FROM TRIGGERS TO ANTIGENS: AN IMMUNE-INSPIRED FRAMEWORK FOR CONTINUAL BACKDOOR DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks are increasingly deployed in dynamic environments where models must continuously learn from evolving data streams, yet many existing backdoor defenses remain fundamentally limited by their reliance on predefined trigger patterns. Such a trigger-centric view fails when attacks become distributed, representation-dependent, or semantically consistent with their target labels, leaving the underlying malicious behavior invisible to conventional detectors. In this work, we reveal a different perspective: backdoors are not characterized by their observable triggers, but by recurring non-self behaviors that persist after normal semantic responses are removed. Through extensive empirical analysis, we discover that samples carrying a recurring backdoor factor across different semantic contents can share a class-conditioned residual, analogous to a pathogen-derived antigen exposed after subtracting healthy self responses. This observation motivates a new formulation of continual backdoor defense as an antigen-to-antibody learning process rather than a trigger identification problem. Building upon this insight, we propose Immune-inspired Semantic Network (ISN), a continual defense framework that learns transferable immune memory against emerging backdoor behaviors. ISN establishes a stable semantic self-reference, extracts attack-invariant antigens through complementary receptors, and transforms validated antibodies into adaptive immune responses including correction, neutralization, and vaccination. Unlike many existing approaches that memorize individual attacks, ISN accumulates reusable immune memory capable of recognizing recurrent threats across tasks and attack families. Empirical results demonstrate that our framework achieves the state-of-the-art performance.

1 INTRODUCTION

Deep neural networks are vulnerable to backdoor attacks, where adversaries implant hidden shortcuts during training to induce trigger-specific misclassification while preserving clean accuracy Nguyen et al. (2025); Gu et al. (2019). This threat is amplified in continual learning, as models must adapt to evolving tasks without prior knowledge of future attacks Guo et al. (2025); Kang et al. (2023); Nguyen et al. (2025); Zhang et al. (2025). Backdoors may persist across tasks, and catastrophic forgetting complicates the distinction between legitimate adaptation and malicious behavior. Therefore, continual backdoor defense requires detecting both current poisoned samples and recurring attack patterns across diverse contexts.

Existing defenses often assume that backdoors can be identified through predefined trigger characteristics Wang et al. (2019). These methods detect anomalous patterns Wang et al. (2019), reverse-engineer triggers Xiang et al. (2021), identify suspicious neurons Wu & Wang (2021), or analyze semantic-label inconsistencies Wang et al. (2024). However, modern attacks may employ distributed triggers, hidden representations, or semantically consistent poisoned samples Bagdasaryan & Shmatikov (2021); Zeng et al. (2021b), challenging the transfer of these detection criteria. In continual learning, evolving task semantics further hinder the reuse of previous detection criteria. Consequently, existing defenses often treat new attacks independently rather than capturing their shared malicious behaviors.

In this work, we ask a different question: *what truly distinguishes backdoored inputs from clean samples, independent of any particular trigger realization?* To answer this question, we exam-

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.

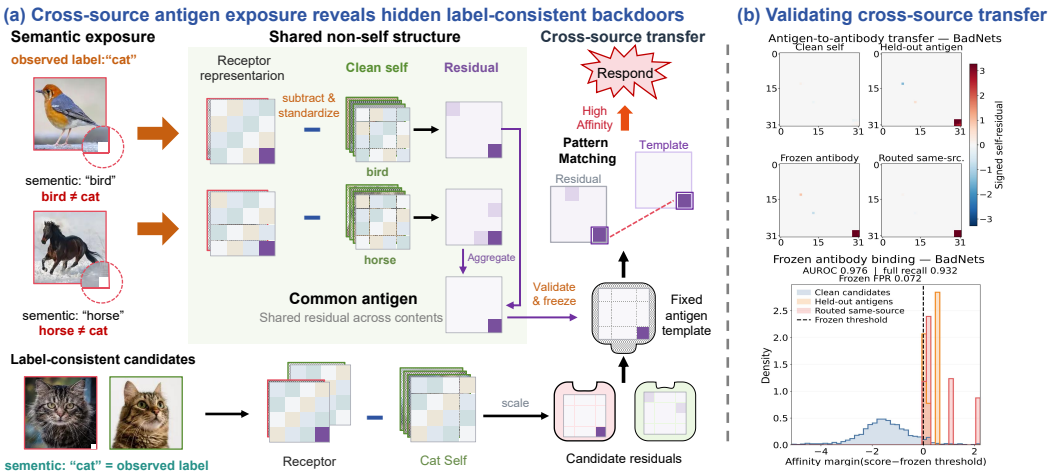


Figure 1: **Empirical motivation for our immune-inspired defense.** (a) Backdoored inputs with diverse semantic contents exhibit a recurring non-self residual after removing class-conditioned clean responses, resembling a shared pathogen antigen. This observation motivates constructing an antibody from exposed cross-source samples to recognize otherwise hidden label-consistent backdoors. (b) BadNets visualizations show that the shared residual persists in same-source samples and yields affinity responses distinct from clean target-class inputs. More details are provided in Appendix E.2.

ine poisoned samples relative to class-conditioned clean references instead of directly inspecting their observable patterns. As illustrated in Fig. 1, removing the expected class-dependent response reveals a striking empirical phenomenon in which although backdoored samples may differ substantially in semantic content, they consistently retain a recurring non-self residual after the clean response is subtracted, whereas this residual is largely absent from clean target-class samples. More importantly, the same residual persists in semantically consistent backdoors that cannot be exposed through semantic-label conflicts alone. This suggests that the common characteristic of infected samples is not a particular trigger appearance, but a shared deviation from normal class-conditioned behavior.

This observation echoes two complementary mechanisms of adaptive immunity. Cells process proteins into peptide fragments and display them in peptide–MHC complexes, allowing pathogen-derived molecular evidence to be recognized within a much larger background of self-derived material Blum et al. (2013). In humoral immunity, germinal centers expand antigen-specific B-cell clones and select immunoglobulins by affinity, producing antibody-secreting plasma cells and memory B cells that support durable and accelerated responses upon re-exposure Victora & Nussenzweig (2022). Thus, immunity retains a reusable recognition program for recurring antigenic features rather than memorizing every infected cell as a whole. Inspired by this analogy, we reformulate continual backdoor defense as an antigen-to-antibody learning process rather than a trigger-identification problem.

Building on this perspective, we propose an **Immune-inspired Semantic Network (ISN)**, a continual backdoor defense framework that discovers recurring malicious behaviors directly from observable semantic conflicts and transfers this knowledge to otherwise hidden infections. ISN first establishes a stable notion of *self* through a frozen semantic encoder, ensuring that abnormality remains comparable throughout continual learning. High-confidence semantic conflicts then serve as exposed antigen seeds, from which complementary immune receptors remove class-conditioned clean responses and aggregate the recurring non-self component into receptor-specific antigen prototypes. After held-out validation, these prototypes are frozen as local antibodies that recognize semantically consistent backdoors through clean-calibrated antigen affinity rather than trigger appearance. Finally, validated antibodies actively guide correction, neutralization, vaccination, and cross-task immune memory, allowing previously learned defenses to generalize to recurrent attacks without assuming predefined trigger patterns.

Our main contributions are summarized as follows: (1) We provide a new empirical perspective showing that diverse backdoored samples share a recurring class-conditioned non-self residual after removing expected clean responses, revealing a common abnormality beyond specific trigger patterns; (2) We formulate continual backdoor defense as an immune-inspired antigen-to-antibody

process and propose ISN, which learns reusable antibodies from observable semantic conflicts to detect hidden label-consistent backdoors. (3) We develop an integrated framework consisting of stable self-modeling, multi-receptor immune perception, adaptive immune response, and cross-task immune memory, enabling continual defense without relying on predefined trigger assumptions.

2 Related Work

Continual learning. Continual learning acquires knowledge from sequential tasks while mitigating catastrophic forgetting. Regularization constrains changes to important parameters, whereas replay revisits stored examples or predictions Kirkpatrick et al. (2017); Rebuffi et al. (2017); Buzzega et al. (2020). Prompt-based and parameter-efficient methods instead freeze pretrained backbones and learn compact task modules Wang et al. (2022b); Smith et al. (2023); Liang & Li (2024); Savadikar et al. (2026). These approaches usually assume benign streams; retaining prior knowledge may therefore also retain malicious shortcuts learned from poisoned tasks.

Backdoor attacks and defenses. Backdoors include patch, warping, frequency-domain, and clean-label triggers Gu et al. (2019); Nguyen & Tran (2021); Wang et al. (2022a); Dao et al. (2026). Existing defenses reconstruct triggers, filter suspicious samples, or purify models Wang et al. (2019); Li et al. (2021a); Zhang et al. (2026); Yang et al. (2026); Tao et al. (2026), but mainly focus on fixed datasets or models. In continual learning, backdoors may persist, transfer through prompts, or adapt to label shifts Nguyen et al. (2025); Guo et al. (2025); Gao & Liu (2025). AACL addresses this issue by using a perceptible pattern to suppress unknown triggers Umer & Polikar (2024), whereas ISN learns class-conditioned antibodies from semantic conflicts and recalls them for cross-task infection recognition. More discussions are in **Appendix** A.

3 Methodology

3.1 Overview and Problem Definition

We propose an immune-inspired continual backdoor defense framework that dynamically discovers and suppresses backdoor behaviors without relying on predefined trigger assumptions. Inspired by biological immune systems, we define four concepts throughout the paper. The *self* is the trusted-clean reference for normal behavior, where frozen semantic coordinates provide task-invariant anchors and clean receptor statistics characterize expected task-specific responses. An *antigen* is a recurrent non-self factor shared by suspicious samples after removing their expected content response. It is not restricted to predefined triggers or explicit pixel patterns. A *receptor* is an observation mechanism that reveals one manifestation of this factor, such as spatial residuals, target-conditioned responses, or frequency deviations. An *antibody* is the data-fitted state learned through a receptor, encoding both a recognition direction and, when available, a physical imprint for mitigation. Its affinity measures how strongly a new sample expresses the learned non-self factor. Thus, the receptor defines *where to look*, the antigen specifies *what recurs*, and the antibody captures *what has been learned*.

Our framework consists of four components: (1) **Innate Immune Layer** establishes a stable self-reference via a frozen semantic encoder and task-specific lightweight experts; (2) **Immune Perception** identifies semantic-conflict seeds and extracts their shared abnormality through complementary receptors; (3) **Immune Response** transforms learned antibodies into adaptive training interventions, including semantic correction, antigen neutralization, and counterfactual vaccination; and (4) **Immune Memory** consolidates validated antibodies and recalls them when similar abnormal responses reappear across tasks. The key mechanism is *antigen sharing*, where label-conflicting samples reveal evidence for detecting and suppressing their label-consistent counterparts, forming a closed loop of self-modeling, antigen discovery, adaptive response, and continual protection.

Problem Formulation. We consider a continual learning scenario with a sequence of tasks $\mathcal{T} = \{1, \ldots, T\}$. At task t, the learner receives a training set $\mathcal{D}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_t}$, where $\mathbf{x}_i$ denotes an input image and y_i is its observed label in the current class set $\mathcal{Y}_t$. All class comparisons below are restricted to $\mathcal{Y}_t$. A poisoned sample may carry either an adversarially altered label or a semantically valid label accompanied by an abnormal predictive shortcut. The objective is to maintain high clean accuracy while suppressing the attack success rate throughout the task sequence.

3.2 Innate Immune Layer

A reliable self-reference is essential for immune recognition. To avoid representation drift in continual learning, we introduce a frozen pretrained vision–language encoder: $\mathbf{z} = f_0(\mathbf{x})$, where $f_0(\cdot)$

remains fixed across tasks. It provides a stable semantic space for comparing prototypes, antigen signatures, and memory representations. The role of the frozen backbone resembles innate immunity. It does not need to recognize every possible pathogen in advance, but instead provides a general understanding of normal biological structures. Similarly, our semantic backbone does not directly identify specific backdoor triggers. It only establishes a stable reference of normal visual semantics.

