	**19.0**

D Extended Related Work

D.1 Backdoor Attacks on Multimodal Models

Early backdoor attacks used fixed visual patches in supervised classifiers (Gu et al., 2019). Later attacks hid the trigger through blended perturbations (Chen et al., 2017), superimposed signals (Barni et al., 2019), image warping (Nguyen & Tran, 2021), or sample-specific patterns (Li et al., 2021c). Multimodal contrastive pre-training adds a new attack surface because image–text pairs are often collected at scale. A small poisoned subset can implant a backdoor (Carlini & Terzis, 2022), and poisoning real web data is feasible (Carlini et al., 2024). Other attacks target self-supervised encoders (Jia et al., 2022; Saha et al., 2022; Zhang et al., 2024) or poison both modalities (Yang et al., 2023b). BadCLIP further uses dual-embedding guidance to hide poisoned samples among clean features (Liang et al., 2024).

Recent attacks also aim to survive downstream fine-tuning. SBL treats backdoor injection as continual learning and uses flatness-aware optimization (Foret et al., 2021; Pham et al., 2024). BadCLIP++ directly trains multimodal backdoors for persistence under downstream adaptation (Liang

et al., 2026). BadCLIP++ is the closest work in its focus on later adaptation, but it studies attack construction. We instead start from a purified checkpoint that has passed the delivery ASR threshold and study whether benign adaptation makes it unsafe again; full utility-gate verification requires records absent from this preview. Our method addresses this post-purification stage from the defense side.

D.2 CLIP Backdoor Detection and Purification

Backdoor defenses for supervised classifiers include trigger reconstruction (Wang et al., 2019), pruning and distillation (Liu et al., 2018; Li et al., 2021b), robust training (Li et al., 2021a), adversarial perturbation and unlearning (Wu & Wang, 2021; Zeng et al., 2022), and clean fine-tuning (Sha et al., 2024). BackdoorBench provides a common evaluation framework for these methods (Wu et al., 2022).

CLIP defenses adapt these ideas to multimodal contrastive learning. CleanCLIP re-aligns image and text features on clean data, while later methods use stronger text augmentation or adversarial examples (Bansal et al., 2023; Xun et al., 2025; Kuang et al., 2024). PAR perturbs and recovers model parameters to remove the trigger response (Singh et al., 2026). InverTune estimates the target, reconstructs a trigger, and removes backdoor behavior through clustering-guided fine-tuning (Sun et al., 2026). Parameter-efficient defenses instead tune visual or class-wise prompts (Zhang et al., 2025; Kong et al., 2026). Other work detects compromised encoders or poisoned samples (Feng et al., 2023; Niu et al., 2025; Wang et al., 2024; Huang et al., 2025), or protects CLIP during pre-training (Yang et al., 2023a; 2024).

Post-training purification methods mainly evaluate the checkpoint produced at the end of defense. They do not optimize its behavior under later clean adaptation. ImmuneCLIP is applied after purification. It uses an inverted trigger to estimate residual backdoor risk, then reduces the risk along downstream update directions and at virtually reachable checkpoints.

D.3 Security During Downstream Adaptation

Fine-tuning is usually studied through utility effects such as feature distortion (Kumar et al., 2022). LoRA lowers the cost of adapting pre-trained models, while OpenCLIP provides openly available CLIP checkpoints (Hu et al., 2022; Cherti et al., 2023). The security effects of continued adaptation are less understood. Persistence-oriented attacks exploit downstream update geometry (Pham et al., 2024; Liang et al., 2026), while benign fine-tuning can also weaken safety alignment in language models (Qi et al., 2024). For CLIP backdoor defense, fine-tuning is usually the defense itself (Sha et al., 2024; Bansal et al., 2023); continued training after purification is not the protected object. We treat downstream adaptation as part of the model's security lifecycle. Our protocol measures failure during this stage, and ImmuneCLIP protects the checkpoint as it moves through it.