Task-specific Immune Experts. To adapt to continual learning while avoiding catastrophic forgetting, we construct a lightweight expert for each task. For task t, the model contains a dedicated parameter-efficient adaptation module, and its prediction is expressed as

$$f_t(\mathbf{x}) = h_{\theta_t}\big(f_{0,\Delta_t}(\mathbf{x})\big), \tag{1}$$

where Δ_t denotes task-specific QKV-LoRA updates inserted into the backbone, f_{0,Δ_t} is the corresponding adapted forward pass, and h_{θ_t} is a cosine classifier parameterized by θ_t. The semantic reference always uses the unadapted f_0. During training on task t, only the current expert parameters are updated, while previous experts remain frozen. This design separates task adaptation from global semantic representation. The frozen backbone provides a stable self-space, whereas $f_t(\mathbf{x})$ represents the current task's class-logit vector and individual experts capture task-specific decision boundaries.

Trusted Clean Anchors. Reliable self-modeling requires uncontaminated references. For each class c, we reserve two disjoint trusted-clean subsets $\mathcal{A}_c^{anchor}$ and $\mathcal{A}_c^{cal}$. The anchor set $\mathcal{A}_c^{anchor}$ is used to construct clean semantic prototypes, estimate normal receptor responses, and warm up task experts, whereas $\mathcal{A}_c^{cal}$ is strictly reserved for final threshold estimation. The two sets are independent and never overlap with evaluation data or poisoned samples. Given the trusted anchor samples, the clean prototype of class c is defined as

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{A}_c^{anchor}|} \sum\nolimits_{\mathbf{x}_i \in \mathcal{A}_c^{anchor}} f_0(\mathbf{x}_i). \tag{2}$$

For an arbitrary sample $\mathbf{x}$, its semantic consistency with label y is measured by the cosine distance

$$d_y(\mathbf{x}) = F_{\cos}\big(f_0(\mathbf{x}), \boldsymbol{\mu}_y\big), \qquad F_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}, \tag{3}$$

where $\|\cdot\|_2$ denotes the L2 norm. The prototype distance itself does not directly indicate whether a sample is poisoned. Instead, it provides the self-deviation measurement required by later immune perception modules. $\boldsymbol{\mu}_c$ represents the semantic center of clean class c. A small $d_y(\mathbf{x})$ means that $\mathbf{x}$ resembles the self-reference of its observed class, whereas a large distance signals semantic incompatibility. Such incompatibility makes a sample a potential abnormal candidate, but is not itself a poison decision.

3.3 Immune Perception Mechanism

Immune perception operationalizes this evidence transfer in three steps: it exposes cross-class seeds, fits their recurring non-self response through complementary receptors, and uses validated antibodies to screen same-class data.

Exposing antigen seeds. We infer the semantic class as $\hat{c}(\mathbf{x}) = \arg\min_c d_c(\mathbf{x})$ and compare each class distance with its empirical clean reference. Let $p_c(\mathbf{x})$ be the upper-tail compatibility score defined in Appendix D.3. A small value indicates that $\mathbf{x}$ is unusually far from class c. For candidate target q, the visible antigen seeds are defined as:

$$\mathcal{S}_q = \{(\mathbf{x}, y) \in \mathcal{D}_t : y = q,\ \hat{c}(\mathbf{x}) \neq q, p_q(\mathbf{x}) \leq \alpha_{\text{lbl}}, p_{\hat{c}(\mathbf{x})}(\mathbf{x}) > \alpha_{\text{sem}}\}. \tag{4}$$

Here y denotes the observed label and q the candidate target shared by the seeds. The two tests reject the observed class and support a semantic alternative, with α_{lbl} and α_{sem} controlling rejection and support strength, respectively. Thus, $\mathcal{S}_q$ captures evidence for candidate target q rather than all abnormal predictions. It is split into disjoint fitting seeds $\mathcal{S}_q^{\text{fit}}$ for estimating the antibody and validation seeds $\mathcal{S}_q^{\text{val}}$ for evaluation. Consistency across diverse seeds, rather than a single conflict, provides evidence for a shared antigen.

Extracting receptor-specific antibodies. Each receptor observes a different manifestation of the same candidate antigen. We denote the spatial, model-conditioned, and frequency receptors by r_{spa}, r_{mc},

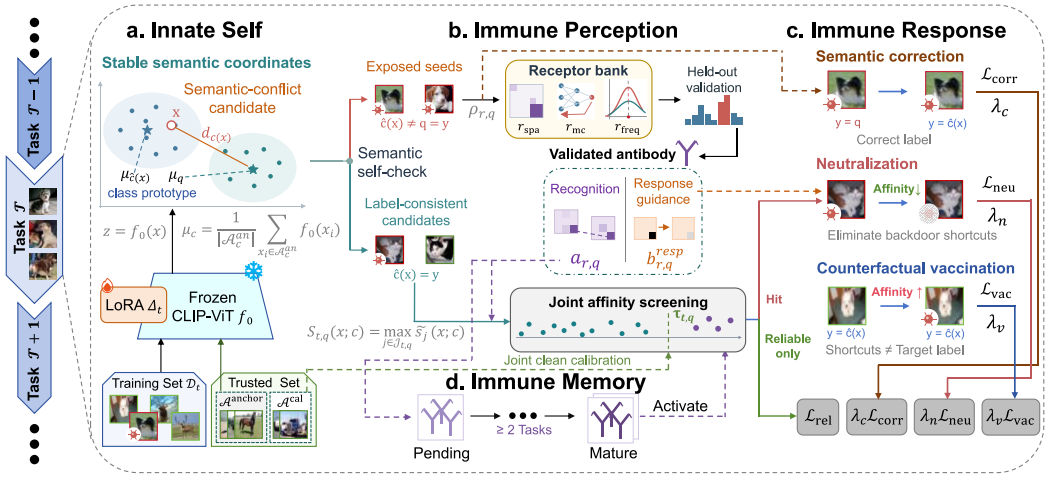


Figure 2: **Overview of ISN.** (a) A frozen CLIP-ViT encoder and trusted clean anchors define stable class-conditioned self references, while task-specific LoRA experts learn sequential tasks. (b) Semantic-conflict seeds reveal candidate non-self responses; complementary receptors fit antibodies that pass held-out validation and jointly calibrated affinity screening. (c) Validated antibodies guide semantic correction, neutralization, and counterfactual vaccination. (d) Recurrent validated antibodies mature into memory cells and can be recalled in later tasks.

and r_{freq}, respectively, and collect these receptor types in $\mathcal{R} = \{r_{\text{spa}}, r_{\text{mc}}, r_{\text{freq}}\}$. For a receptor type $r \in \mathcal{R}$, $\Psi_{r,q}(\mathbf{x})$ denotes its observation of image $\mathbf{x}$ for candidate target q. The observation associated with each receptor is specified in Table 1. We first remove the clean response of content class c:

$$\rho_{r,q}(\mathbf{x}; c) = \mathcal{P}_r\left(\frac{\Psi_{r,q}(\mathbf{x}) - b_{r,q,c}}{\sigma_{r,q,c} + \epsilon}\right), \qquad c = \hat{c}(\mathbf{x}), \tag{5}$$

where $b_{r,q,c}$ and $\sigma_{r,q,c}$ are the clean center and robust scale, $\epsilon > 0$ prevents division by zero, and $\mathcal{P}_r$ clips or normalizes the residual. The centered, scaled quantity measures departure from the expected response of class c in units of clean variation. Spatial and frequency receptors preserve this magnitude up to clipping. The model-conditioned receptor normalizes it to compare directions. Thus $\rho_{r,q}(\mathbf{x}; c)$ describes a candidate non-self response, whose repetition across seeds must still be validated. We use $c = \hat{c}(\mathbf{x})$ for untrusted images and the known class for trusted references. The fitting residuals are then consolidated into an antibody and its binding score is defined as:

$$\begin{aligned} \mathsf{Ab}_{r,q} &= (a_{r,q}, b^{\text{resp}}_{r,q}), \qquad a_{r,q} = \text{Norm}\big(\text{Agg}_r\{\rho_{r,q}(\mathbf{x}; \hat{c}(\mathbf{x})) : (x, y) \in \mathcal{S}^{\text{fit}}_q\}\big), \\ s_{r,q}(\mathbf{x}; c) &= \langle \rho_{r,q}(\mathbf{x}; c), a_{r,q} \rangle, \end{aligned} \tag{6}$$

Here $\mathsf{Ab}_{r,q}$ denotes the complete antibody fitted through receptor r for candidate target q; it is a structured detector state, not a scalar score. Its first component $a_{r,q}$ is the shared residual direction used for recognition. Its second component $b^{\text{resp}}_{r,q}$ stores a physical imprint for constructing response views, and equals $\varnothing$ when no direct image edit is available. This distinction preserves the support and amplitude that normalization removes from $a_{r,q}$.

The operator Agg_r maps fitting-seed residuals to a representative pattern by aggregating recurring coordinates and selecting spatial/frequency support according to Table 1. $\text{Norm}(v) = v/\max(\|v\|_2, \epsilon)$ removes scale, and degenerate aggregates are discarded. The affinity $s_{r,q}(\mathbf{x}; c)$ is the inner product between the flattened residual and the aggregated direction: positive values indicate alignment, near-zero values weak evidence, and negative values opposing residuals. For normalized model-conditioned residuals, it is a cosine score in $[-1, 1]$; spatial and frequency affinities also encode residual magnitude. None represents a poison probability, so decisions rely on clean-calibrated thresholds.

The spatial receptor searches for a pattern repeated at consistent image locations. The model-conditioned receptor instead observes how a temporary expert, trained on the task data and then frozen, responds to target q. It can reveal a distributed pattern that is inconspicuous in pixel space. The frequency receptor searches for sparse transform-domain deviations, balancing the source classes so that one does not dominate the template. Appendix D.4 gives their construction details.

Table 1: Receptor-specific instances of Eqs. (5)–(6). The observation and aggregation determine recognition; the stored imprint connects each antibody to the response operators in Section 3.4.

Receptor	Observation $\Psi_{r,q}$	Aggregation Agg_r	Imprint $b^{\mathrm{resp}}_{r,q}$
r_{spa}	RGB pixels	Median residual on a stable local support	Binary mask and RGB pattern
r_{mc}	Pooled activation–gradient magnitudes for target q	Median residual direction	$\varnothing$; affinity guides transformation search
r_{freq}	Non-DC YCrCb DCT coefficients	Source-class medians followed by sparse support selection	Signed perturbation in raw DCT units

Validation and hidden-infection screening. An antibody is locally validated only when it separates held-out seeds from content-matched trusted-clean examples under fixed ranking, effect-size, and low-false-positive criteria. Passing affinities are standardized by their clean reference so that current and recalled antibodies can be compared. Let $\mathcal{J}_{t,q}$ index the strongest locally validated antibody, if any, and the mature cells admitted for current target q, and let $\widetilde{s}_j$ denote antibody j's standardized affinity. We screen label-consistent examples by

$$\begin{aligned} S_{t,q}(\mathbf{x}; c) &= \max_{j \in \mathcal{J}_{t,q}} \widetilde{s}_j(x; c), \\ H_{t,q}(\mathbf{x}, y) &= \mathbb{I}[y = \hat{c}(x) = q]\, \mathbb{I}[p_q(\mathbf{x}) > \alpha_{\mathrm{sem}}]\, \mathbb{I}[S_{t,q}(\mathbf{x}; q) > \tau_{t,q}]. \end{aligned} \tag{7}$$

The threshold $\tau_{t,q}$ is calibrated on the maximum score of independent clean class-q examples, controlling the joint decision rather than individual antibodies. An empty antibody set gives no hit, with $S_{t,q} = -\infty$. $S_{t,q}$ denotes the strongest standardized antigen evidence for $\mathbf{x}$, and $H_{t,q} = 1$ indicates a same-class hit only when semantic agreement and abnormal affinity hold jointly; $\mathbb{I}[\cdot]$ is the indicator function. A hit flags a suspicious training view, while $H_{t,q} = 0$ only means no flag is produced, not that the sample is clean. Semantic consistency provides the retained label, and antibody evidence determines intervention.

3.4 Immune Response Mechanism

Recognition becomes protective only when it changes the learner's supervision. We therefore couple each antibody to three complementary responses: *correction* repairs an exposed label conflict, *neutralization* removes antigen evidence from a detected same-class infection, and *vaccination* places that evidence on a non-target donor. The latter two act in opposite affinity directions, while each preserves its sample's semantic label.

Let $\mathcal{C}_t$ contain the high-confidence cross-class conflicts in Eq. (4) and their observed label is replaced by the confidence-controlled pseudo-label $\hat{c}(\mathbf{x})$. Let $\mathcal{H}_t$ contain detected same-class tuples $(\mathbf{x}, y, q, j)$ with $y = \hat{c}(\mathbf{x}) = q$, and let $\mathcal{V}_t$ contain reliable donor tuples $(\mathbf{x}, y, q, j)$ with $y = \hat{c}(\mathbf{x}) \neq q$. For a detected sample, j indexes the strongest acting antibody, with ties resolved by a fixed rule; for a donor, j indexes the antibody used to construct its vaccination. In the losses below, $\mathcal{H}_t$ and $\mathcal{V}_t$ include only tuples with accepted edits. For antibody j, N_j lowers affinity on an infection, whereas V_j raises affinity on a donor. Every generated view $\mathbf{x}'$ must obey

$$\hat{c}(\mathbf{x}') = \hat{c}(\mathbf{x}) = y, \qquad p_y(\mathbf{x}') > \alpha_{\mathrm{sem}}, \qquad F_{\cos}(f_0(\mathbf{x}'), f_0(\mathbf{x})) \leq \delta, \tag{8}$$

where δ bounds semantic drift in the frozen self-space. Neutralization is accepted only when $\widetilde{s}_j(N_j(\mathbf{x}); y)$ decreases by a fixed margin and vaccination must increase $\widetilde{s}_j(V_j(\mathbf{x}); y)$ beyond both that margin and the clean screening threshold. The first equality preserves the inferred class, the compatibility test $p_y(\mathbf{x}')$ keeps the edit within clean class-y variation, and the cosine constraint prevents a visually mild edit from causing a large semantic displacement. Failed edits are discarded. The three response losses make their directions explicit, expressed as:

$$\begin{aligned} &\mathcal{L}_{\mathrm{corr}} = \mathbb{E}_{(\mathbf{x},y)\in\mathcal{C}_t} \ell_{\mathrm{CE}}(f_t(\mathbf{x}), \hat{c}(\mathbf{x})), \qquad \mathcal{L}_{\mathrm{neu}} = \mathbb{E}_{(\mathbf{x},y,q,j)\in\mathcal{H}_t} \ell_{\mathrm{CE}}(f_t(N_j(\mathbf{x})), y), \\ &\mathcal{L}_{\mathrm{vac}} = \mathbb{E}_{(\mathbf{x},y,q,j)\in\mathcal{V}_t} \ell_{\mathrm{CE}}(f_t(V_j(\mathbf{x})), y), \end{aligned} \tag{9}$$

where ℓ_{CE} is cross-entropy and expectations denote empirical means. Minimization reinforces the retained semantic label. Correction removes erroneous antigen–q associations, neutralization weakens the antigen while preserving the target, and vaccination generates the counterfactual *antigen present, target absent*. Spatial/frequency antibodies edit image/DCT evidence, while model-conditioned antibodies select admissible transformed views. Appendix D.6 details the operators and tests.

Table 2: Matched-backbone BadNets and WaNet results on CIFAR-100 and TinyImageNet. All use frozen CLIP ViT-B/16 task-LoRA backbone. ASR is clean-conditioned Task-IL attack success rate; CA is clean accuracy. Values are percentages (mean ± standard deviation over three independent runs unless marked). **None** columns report single runs. Lower ASR and higher CA are better.

Method	CIFAR-100					TinyImageNet				
	None	BadNets		WaNet		None	BadNets		WaNet	
	CA	ASR	CA	ASR	CA	CA	ASR	CA	ASR	CA
Vanilla LoRA	97.44	92.45±0.20	97.36±0.04	43.47±2.49	95.25±0.07	95.10	97.18±0.11	94.97±0.08	92.42±0.44	94.60±0.39
LoRA-AACL	97.41	**0.00**±0.00	88.78±0.51	41.18±3.96	94.06±0.41	95.12	97.44±0.25	94.72±0.19	92.02±2.13	94.52±0.07
LoRA-ABL	97.33	92.04±0.16	96.22±0.18	42.93±1.08	91.77±0.27	93.68	96.84±0.36	93.44±0.15	90.23±0.35	93.55±0.07
LoRA-I-BAU	96.54	91.35±0.05	94.31±0.80	27.50±6.22	94.88±0.90	93.64	96.88±0.09	93.11±0.50	87.91±0.80	92.53±0.84
LoRA-PDB	96.30	0.16±0.01	96.27±0.21	0.62±0.08	95.75±0.30	94.38	4.43±4.38	93.16±0.18	**0.20**±0.09	93.37±0.01
LoRA-LMR	98.26	85.77±3.54	98.08±0.06	0.16±0.02	**98.04**±0.09	96.39	89.58±3.20	95.84±0.10	57.06±0.69	95.66±0.15
Ours	97.81	0.01±0.01	**98.09**±0.08	**0.14**±0.04	97.74±0.00	96.92	**0.01**±0.02	**96.78**±0.08	0.32±0.35	**96.63**±0.01

Table 3: BadNets and WaNet results on CIFAR-100 and TinyImageNet for native-backbone continual-learning methods and Ours (frozen CLIP ViT-B/16 task-LoRA).

Method	CIFAR-100					TinyImageNet				
	None	BadNets		WaNet		None	BadNets		WaNet	
	CA	ASR	CA	ASR	CA	CA	ASR	CA	ASR	CA
DER++	64.43	76.50±0.19	55.61±0.21	10.69±1.84	55.62±0.28	51.91	76.66±5.41	40.45±0.88	36.55±5.62	40.77±1.33
L2P	94.35	92.08±0.42	97.19±0.29	28.51±3.90	92.34±0.55	95.87	97.62±0.41	96.05±0.14	85.62±0.85	94.47±0.12
CODA-Prompt	94.66	80.19±0.24	95.73±0.17	18.77±2.44	93.96±0.75	95.96	85.35±0.17	95.91±0.00	68.94±0.98	94.84±0.15
InfLoRA	96.77	43.73±0.25	96.02±0.05	9.32±0.23	94.47±0.02	95.31	60.88±3.79	95.53±0.17	67.12±0.55	94.16±0.00
EWC-LoRA	96.83	92.22±0.02	97.50±0.05	43.25±2.31	92.50±0.55	96.61	96.43±0.91	96.64±0.05	95.58±0.28	95.83±0.25
KeepLoRA	97.82	88.42±1.84	97.84±0.02	31.83±1.71	93.38±0.69	97.42	95.81±0.02	**97.10**±0.04	89.30±0.97	95.87±0.01
Ours	97.81	**0.01**±0.01	**98.09**±0.08	**0.14**±0.04	**97.74**±0.00	96.92	**0.01**±0.02	96.78±0.08	**0.32**±0.35	**96.63**±0.01

Let $\mathcal{B}_t$ contain reliable originals assigned to neither correction nor neutralization, with $\mathcal{L}_{\mathrm{rel}} = \mathbb{E}_{(x,y)\in\mathcal{B}_t}\ell_{\mathrm{CE}}(f_t(x), y)$. The task expert minimizes the following loss function

$$\mathcal{L}_t = \mathcal{L}_{\mathrm{rel}} + \lambda_c\mathcal{L}_{\mathrm{corr}} + \lambda_n\mathcal{L}_{\mathrm{neu}} + \lambda_v\mathcal{L}_{\mathrm{vac}}, \tag{10}$$

where nonnegative weights λ_c, λ_n, and λ_v control the three supervision terms. Zero weights disable the corresponding losses, and larger weights increase their emphasis. Empty sets yield zero expectations. Originals use at most one route, while vaccination provides counterfactual views. Ambiguous samples and failed neutralizations are excluded, and antibodies are fixed before optimization to prevent self-altering evidence.

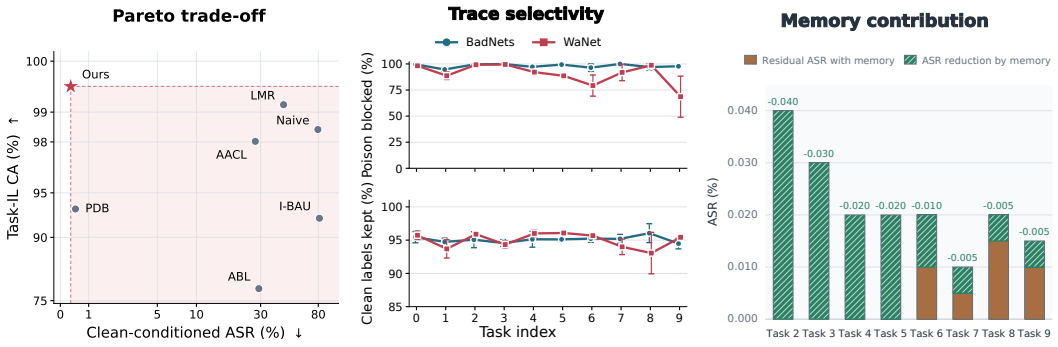
3.5 Immune Memory Mechanism

Immune memory carries a learned antibody forward by progressively extending the evidence that supports its use. A sufficiently supported local antibody first seeds a *pending cell*, whose stored pattern becomes a reference for recognizing recurrence in later tasks. Such recurrence consolidates the same cell into *mature memory*; its antibody can then be recalled to screen new samples and become *active* on a compatible match. The stages therefore operate on one evolving cell: generation records an initial observation, maturation establishes cross-task support, and activation turns that accumulated evidence into a current intervention.

Generation: from a local antibody to a pending cell. The selected local antibody from Section 3.3 supplies the material for memory: its detection direction and response imprint are estimated from fitting seeds, and its reliability is assessed on validation seeds against content-matched clean references. It is admitted to memory only if its ranking, separation, and recall pass the local validation criteria and at least m_{mem} validation seeds support that assessment. The additional support requirement prevents a small locally successful seed set from establishing a persistent record. The explicit admission rule is given in Eq. (28). At task end, an admitted proposal with no matching stored pattern initializes

$$m_j = (r_j, \mathsf{Ab}_j, e_j), \qquad \mathsf{Ab}_j = (a_j, b_j^{\mathrm{resp}}), \qquad e_j = (n_j, \bar{\kappa}_j, \chi_j). \tag{11}$$

Here r_j specifies the observation mechanism, Ab_j stores its detection direction and optional response imprint, and e_j records the evidence for reuse: n_j counts distinct supporting tasks, $\bar{\kappa}_j$ is the weakest



(a) **(b)** **(c)**

Figure 3: Three complementary analyses of ISN: (a) architecture-controlled ASR–CA trade-off, (b) task-wise routing selectivity, and (c) task-wise ASR reduction from mature memory. Results are averaged over three independent runs.

accepted cross-task template agreement, and $\chi_j \in \{0, 1\}$ records whether the required seed–clean separation tests were passed. A new cell has $e_j = (1, 1, 1)$: one supporting task, an initial agreement value, and passed local separation tests. Since $n_j = 1$, it remains pending. Generation thus produces both a pattern to compare against future antibodies and an evidence record to update; a proposal matching an existing cell instead enters the consolidation step below.

Maturation: from recurrence to recall eligibility. The pending cell produced above becomes the reference for the next encounter. When a later task supplies another proposal passing the same admission rule, its direction is compared with the stored a_j within the same receptor family. A sufficiently similar proposal is consolidated only if the stored and updated directions also separate the new validation seeds from clean references. This encounter updates Ab_j and the existing evidence record e_j: it adds one task to n_j, records the cross-task agreement in $\bar{\kappa}_j$, and checks clean separation through χ_j. Repeated batches within one task add no task support. The resulting cell becomes eligible for future recall when

$$\mathrm{Mature}(m_j) = \mathbb{I}[n_j \geq 2]\, \mathbb{I}[\bar{\kappa}_j \geq \gamma]\, \mathbb{I}[\chi_j = 1]\,, \tag{12}$$

where Mature is a binary eligibility function in which a value of 1 means that the cell has passed all three requirements and may be considered for recall. A value of 0 keeps it ineligible. The threshold $\gamma \in (0, 1]$ specifies the minimum cosine agreement and larger γ demands more similar recurrent patterns. $\chi_j = 1$ means that clean separation is supported, and $\chi_j = 0$ withholds eligibility. Thus, recurrence alone does not establish mature memory, and maturity alone does not trigger a response. Updates are committed at task end and become available to subsequent tasks. Appendix D.7 specifies the matching and evidence updates.

Activation: from eligible memory to a current response. Maturation stores antibodies with sufficient evidence for recall. In later tasks, the detector a_j is reused for screening and the imprint b_j^{resp} for response, while pending cells await further evidence. The memory bank remains frozen; current clean anchors recalibrate receptor centers and scales. Mature cells are evaluated under each label context q, and compatible cells form $\mathcal{J}_{t,q}^{\mathrm{mem}}$ before joining the local antibody set $\mathcal{J}_{t,q}$ for calibration of $\tau_{t,q}$. Model-conditioned cells additionally require current held-out separation; otherwise they remain inactive. For a stored cell j, define

$$\begin{aligned} s_{j,q}^{(t)}(x; c) &= \langle \rho_{r_j,q}^{(t)}(x; c), a_j \rangle, \\ I_{t,j,q}^{\mathrm{mem}}(x, y) &= \mathbb{I}[j \in \mathcal{J}_{t,q}^{\mathrm{mem}}]\, H_{t,q}(x, y)\, \mathbb{I}[\tilde{s}_{j,q}^{(t)}(x; q) > \tau_{t,q}]. \end{aligned} \tag{13}$$

The raw affinity $s_{j,q}^{(t)}$ measures alignment with the stored direction after removing the current class response. Its standardized counterpart $\tilde{s}_{j,q}^{(t)}$ measures evidence relative to current clean scores. $I_{t,j,q}^{\mathrm{mem}} = 1$ means that an eligible, compatible memory cell matches a semantically consistent sample above the calibrated threshold. A value of 0 means that this cell does not activate on this sample. The membership test incorporates maturity and current compatibility, while $H_{t,q}$ supplies the semantic and screening conditions from Eq. (7).

An activation flags the current view for the response in Section 3.4 while retaining the label supported by $\hat{c}(x)$. For multiple matches, the highest standardized affinity selects the response operator.

Its imprint generates a neutralized view and vaccinations on reliable non-target donors; otherwise, the view is quarantined. Task-specific labels, statistics, and thresholds are discarded, while memory retains reusable antibodies and maturation evidence. Activations reuse this evidence without increasing n_j, and only seed-derived proposals admitted by the generation rule consolidate the cell. Mature cells may remain dormant and reactivate when compatible evidence reappears.

4 EXPERIMENTS

We provide the detailed experiment settings in **Appendix** E.1.

4.1 MAIN RESULTS

Architecture-controlled comparison. Table 2 compares methods under the same frozen CLIP–LoRA backbone. ISN achieves strong robustness while maintaining high accuracy. On CIFAR-100, ISN reduces BadNets and WaNet ASR to 0.01% and 0.14%, with CA of 98.09% and 97.74%, respectively. On TinyImageNet, ISN achieves 0.01% BadNets ASR with 96.78% CA. For WaNet, LoRA-PDB obtains slightly lower ASR (0.20% vs. 0.32%), but with lower CA (93.37% vs. 96.63%).

Table 4: LTB results on CIFAR-100 and TinyImageNet.

Method	CIFAR-100		TinyImageNet	
	ASR	CA	ASR	CA
Vanilla LoRA	82.88±6.12	97.63±0.07	98.98±1.02	95.03±0.11
LoRA-AACL	26.04±18.96	97.31±0.20	12.63±8.20	94.89±0.05
LoRA-ABL	91.20±2.55	96.09±0.12	97.77±0.15	93.75±0.00
LoRA-I-BAU	80.23±5.48	95.84±0.21	100.00±0.00	92.41±0.19
LoRA-PDB	**0.00**±0.00	96.46±0.27	**0.00**±0.00	93.52±0.17
LoRA-LMR	65.48±24.10	**98.23**±0.06	70.37±27.51	95.85±0.10
Ours	2.60±1.57	98.03±0.01	3.24±1.57	96.79±0.21
L2P	2.27±1.11	97.63±0.02	**0.00**±0.00	96.72±0.05
CODA-Prompt	12.64±4.04	97.10±0.10	4.88±2.10	96.42±0.07
InfLoRA	40.96±14.76	97.60±0.15	12.76±10.63	96.54±0.03
EWC-LoRA	3.50±2.50	97.72±0.05	**0.00**±0.00	96.66±0.10
KeepLoRA	30.04±21.96	**98.47**±0.08	4.17±5.89	**97.11**±0.04

Comparison with continual learners. Table 3 compares methods under native architectures and protocols. ISN achieves the lowest mean ASR across datasets and attacks while preserving competitive CA. For example, KeepLoRA obtains 97.10% CA but retains 95.81% ASR on TinyImageNet/BadNets, whereas ISN achieves 96.78% CA with only 0.01% ASR, showing that clean accuracy alone does not guarantee backdoor robustness.

Persistent training-level attacks. Table 4 evaluates LTB after the complete sequence. On CIFAR-100, ISN achieves 2.60% ASR and 98.03% CA, versus LoRA-LMR's 65.48% and 98.23%. On TinyImageNet, ISN attains 3.24% ASR with 96.79% CA. Although LoRA-PDB reports zero ASR on both datasets, its CA is lower at 96.46% and 93.52%, respectively, showing a trade-off between robustness and clean accuracy.

Table 5: Ablation on CIFAR-10 under BadNets. Final Task-IL CA and clean-conditioned ASR (%).

Variant	CA ↑	ASR ↓
Full	99.59±0.08	**0.07**±0.04
w/o Perception	99.61±0.07	29.52±7.57
w/o Response	**99.71**±0.00	94.00±0.57
w/o Memory	99.65±0.06	9.30±12.83

4.2 ABLATION STUDY

Table 5 evaluates immune components under BadNets. Removing all receptor pathways while keeping semantic screening increases ASR from 0.07% to 29.52%, while disabling the full response raises it to 94.00%, demonstrating the importance of detection-to-intervention translation. Removing memory increases mean ASR to 9.30% with higher variance, while CA remains above 99.5%.

Architecture-controlled trade-off. Figure 3(a) compares methods under the same frozen CLIP ViT-B/16 backbone and LoRA architecture. Averaged over BadNets and WaNet on seq-CIFAR10, ISN achieves 0.10% clean-conditioned ASR and 99.60% Task-IL accuracy, without sacrificing clean performance.

Selective routing and response. Figure 3(b) shows training actions on seq-CIFAR100. ISN blocks most BadNets samples through relabeling, neutralization, or quarantine, while preserving original labels for 93–96% of clean samples. Whiskers denote ranges over three runs.

Contribution of immune memory. Figure 3(c) evaluates memory after cross-task consolidation. Memory reduces ASR across Tasks 2–9, eliminating mean ASR on Tasks 2–5 and reducing late-task ASR from 0.0219% to 0.0050% (77.1% reduction) with only -0.034 percentage points clean accuracy change.

5 CONCLUSION

In this work, we propose an immune-inspired perspective for continual backdoor defense, revealing that backdoors are characterized by recurring non-self behaviors beyond specific triggers. Our Immune-inspired Semantic Network (ISN) learns transferable antibodies for detection, mitigation, and continual protection without relying on predefined trigger assumptions.

AI USE STATEMENT

Generative AI tools were used for language editing. All technical claims, implementations, experiments, and reported numbers were verified by the authors.

REFERENCES

Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1505–1521, 2021.

Janice S Blum, Pamela A Wearsch, and Peter Cresswell. Pathways of antigen processing. *Annual review of immunology*, 31(1):443–473, 2013.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025*, 2021.

Thinh Dao, Khoa D Doan, and Kok-Seng Wong. Clean-label physical backdoor attacks with data distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3515–3523, 2026.

Rui Gao and Weiwei Liu. Red alarm: Controllable backdoor attack in continual learning. *Neural Networks*, 188:107479, 2025.

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

Zhen Guo, Abhinav Kumar, and Reza Tourani. Persistent backdoor attacks in continual learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6379–6397, 2025.

Siteng Kang, Zhan Shi, and Xinhua Zhang. Poisoning generative replay in continual learning to promote forgetting. In *International Conference on Machine Learning*, pp. 15769–15785. PMLR, 2023.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proc. of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto, 2009.

Ya Le and Xuan Yang. Tiny imageNet visual recognition challenge. Technical report, Univ. of Stanford, 2015.

Jianwei Li and Jung-Eun Kim. Purifying generative llms from backdoors without prior knowledge or clean reference. *arXiv preprint arXiv:2603.13461*, 2026.

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

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021b.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647. IEEE, 2024.

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*, pp. 273–294. Springer, 2018.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. Keeplora: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 136017–136040, 2026.

Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

Trang Nguyen, Anh Tran, and Nhat Ho. Attack on prompt: Backdoor attack in prompt-based continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19686–19694, 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

Chinmay Savadikar, Michelle Dai, and Tianfu Wu. Cheem: Continual learning by reuse, new, adapt and skip-a hierarchical exploration-exploitation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25066–25076, 2026.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11909–11919, 2023.

Guanhong Tao, Siyuan Cheng, Guangyu Shen, Yingqi Liu, Shengwei An, Zhuo Zhang, Zhenting Wang, Hanxi Guo, and Xiangyu Zhang. Mitigating backdoor attacks via trigger reconstruction and model hardening. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 580–590. IEEE, 2026.

Muhammad Umer and Robi Polikar. Adversary aware continual learning. *IEEE Access*, 12:126108–126121, 2024.

Gabriel D Victora and Michel C Nussenzweig. Germinal centers. *Annual review of immunology*, 40:413–442, 2022.

Bo Wang, Fei Yu, Fei Wei, Yi Li, and Wei Wang. Invisible intruders: Label-consistent backdoor attack using re-parameterized noise trigger. *IEEE Transactions on Multimedia*, 26:10766–10778, 2024.

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 *2019 IEEE symposium on security and privacy (SP)*, pp. 707–723. IEEE, 2019.

Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision*, pp. 396–413. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Mitigating backdoor attack by injecting proactive defensive backdoor. *Advances in Neural Information Processing Systems*, 37:80674–80705, 2024.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021.

Zhen Xiang, David J Miller, and George Kesidis. Reverse engineering imperceptible backdoor attacks on deep neural networks for detection and training set cleansing. *Computers & Security*, 106:102280, 2021.

Zhiguo Yang, Dongsheng Xu, Ruizhi Zhong, Jiacheng Pi, Xingxing Huang, and Wenjie Ruan. Logit-margin repulsion for backdoor defense. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34918–34928, 2026.

Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021a.

Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16453–16461. IEEE, 2021b.

Mingda Zhang, Mingli Zhu, Zihao Zhu, Li Shen, and Baoyuan Wu. Reliable poisoned sample detection against backdoor attacks enhanced by sharpness aware minimization. In *International Conference on Learning Representations*, volume 2026, pp. 123199–123229, 2026.

Xueyi Zhang, Peiyin Zhu, Chengwei Zhang, Zhiyuan Yan, Jikang Cheng, Mingrui Lao, Siqi Cai, and Yanming Guo. Generalization-preserved learning: Closing the backdoor to catastrophic forgetting in continual deepfake detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3798–3808. IEEE, 2025.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations*, volume 2026, pp. 42895–42919, 2026.

A Additional Related Work

A.1 Continual Learning Strategies

Regularization and replay. Elastic Weight Consolidation penalizes changes to parameters estimated to be important for previous tasks Kirkpatrick et al. (2017). Replay-based approaches instead revisit earlier information: iCaRL retains class exemplars Rebuffi et al. (2017), Gradient Episodic Memory constrains updates using gradients on stored examples Lopez-Paz & Ranzato (2017), and DER++ replays both labels and past logits Buzzega et al. (2020). ER-ACE applies an asymmetric loss to incoming and replayed examples to reduce disruptive representation changes at task boundaries Caccia et al. (2021). These mechanisms improve retention, but retention itself is not a test of whether a stored example is benign. If poisoned examples enter a replay buffer, subsequent updates may encounter them again.

Pretrained and parameter-efficient learners. L2P selects prompts from a learned pool, whereas CODA-Prompt decomposes prompt components to support rehearsal-free adaptation Wang et al. (2022b); Smith et al. (2023). InfLoRA uses low-rank task adaptation Liang & Li (2024); CHEEM chooses among reusing, adding, adapting, and skipping model components Savadikar et al. (2026). Such designs limit interference with previous tasks but do not, by themselves, prevent a current task module from learning a poisoned shortcut. ISN therefore combines task-specific LoRA experts with a separate, unadapted semantic reference and explicit screening of the current task's data.

A.2 Backdoor Attacks in Static and Continual Learning

BadNets associates a fixed patch with an attacker-chosen label Gu et al. (2019). WaNet replaces a localized patch with smooth geometric warping, while FTrojan embeds perturbations in frequency coefficients Nguyen & Tran (2021); Wang et al. (2022a). Clean-label physical attacks can poison target-class examples without altering their labels Dao et al. (2026). Such examples explain why a semantic-label conflict can expose some poisoned samples but cannot serve as a complete backdoor detector.

Continual learning creates an additional persistence problem. Attack on Prompt manipulates prompt selection and training to make a trigger effective across incremental updates Nguyen et al. (2025).

Guo et al. introduce Blind Task Backdoor, which alters the loss computation, and Latent Task Backdoor, which affects one task while the backdoor remains effective after subsequent benign tasks Guo et al. (2025). CBACL uses controllable triggers to address label shift in sequential learning Gao & Liu (2025). These mechanisms motivate defenses that can retain validated malicious-behavior evidence across tasks without assuming a fixed visible trigger.

A.3 Backdoor Defense Mechanisms

Detection and post-training removal. Neural Cleanse searches for unusually small reconstructed triggers Wang et al. (2019), and ABL exploits training dynamics to separate suspicious examples before relearning Li et al. (2021a). Fine-Pruning combines pruning and clean fine-tuning Liu et al. (2018); NAD uses a clean subset to distill attention from a repaired teacher Li et al. (2021b); and I-BAU formulates backdoor removal as implicit-hypergradient adversarial unlearning Zeng et al. (2021a). Recent methods address weaker poison signals through sharpness-aware training Zhang et al. (2026), enlarge benign-versus-target logit margins before pruning Yang et al. (2026), or reconstruct malicious inputs and harden the model against them Tao et al. (2026). These methods differ in their access to clean data and poisoned models, and are generally formulated for a fixed training stage or a trained model rather than reusable cross-task memory.

Proactive and continual defenses. PDB injects a defensive backdoor during training and uses a defensive trigger with a reversible label mapping at inference Wei et al. (2024). AACL specifically addresses continual learning by training with a stronger perceptible decoy pattern and applying it to test inputs to overpower an unknown imperceptible trigger Umer & Polikar (2024). A separate immunization-inspired method finds common backdoor signatures across synthetic model variants for post-training purification of generative LLMs Li & Kim (2026). ISN instead fits antibodies from class-conditioned residuals of observable semantic conflicts, validates them against clean references, and recalls them when compatible abnormal responses reappear in later image tasks.

A.4 Trusted Clean References and the Scope of ISN

Access to trusted data is a material part of a defense's threat model. Fine-Pruning and NAD use clean examples for model repair Liu et al. (2018); Li et al. (2021b); ISN also requires small, disjoint trusted-clean anchor and calibration sets for each class. The anchors establish semantic prototypes and normal receptor responses, while the calibration set controls screening thresholds. Antibody fitting uses suspicious examples from the untrusted task stream, with a separate held-out seed split for validation. Consequently, ISN does not claim clean-data-free operation. If a new attack produces too few reliable semantic conflicts, ISN does not establish a new antibody; a previously matured compatible antibody can still be recalled. This distinction separates the evidence required to discover a new antigen from the evidence required to recognize one already learned.

B Threat Model and Defense Scope

Continual-learning interface. We consider a class-disjoint Task-IL stream $\mathcal{T} = \{1, \ldots, T\}$ in which task t provides the observed training set $\mathcal{D}_t$ and class set $\mathcal{Y}_t$. Each task is visited once, although multiple optimization epochs are allowed while its data are available. The task identifier is known during both training and evaluation, and predictions are restricted to $\mathcal{Y}_t$. The frozen semantic encoder f_0 is shared across tasks, whereas the LoRA update Δ_t and classifier h_{θ_t} are task-specific. Once task t is complete, its expert is frozen. Therefore, ISN addresses protection of a sequential Task-IL learner rather than task discovery or Class-IL routing.

Adversarial capabilities. The untrusted portion of $\mathcal{D}_t$ may contain an unknown mixture of clean and poisoned examples. For data-level attacks, the adversary may select a subset of the current stream, transform its inputs, and either retain or replace their observed labels with a target $q \in \mathcal{Y}_t$. This includes both label-conflicting poisons and label-consistent target-class infections. The trigger may be localized, distributed, frequency-based, or dependent on the current model response, and the adversary may choose which tasks, source classes, and target classes to attack. We do not assume that the defense knows the attack family, trigger, poisoned indices, source class, target class, or poisoning rate. Our LTB evaluation additionally considers the prescribed training-level attacker, which inserts its malicious objective and constrains a selected parameter subset during the attacked

Table 6: Centralized threat model and information available to ISN.

Item	Assumption or access
Learning protocol	Class-disjoint Task-IL; task identity is available; one sequential visit per task.
Untrusted stream	May contain clean, relabeled, or label-consistent poisoned samples at an unknown rate.
Attacker knowledge	May know the model and defense design; receives no access to trusted subsets or their membership.
Defender knowledge	No attack identity, trigger, poison indicator, source class, target class, or poisoning rate.
Trusted resource	Disjoint clean anchor and calibration sets for every class; neither overlaps evaluation data.
Persistent state	Frozen backbone, completed task experts, and bounded antibody memory; no earlier untrusted stream.
Evaluation threat	Triggered non-target test inputs; poison annotations are used only to compute metrics.

task; the wrapper does not alter the trusted references or ISN's screening and response rules. The defense does not rely on secrecy of its architecture, but the reported attacks are standard non-adaptive instantiations rather than attacks optimized end-to-end against ISN.

Defender access and trusted resources. For every class c, the defender has two small, disjoint trusted-clean sets, $\mathcal{A}_c^{\text{anchor}}$ and $\mathcal{A}_c^{\text{cal}}$, sampled from the training split and removed from the untrusted stream. The attacker cannot modify their images or labels. Anchors may define semantic prototypes, robust self-statistics, and antibody-development references; calibration examples are accessed only after the acting antibody family has been fixed and determine its final threshold. Apart from these sets, all current-task examples and their observed labels are treated as untrusted. Poison identities, original labels, attack targets, and triggered test counterparts are unavailable to the method and are used only for post-hoc evaluation. The defender also retains the frozen pretrained encoder, completed task experts, and the bounded immune-memory bank, but does not retain earlier untrusted task streams.

Security objectives and non-goals. The primary objectives are to preserve final Task-IL clean accuracy and to suppress clean-conditioned attack success throughout the task sequence. ISN also aims to prevent a validated recurrent non-self response from being learned as a target shortcut and to reuse that evidence when a compatible response reappears. These objectives do not constitute a certified guarantee against arbitrary adaptive attacks. In particular, the present formulation does not cover corruption of trusted anchors or calibration data, malicious modification of the pretrained encoder or defense code, absence of the task identifier, or test-time attacks unrelated to the learned backdoor association. A previously unseen attack that yields too few reliable semantic conflicts may not create a local antibody; unless compatible mature memory is available, ISN abstains rather than inferring an unsupported antigen.

C End-to-End Learning and Inference

C.1 Task-wise ISN Training Procedure

Algorithm 1 summarizes the complete method defined in Sections 3.2–3.5 and Appendix D. The operators Fit, Validate, Recall, and Commit are shorthand for the receptor-specific construction, held-out validation, current-task memory admission, and task-end memory update specified in Appendices D.4, D.5, and D.7. They use no ground-truth poison indicators.

The ordering in Algorithm 1 enforces three separation rules. First, fitting seeds never score the antibody that they alone validate; selection uses the disjoint validation split. Second, the acting antibody set is fixed before the calibration split is accessed, and calibration determines only the final joint threshold. Third, responses generated by an antibody cannot create support for that antibody: memory proposals are derived only from the original semantic-conflict seed pathway, and all mem-

Algorithm 1 End-to-end task-wise training of ISN

Require: Task stream $\{\mathcal{D}_t, \mathcal{Y}_t\}_{t=1}^{T}$; frozen encoder f_0; trusted sets $\{\mathcal{A}_c^{\text{anchor}}, \mathcal{A}_c^{\text{cal}}\}_c$; receptor family $\mathcal{R}$
Ensure: Frozen task experts $\{f_t\}_{t=1}^{T}$ and immune-memory bank $\mathcal{M}$

Initialize $\mathcal{M} \leftarrow \varnothing$
for $t = 1$ to T **do**
  Initialize task expert f_t and temporary heating expert g_t; freeze f_0 and all previous experts
  Build semantic prototypes and clean receptor statistics from $\{\mathcal{A}_c^{\text{anchor}}\}_c$; warm up the current task head
  Train g_t for its fixed schedule on the observed current-task supervision and then freeze it
  for each candidate target $q \in \mathcal{Y}_t$ **do**
    Expose $\mathcal{S}_q$ using Eq. (4) and split it once into $\mathcal{S}_q^{\text{fit}}$ and $\mathcal{S}_q^{\text{val}}$
    Set the local acting set $\mathcal{J}_{t,q}^{\text{local}} \leftarrow \varnothing$
    if $\mathcal{S}_q$ has the prescribed fitting and validation support **then**
      for each receptor $r \in \mathcal{R}$ **do**
        $\mathsf{Ab}_{r,q} \leftarrow \mathrm{Fit}(r, q, \mathcal{S}_q^{\text{fit}}, \{\mathcal{A}_c^{\text{anchor}}\}_c, g_t)$
        Evaluate $(\mathrm{AUC}_{r,q}, E_{r,q}, R_{r,q})$ on held-out seeds and content-matched clean references
      end for
      Select the strongest receptor passing all fixed validation cutoffs; if none passes, keep $\mathcal{J}_{t,q}^{\text{local}} = \varnothing$
    end if
    $\mathcal{J}_{t,q}^{\text{mem}} \leftarrow \mathrm{Recall}(\mathcal{M}, q, \{\mathcal{A}_c^{\text{anchor}}\}_c, \mathcal{S}_q^{\text{val}}, g_t)$
    $\mathcal{J}_{t,q} \leftarrow \mathcal{J}_{t,q}^{\text{local}} \cup \mathcal{J}_{t,q}^{\text{mem}}$
    Calibrate $\tau_{t,q}$ on $\mathcal{A}_q^{\text{cal}}$ using the maximum standardized score in Eq. (23); an unsupported or empty family abstains
  end for
  Freeze all acting antibodies, thresholds, and response operators before final expert optimization
  for each optimization epoch and mini-batch $\mathcal{B} \subset \mathcal{D}_t$ **do**
    Route high-confidence semantic conflicts to $\mathcal{C}_t$ and reliable originals to $\mathcal{B}_t$
    Apply Eq. (7) to route label-consistent hits to candidate neutralization and reliable non-target donors to candidate vaccination
    Admit only response views satisfying Eqs. (8) and (27); quarantine ambiguous samples and failed neutralizations
    Update only the current expert by minimizing Eq. (10)
  end for
  Form memory proposals only from locally validated seed-derived antibodies satisfying Eq. (28)
  $\mathcal{M} \leftarrow \mathrm{Commit}(\mathcal{M}, \text{proposals})$ by same-receptor matching, validated consolidation, maturation, and the bounded-bank rule
  Store (Δ_t, θ_t); discard g_t, current seed partitions, task-specific self-statistics, thresholds, and generated response views
end for
return $\{f_t\}_{t=1}^{T}, \mathcal{M}$

ory changes are staged until task end. Consequently, a sample-level activation can guide the current expert without recursively maturing the cell that produced it.

C.2 Inference Procedure

ISN uses the standard Task-IL inference interface. Given a test image $\mathbf{x}$ and its task identifier t, the learner activates only the frozen expert of that task and restricts prediction to $\mathcal{Y}_t$:

$$\widehat{y}(\mathbf{x}, t) = \arg\max_{c \in \mathcal{Y}_t} \left[h_{\theta_t}(f_{0,\Delta_t}(\mathbf{x}))\right]_c . \tag{14}$$

The same forward path is used for clean and triggered test inputs. Semantic screening, antibody affinity, neutralization, vaccination, and quarantine are training-time mechanisms: they shape (Δ_t, θ_t) but are not applied to test images. The temporary heating expert, seed sets, clean response statistics, and task-specific thresholds are therefore unnecessary at inference. Likewise, the immune-memory bank is retained for defending future training tasks, not for routing predictions among completed experts. Thus, for a known task identifier, ISN adds no antibody search or image-transformation branch to the deployed task-expert forward pass.

Algorithm 2 Task-IL inference with a trained ISN expert

Require: Test image $\mathbf{x}$; task identifier t; frozen f_0; task parameters (Δ_t, θ_t); task class set $\mathcal{Y}_t$
Ensure: Predicted label $\widehat{y} \in \mathcal{Y}_t$
1: $\mathbf{z}_t \leftarrow f_{0,\Delta_t}(\mathbf{x})$
2: $\boldsymbol{\ell}_t \leftarrow h_{\theta_t}(\mathbf{z}_t)$
3: **return** $\arg\max_{c \in \mathcal{Y}_t} [\boldsymbol{\ell}_t]_c$

D Implementation Details

D.1 Notation and Protocol

Table 7 summarizes the symbols shared across the method and Appendix D. We use one notation for each conceptual role: y always denotes an observed label, q a candidate attack target, and $\hat{c}(x)$ the class inferred from the frozen semantic self-space.

Table 7: Principal notation used by ISN.

Symbol	Meaning
$t, \mathcal{D}_t$	Current task and its observed training set
f_0, f_t, g_t	Frozen semantic encoder, task expert, and heating expert
$y, q, \hat{c}(x)$	Observed label, candidate target, and inferred content class
$\mathcal{A}_c^{anchor}, \mathcal{A}_c^{cal}$	Trusted anchor and calibration sets of class c
$\mathcal{S}_q^{\text{fit}}, \mathcal{S}_q^{\text{val}}$	Modeling and validation antigen seeds for target q
$\mathcal{R} = \{r_{\text{spa}}, r_{\text{mc}}, r_{\text{freq}}\}$	Spatial, model-conditioned, and frequency receptor types
$r \in \mathcal{R}, \Psi_{r,q}$	Selected receptor type and its observation of an input
$r_j \in \mathcal{R}$	Receptor type stored in memory cell j
$b_{r,q,c}, \sigma_{r,q,c}$	Clean response center and natural clean-response scale
$\rho_{r,q}, a_{r,q}, s_{r,q}$	Standardized non-self residual, detection direction, and affinity
$b_{r,q}^{\text{resp}}$	Optional physical imprint used to construct a response
N_j, V_j	Neutralization and vaccination operators of antibody j
$\alpha_{\text{lbl}}, \alpha_{\text{sem}}$	Observed-label rejection and inferred-class compatibility cutoffs
$\alpha, \tau_{t,q}$	Nominal clean false-positive level and its calibrated score threshold
δ, η	Maximum semantic drift and minimum directional affinity change
$\lambda_c, \lambda_n, \lambda_v$	Weights of correction, neutralization, and vaccination losses
$\mathsf{Ab}_{r,q}, \mathrm{Agg}_r$	Structured antibody state and receptor-specific seed aggregation
$m_j = (r_j, \mathsf{Ab}_j, e_j)$	Memory cell: receptor, antibody, and maturation evidence
$\mathrm{Mature}(m_j)$	Binary recall eligibility: 1 eligible, 0 ineligible
$\mathcal{J}_{t,q}^{\text{mem}}$	Mature cells compatible with current task t and label q
$I_{t,j,q}^{\text{mem}}(x, y)$	Binary activation of cell j on a current sample
$n_j, \bar{\kappa}_j, \chi_j$	Distinct-task support, template agreement, and clean-rejection state

The task expert is selected by the Task-IL task identifier. Experimental figures and logs use zero-based task indices; index 2 is the third task, consistent with recall beginning at task 3 in the one-based method notation. For every class, the anchor and calibration subsets are disjoint and fixed before model training. Anchor samples may be used for prototype and self-statistic estimation; calibration samples are never used to fit or select an antibody. All per-task antibodies and response operators are fixed before the final optimization of f_t; response views are instantiated from these fixed operators. Appendix B specifies the computations omitted from the main text. Unless otherwise stated, medians, absolute deviations, and divisions are applied coordinate-wise. All hyperparameters, including support budgets and acceptance cutoffs, are fixed across experiments. Throughout Appendix B, $\epsilon > 0$ is a small numerical stabilizer, TopK returns a support set rather than sorted values, and Norm denotes L2 normalization. A vector is discarded when its norm is at most ϵ; all retained directions have unit norm. vec flattens a tensor, $\mathrm{reshape}$ restores its receptor coordinates, and $\odot$ is coordinate-wise multiplication. The two clean splits remain separate: cross-fitted anchor scores support development decisions, and calibration scores determine only the final threshold.

D.2 Innate Immune Layer

The pretrained parameters of f_0 remain frozen. For an attention projection W_ℓ at layer ℓ, task t introduces a low-rank update

$$W_{\ell,t} = W_\ell + \frac{\alpha_{\mathrm{L}}}{r_{\mathrm{L}}} B_{\ell,t} A_{\ell,t}, \tag{15}$$

where $A_{\ell,t} \in \mathbb{R}^{r_{\mathrm{L}} \times d_{\mathrm{in}}}$ and $B_{\ell,t} \in \mathbb{R}^{d_{\mathrm{out}} \times r_{\mathrm{L}}}$ are the task-specific LoRA factors, r_{L} is the LoRA rank, and α_{L} is its scale; d_{in} and d_{out} are the input and output projection dimensions. $W_{\ell,t}$ is therefore the effective projection used by expert t, while $B_{\ell,t}A_{\ell,t}$ represents only that task's low-rank departure from the shared pretrained projection W_ℓ. The update is inserted into the designated QKV projections, and a task-specific cosine head h_{θ_t} is trained jointly with these factors. Earlier task factors and heads remain frozen.

Class prototypes use the unadapted output of f_0, as in the main text. In implementation, both the encoded sample and prototype are L2-normalized before their cosine distance is evaluated. The anchor subset also warms up the task head and supplies the clean self-statistics below; the independent calibration subset is accessed only after receptor selection.

D.3 Semantic-conflict Seeds

For class c, the empirical upper-tail compatibility score used in seed discovery is

$$\begin{aligned} \mu_c^{(-v)} &= \frac{1}{|\mathcal{A}_c^{\mathrm{anchor}}| - 1} \sum_{u \in \mathcal{A}_c^{\mathrm{anchor}} \setminus \{v\}} f_0(u), \qquad d_c^{(-v)}(v) = F_{\cos}(f_0(v), \mu_c^{(-v)}), \\ p_c(x) &= \frac{1 + \sum_{v \in \mathcal{A}_c^{\mathrm{anchor}}} \mathbb{I}[d_c^{(-v)}(v) \geq d_c(x)]}{|\mathcal{A}_c^{\mathrm{anchor}}| + 1}. \end{aligned} \tag{16}$$

Here $\mu_c^{(-v)}$ is a leave-one-out clean prototype and $d_c^{(-v)}(v)$ is anchor v's distance from a center that it did not help construct. This prevents an artificially small reference distance caused by self-inclusion. $|\cdot|$ denotes set cardinality and $\mathbb{I}[\cdot]$ is the indicator function. Thus, $p_c(x)$ represents the empirical fraction of clean class-c examples at least as distant as $\mathbf{x}$: a small score means that $\mathbf{x}$ is atypical for class c, not that it has a small posterior probability of belonging to c. The finite-sample correction prevents a zero value. This leave-one-out construction is a semantic screening heuristic, not an exact conformal p-value: the anchor and query distances use different fitted prototypes. At least two anchors are required. A sample contributes to $\mathcal{S}_q$ only when it is atypical for observed label q, compatible with inferred class $\hat{c}(x)$, and $\hat{c}(x) \neq q$. A candidate target is considered only after its seed set reaches the prescribed minimum support. The accepted seeds are divided once into $\mathcal{S}_q^{\mathrm{fit}}$ and $\mathcal{S}_q^{\mathrm{val}}$; the latter is not used for masks, frequency supports, or antibody templates. For the model-conditioned receptor, these validation examples may have appeared during heating. Their held-out status refers to template fitting; it does not establish transfer to same-class infections, which is evaluated separately.

D.4 Receptor-specific Antibody Construction

Robust self-statistics. For an observation $\Psi_{r,q}$ and semantic class c, the trusted-clean center and scale are

$$\begin{aligned} b_{r,q,c} &= \mathrm{Med}_{v \in \mathcal{A}_c^{anchor}} \Psi_{r,q}(v), \\ \sigma_{r,q,c} &= 1.4826\, \mathrm{Med}_{v \in \mathcal{A}_c^{anchor}} |\Psi_{r,q}(v) - b_{r,q,c}|. \end{aligned} \tag{17}$$

The class-wise scale above applies unless a receptor-specific definition overrides it. In particular, the frequency receptor below uses a shared scale, $\sigma_{r_{\mathrm{freq}},q,c} = \sigma_{r_{\mathrm{freq}}}$. The factor 1.4826 places the median absolute deviation on a scale comparable to a standard deviation under a Gaussian reference. A positive scale floor is used before division. Hence $b_{r,q,c}$ represents the typical clean response of class c, whereas $\sigma_{r,q,c}$ represents its normal coordinate-wise variation. Subtracting this center and dividing by the scale expresses departures in units of clean variability. The subsequent $\mathcal{P}_r$ determines whether the receptor preserves residual magnitude or retains only its direction. For the spatial and frequency receptors, $\Psi_{r,q}$ is independent of q before target-specific masking; for the model-conditioned receptor, it depends on the target margin.

Spatial mask and imprint. For the spatial receptor, $\Psi_{r_{\mathrm{spa}},q}(x) = x$ and $\mathcal{P}_{r_{\mathrm{spa}}}$ clips standardized RGB residuals to $[-8, 8]$ coordinate-wise. This limits isolated extreme pixels before estimating a recurring pattern. Let $\delta_q(u) = \mathrm{Med}_{(x,y)\in\mathcal{S}_q^{\mathrm{fit}}} \rho_{r_{\mathrm{spa}},q}(x; \hat{c}(x), u)$ be the shared signed residual at location u. Its stability score and support mask are

$$\begin{aligned} P_q(u) &= \frac{\|\delta_q(u)\|_2}{\mathrm{Med}_{(x,y)\in\mathcal{S}_q^{\mathrm{fit}}} \|\rho_{r_{\mathrm{spa}},q}(x; \hat{c}(x), u) - \delta_q(u)\|_2 + \epsilon}, \\ \Omega_{r_{\mathrm{spa}},q} &= \mathrm{TopK}_u(P_q(u), K_{r_{\mathrm{spa}}}), \qquad M_q(u) = \mathbb{I}[u \in \Omega_{r_{\mathrm{spa}},q}], \\ a_{r_{\mathrm{spa}},q} &= \mathrm{Norm}\,\mathrm{vec}(M_q \odot \delta_q), \qquad b^{\mathrm{resp}}_{r_{\mathrm{spa}},q} = (M_q, I_q). \end{aligned} \tag{18}$$

$P_q(u)$ is large when seeds share a strong residual at position u with little seed-to-seed variation. $\Omega_{r_{\mathrm{spa}},q}$ is the selected location set; its binary mask has $M_q(u) = 1$ at selected positions and 0 elsewhere, and is broadcast over color channels. Consequently, $\mathrm{Agg}_{r_{\mathrm{spa}}}$ in Eq. (6) is precisely $\mathrm{vec}(M_q \odot \delta_q)$ before normalization. For an $H \times W$ image, the default budget is $K_{r_{\mathrm{spa}}} = \max(4, \lceil 0.01HW \rceil)$. Here I_q is the coordinate-wise median RGB value of fitting seeds on M_q. Thus the unit vector $a_{r_{\mathrm{spa}},q}$ is used only for detection, whereas the binary mask and pixel-domain imprint preserve the support and amplitude used for response.

Heating and contribution extraction. The temporary heating expert is trained for a fixed schedule to minimize

$$\mathcal{L}_{\mathrm{heat}}(\omega) = \frac{1}{|\mathcal{D}_t|} \sum_{(x,y)\in\mathcal{D}_t} \ell_{\mathrm{CE}}(g_{t,\omega}(x), y). \tag{19}$$

Here ω denotes the temporary expert parameters and $g_{t,\omega}(x)$ its logits. $\mathcal{L}_{\mathrm{heat}}$ measures its fit to the observed task supervision; the resulting parameters are used only to extract model-conditioned evidence. A fixed training schedule does not assume convergence to a global minimizer. After the fixed heating schedule, g_t is frozen. For target q, we compute $m_q(x) = g_{t,q}(x) - \max_{j\neq q} g_{t,j}(x)$ and use the activation-gradient product at the designated layer. The diagnostic expert is intentionally trained on the observed task supervision: $m_q(x)$ then represents how strongly its current decision favors q over the best alternative, exposing features that participate in that association. It is not the final task expert used for evaluation. Absolute contributions are summed over channels and adaptively pooled to 4×4, yielding

$$\begin{aligned} \Psi_{r_{\mathrm{mc}},q}(x) &= \mathrm{vec}\left[\mathrm{Pool}_{4\times4}\left(\sum_k \left| A_k(x) \odot \frac{\partial m_q(x)}{\partial A_k(x)} \right|\right)\right], \\ a_{r_{\mathrm{mc}},q} &= \mathrm{Norm}\Big(\mathrm{Med}_{(x,y)\in\mathcal{S}_q^{\mathrm{fit}}} \rho_{r_{\mathrm{mc}},q}(x; \hat{c}(x))\Big). \end{aligned} \tag{20}$$

$A_k(x)$ is channel k of the designated intermediate layer, and $\mathrm{Pool}_{4\times4}$ is adaptive average pooling to 16 spatial bins. The activation–gradient product is a first-order sensitivity measure of the target margin. Taking its absolute value retains contribution magnitude from features that either support or oppose q; it is not a signed or causal attribution. Thus $\Psi_{r_{\mathrm{mc}},q}$ describes the spatial distribution of sensitivity to the target decision. Here $\mathrm{Agg}_{r_{\mathrm{mc}}}$ is the coordinate-wise median. Clean centers and scales follow Eq. (17), and $\mathcal{P}_{r_{\mathrm{mc}}} = \mathrm{Norm}$ makes the residual a directional signature. Because this descriptor has no direct image inverse, $b^{\mathrm{resp}}_{r_{\mathrm{mc}},q} = \varnothing$; the learned affinity guides the transformation search in Appendix D.6. The same layer and pooling grid are used in every task.

Frequency support. Images are converted to YCrCb before applying an orthonormal channel-wise 2D DCT, so $\Psi_{r_{\mathrm{freq}},q}(x) = \mathrm{vec}(D(x))$. DC terms are removed. After subtracting the class-specific center, this receptor estimates one coordinate-wise robust scale from all class-centered trusted residuals. Writing $e(v) = \mathrm{vec}(D(v)) - b_{r_{\mathrm{freq}},c}$ for a clean anchor v of class c, this scale is $\sigma_{r_{\mathrm{freq}}} = 1.4826\,\mathrm{Med}_v\,|e(v) - \mathrm{Med}_{v'}\,e(v')|$ with a positive numerical floor. Pooling across classes gives the receptor a common scale for signed frequency deviations. Standardized residuals are clipped to $[-50, 50]$ before aggregation. For every source class represented among target-q seeds, we compute

$$\begin{aligned} u_{q,c} &= \mathrm{Med}_{(x,y)\in\mathcal{S}_{q,c}^{\mathrm{fit}}} \rho_{r_{\mathrm{freq}},q}(x; c), & u_q &= \mathrm{Med}_{c\in\mathcal{K}_q}\, u_{q,c}, \\ \Omega_{r_{\mathrm{freq}},q} &= \mathrm{TopK}(|u_q|, K_{r_{\mathrm{freq}}}), & w_q &= \Pi_{\Omega_{r_{\mathrm{freq}},q}} u_q, \\ a_{r_{\mathrm{freq}},q} &= \mathrm{Norm}(w_q), & b^{\mathrm{resp}}_{r_{\mathrm{freq}},q} &= \mathrm{reshape}\big(w_q \odot (\sigma_{r_{\mathrm{freq}}} + \epsilon)\big). \end{aligned} \tag{21}$$

$\mathcal{S}^{\text{fit}}_{q,c}$ contains seeds inferred as source class c, and $\mathcal{K}_q$ is the set of represented source classes. $u_{q,c}$ describes the typical residual within source class c; u_q extracts its consensus across sources. The two medians give every source class equal influence, reducing dominance by the largest source group. w_q is the sparse signed consensus in standardized coefficient units; its support and sign specify which frequencies deviate and in which direction. Thus $\text{Agg}_{r_{\text{freq}}}$ outputs w_q. Π_Ω zeros coefficients outside support Ω, and $K_{r_{\text{freq}}} = \min(16, d_D)$, where d_D is the number of retained non-DC coefficients. The unit vector $a_{r_{\text{freq}},q}$ supports affinity scoring; multiplying by the clean scale restores an estimated perturbation to raw DCT units for response. Because clipping bounds the residual, this imprint need not recover the original attack perturbation exactly. Omitted DC entries are restored as zeros when reshaping. Zero templates are discarded. The same color conversion, DCT normalization, inverse conversion, and clipping conventions are used throughout training.

D.5 Antibody Validation and Calibration

For receptor r and candidate target q, let $S^+ = \{s_{r,q}(x; \hat{c}(x)) : (x, y) \in \mathcal{S}^{\text{val}}_q\}$, and let S^- contain affinities of content-matched trusted anchors from the source classes represented in S^+. A clean negative is scored by cross-fitting so that it does not estimate its own center or scale. We evaluate held-out ranking, robust separation, and recall at a clean-reference threshold:

$$\begin{aligned} \text{AUC}_{r,q} &= \frac{1}{|S^+||S^-|} \sum_{u \in S^+} \sum_{v \in S^-} \left(\mathbb{I}[u > v] + \tfrac{1}{2}\mathbb{I}[u = v] \right), \\ E_{r,q} &= \frac{\text{Med}(S^+) - \text{Med}(S^-)}{1.4826\,\text{MAD}(S^-) + \epsilon}, \\ R_{r,q} &= \frac{1}{|S^+|} \sum_{u \in S^+} \mathbb{I}[u > Q_\alpha(S^-)]. \end{aligned} \tag{22}$$

$\text{AUC}_{r,q}$ is the pairwise probability that a validation seed outranks a clean sample, with ties receiving half credit. $E_{r,q}$ is the median seed–clean separation normalized by the robust clean scale, and $R_{r,q}$ is validation recall above the clean threshold. Numerically, $\text{AUC}_{r,q} = 0.5$ indicates chance ranking, a larger positive $E_{r,q}$ indicates that the typical seed is separated by more clean-scale units, and $R_{r,q}$ measures how much held-out antigen evidence survives at the chosen operating point. The three quantities therefore test ordering, separation magnitude, and usable coverage rather than repeating the same criterion. Here $\text{MAD}(S) = \text{Med}_{v \in S} |v - \text{Med}(S)|$. For n ordered scores $u_{(1)} \le \cdots \le u_{(n)}$, the finite-sample operator is $Q_\alpha(\{u_i\}) = u_{(k)}$ for $k = \lceil (n+1)(1-\alpha) \rceil \le n$, and $+\infty$ otherwise. Here $\alpha \in (0, 1)$ is the nominal clean false-positive level; a smaller α imposes a higher threshold. Empty score collections also yield $+\infty$, causing abstention, and empty validation groups cannot admit an antibody. An antibody is validated when $\text{AUC}_{r,q} \ge \eta_{\text{auc}}$, $E_{r,q} \ge \eta_{\text{eff}}$, and $R_{r,q} \ge \eta_{\text{rec}}$. The three cutoffs are fixed validation hyperparameters. Passing receptors are ranked by AUC, then recall R, then effect size E, with a fixed receptor order resolving ties. If none passes, there is no local antibody for that target.

For every selected or recalled antibody j, let $S^-_{j,q}$ contain current class-q anchor affinities, scored by cross-fitting. These scores define the clean baseline $\mu^s_{j,q} = \text{Med}(S^-_{j,q})$ and spread $\sigma^s_{j,q} = \max\{1.4826\,\text{MAD}(S^-_{j,q}), \epsilon\}$. The comparable score is $\widetilde{s}_j(x; c) = (s_j(x; c) - \mu^s_{j,q})/\sigma^s_{j,q}$. Zero denotes the clean baseline; a positive value denotes excess affinity in clean-scale units. This reference is fixed for a target-q antibody even when it scores a donor after subtracting the donor's content-class response. Unlike seed validation negatives, these standardization references are available even when the current task contains no seeds. After the acting set $\mathcal{J}_{t,q}$ is fixed, the calibration subset is accessed once:

$$\tau_{t,q} = Q_\alpha\left(\left\{ \max_{j \in \mathcal{J}_{t,q}} \widetilde{s}_j(v; q) : v \in \mathcal{A}^{\text{cal}}_q \right\} \right). \tag{23}$$

This is the threshold used in Eq. (7). Calibrating the maximum directly accounts for the current and recalled antibodies in the same decision and avoids inflating the clean false-positive rate through an independent OR rule. Operationally, $\tau_{t,q}$ is the largest combined antigen evidence normally tolerated for clean class-q examples at level α; a score above it is unusual for the current self-environment. The finite-sample clean false-positive interpretation requires exchangeable calibration and future clean examples and a scoring rule fixed before calibration. It does not guarantee detection of every

poison or control errors aggregated across all tasks. With fewer than 19 calibration examples and $\alpha = 0.05$, the rule returns $+\infty$ and abstains.

D.6 Immune Response Operators

Each direct response uses the physical imprint paired with its detection direction. For the spatial receptor,

$$\begin{aligned} N_{r_{\mathrm{spa}},q}(x) &= (1 - M_q) \odot x + M_q \odot \mu(x), \\ V_{r_{\mathrm{spa}},q}(x) &= (1 - M_q) \odot x + M_q \odot I_q. \end{aligned} \tag{24}$$

where $(M_q, I_q) = b^{\mathrm{resp}}_{r_{\mathrm{spa}},q}$ and $\mu(x)$ is the channel-wise image mean broadcast over spatial locations. Neutralization suppresses the learned support; vaccination transplants its fitted pixel-domain imprint. For the frequency receptor, let $B_q = b^{\mathrm{resp}}_{r_{\mathrm{freq}},q}$ be the sparse raw-coefficient imprint. Separate strengths β_n and β_v are selected from a fixed grid:

$$\begin{aligned} N_{r_{\mathrm{freq}},q}(x) &= \mathrm{clip}_{[0,1]}\, D^{-1}(D(x) - \beta_n B_q), \\ V_{r_{\mathrm{freq}},q}(x) &= \mathrm{clip}_{[0,1]}\, D^{-1}(D(x) + \beta_v B_q). \end{aligned} \tag{25}$$

Here D^{-1} includes the inverse DCT and inverse color conversion. The opposite signs remove and controllably reintroduce the same frequency antigen; β_n and β_v are edit strengths, distinct from the loss weights λ_n and λ_v. Both edit strengths are nonnegative: zero produces the original image and cannot satisfy the strict affinity change test, while larger values apply a stronger signed perturbation. Clipping keeps the resulting image in the valid pixel range. For a model-conditioned antibody, direct subtraction is unavailable. We therefore search a fixed transformation family $\mathcal{G}(x)$. Let $\mathcal{G}_{\mathrm{sem}}(x, y)$ be the subset satisfying Eq. (8):

$$\begin{aligned} N_{r_{\mathrm{mc}},q}(x) &\in \operatorname*{arg\,min}_{x' \in \mathcal{G}_{\mathrm{sem}}(x,y)} \widetilde{s}_{r_{\mathrm{mc}},q}(x'; y), \\ V_{r_{\mathrm{mc}},q}(x) &\in \operatorname*{arg\,max}_{x' \in \mathcal{G}_{\mathrm{sem}}(x,y)} \widetilde{s}_{r_{\mathrm{mc}},q}(x'; y). \end{aligned} \tag{26}$$

The minimum-affinity candidate is the neutralized view and the maximum-affinity candidate is the vaccinated view. In our implementation, $\mathcal{G}$ contains a 3×3 binomial blur mixed equally with the input, $0.9\times$ down-up sampling, one-pixel crop-resize, contrast scaling by 0.95, brightness offset 0.02, and two opposite half-pixel sinusoidal elastic warps. If no candidate is semantically admissible, that response branch abstains.

Across all receptor families, an edited view is used only when its standardized affinity changes in the intended direction:

$$\widetilde{s}_j(N_j(x); y) < \widetilde{s}_j(x; y) - \eta, \qquad \widetilde{s}_j(V_j(x); y) > \max\{\widetilde{s}_j(x; y) + \eta, \tau_{t,q}\}, \tag{27}$$

where $\eta > 0$ is the minimum affinity change. Neutralization applies only to hits with $y = q$; vaccination applies only to donors with $y \neq q$. The second test makes the vaccinated view antigen-positive relative to the target-class clean reference used for $\tau_{t,q}$; it is not a donor-class false-positive guarantee. Accepted views receive the labels specified by Eq. (9); rejected views supply no supervised update. These constraints admit changes in the measured antigen response; they do not establish that an unknown trigger has been exactly removed or reconstructed.

For completeness, reliable originals satisfy $\hat{c}(x) = y$, $p_y(x) > \alpha_{\mathrm{sem}}$, and receive no screening hit. Their set $\mathcal{B}_t$ supplies ordinary supervision. Conflicts in $\mathcal{C}_t = \bigcup_q \mathcal{S}_q$ receive only the corrected label. Detected originals are replaced by accepted neutralizations, or quarantined if no edit passes. Donors are drawn from reliable originals or trusted anchor examples with $y \neq q$; accepted vaccinations are additional views. Each response loss averages only its accepted set, and an empty set has loss zero. These routes keep the original, corrected, and neutralized supervision disjoint.

D.7 Memory Generation, Maturation, and Activation

Generation and task-end staging. Generation starts from the receptor selected by local validation for a candidate target q. Its antibody $\mathsf{Ab}_{r,q}$ contains the direction and imprint estimated from $\mathcal{S}^{\mathrm{fit}}_q$. Its

admission evidence is computed from $\mathcal{S}_q^{\mathrm{val}}$ and content-matched clean references, using the metrics in Eq. (22). For each selected antibody, define

$$G_{t,r,q} = \mathbb{I}[|\mathcal{S}_q^{\mathrm{val}}| \geq m_{\mathrm{mem}}]\,\mathbb{I}[\mathrm{AUC}_{r,q} \geq \eta_{\mathrm{auc}}] \cdot \mathbb{I}[E_{r,q} \geq \eta_{\mathrm{eff}}]\,\mathbb{I}[R_{r,q} \geq \eta_{\mathrm{rec}}]. \tag{28}$$

$G_{t,r,q}$ is the binary admission gate for a memory proposal. The first factor requires at least m_{mem} held-out validation seeds (20 by default). The other factors require that the proposed pattern rank seeds above clean examples, exhibit sufficient typical separation, and retain enough seed recall at the clean-reference operating point. These are the local validation tests reused as memory evidence; generation adds a minimum sample-support requirement, whereas maturation will add distinct-task support. A value of 1 admits the antibody for memory matching; a value of 0 allows neither a new cell nor a task-support increment. An antibody failing only the memory support requirement may still guide the current local response.

Write an admitted proposal as $\mathsf{Ab}_t = (a_t, b_t^{\mathrm{resp}})$ with receptor type $r_t \in \mathcal{R}$. Thus r_t identifies the current proposal's receptor, whereas r_j identifies the receptor stored in cell j; $r_j = r_t$ restricts memory matching to proposals and cells that use the same observation mechanism. It is routed to the matching step below: an unmatched proposal generates a pending cell, while a matching proposal supplies new evidence for an existing cell. All changes are staged until task end, so the bank used for current screening stays frozen. Neither samples selected by an antibody's own affinity nor its generated response views are counted in $\mathcal{S}_q^{\mathrm{val}}$. Thus an antibody cannot create its own admission or maturation evidence through repeated activation.

Matching and consolidation. Each proposal admitted by Eq. (28) is compared with the pending or mature cells produced by earlier tasks, using only cells from the same receptor family and the same descriptor shape. Their coordinate conventions are fixed: image coordinates for spatial directions, DCT coordinates for frequency directions, and the prescribed pooled grid for model-conditioned directions. For unit directions, matching uses

$$j^\star = \arg\max_{j: r_j = r_t} \langle a_t, a_j \rangle, \qquad \kappa_{t,j^\star} = \langle a_t, a_{j^\star} \rangle. \tag{29}$$

Here $j^\star$ identifies the closest stored pattern and $\kappa_{t,j^\star} \in [-1, 1]$ measures directional agreement. A score near 1 indicates aligned patterns, 0 orthogonal patterns, and -1 opposite patterns. Same-family comparison does not by itself establish that the antigens are identical: the stored direction must also pass the current seed–clean separation tests in Appendix D.5.

If no compatible cell exists, or the best agreement is below γ, the proposal initializes a pending cell with $\mathsf{Ab}_j = \mathsf{Ab}_t$, $n_j = 1$, $\bar{\kappa}_j = 1$, and $\chi_j = 1$. The initial $\bar{\kappa}_j = 1$ is a neutral initialization for the running minimum, not evidence of recurrence; $n_j = 1$ prevents premature maturity. For a matched proposal with $\kappa_{t,j^\star} \geq \gamma$, the tentative update is

$$\begin{aligned} a_{j^\star}^+ &= \mathrm{Norm}(n_{j^\star} a_{j^\star} + a_t), \quad (b_{j^\star}^{\mathrm{resp}})^+ = b_t^{\mathrm{resp}}, \\ n_{j^\star}^+ &= n_{j^\star} + 1, \qquad\qquad \bar{\kappa}_{j^\star}^+ = \min(\bar{\kappa}_{j^\star}, \kappa_{t,j^\star}), \\ \chi_{j^\star}^+ &= \chi_{j^\star} \wedge \chi_{t,j^\star}. \end{aligned} \tag{30}$$

The superscript $+$ denotes a proposed new state, and $\wedge$ is logical conjunction. $\chi_{t,j} = 1$ requires both the stored direction and the tentatively consolidated direction to pass the current validation criteria; otherwise the update is rejected and no task support is added. This checks the detector that would actually be stored, rather than assuming that averaging two validated directions preserves separation. Each cell receives at most one accepted consolidation per task, selecting the strongest validated proposal if several match.

The update weights the existing direction by its accumulated task support. It is a normalized running-direction update, not an exact mean of all historical vectors. The latest independently fitted physical imprint is retained; it must still pass current response checks when reused. n_j therefore counts supporting tasks, $\bar{\kappa}_j$ records the weakest accepted cross-task match, and χ_j records accepted separation evidence. The binary rule in Eq. (12) turns these records into recall eligibility. A cell first supported by task 1 remains pending; a validated match in task 2 can make it mature at that task's end; it becomes available for recall from task 3 onward.

Current-task admission and sample activation. The cells made eligible by the preceding consolidation step now supply the recalled detectors and response imprints. At a new task, each such mature cell is considered under every current label $q \in \mathcal{Y}_t$; q is not retrieved from a historical target label. Current anchors rebuild class centers and scales. Spatial and frequency cells use their stored direction in the fixed receptor coordinates, with scores restandardized against current clean class-q anchors. Their physical imprints retain their original pixel or DCT units. Finite descriptors and sufficient clean references are required. A model-conditioned cell additionally must separate current target-q validation seeds from their content-matched clean references under the new heating expert. If such evidence is unavailable or the test fails, the cell is excluded for that label; spatial and frequency recall can still operate without a new local seed set.

The admitted cells constitute $\mathcal{J}_{t,q}^{\mathrm{mem}}$. Together with any current local antibody they form $\mathcal{J}_{t,q}$, which is fixed before Eq. (23) estimates the threshold for their maximum score. A cell can be mature but excluded from this set, or included but inactive on an individual sample. Equation (13) distinguishes these cases: activation requires current admission, semantic agreement, and an above-threshold match to that cell.

A flagged sample is assigned to the strongest acting antibody for response. The label supported by the semantic reference is retained, and every edit must satisfy Appendix D.6; a failed edit leads to quarantine. These screened or edited samples never create new task support for the memory that selected them. A bounded bank prioritizes mature cells, then larger task support, then stronger validation evidence. Lack of activation in a new task does not erase a mature cell: it remains dormant until compatible evidence is encountered.

D.8 Default Implementation Values

Table 8 lists the fixed values used in our implementation. They are selected before attack-specific evaluation and are shared across attack types; dataset-specific training schedules are reported with the experimental protocol.

Table 8: Default hyperparameters for the immune modules.

Hyperparameter	Value
LoRA rank / scale $(r_{\mathrm{L}}, \alpha_{\mathrm{L}})$	$8/16$
Seed fitting fraction / minimum seed support	$0.6/6$
Heating epochs / width / learning rate	$100/64/0.1$
Spatial support fraction / frequency budget	$0.01/16$
Direct response strength grid $\mathcal{B}$	$\{1.0\}$
Validation cutoffs $(\eta_{\mathrm{auc}}, \eta_{\mathrm{eff}}, \eta_{\mathrm{rec}})$	$(0.65, 0.5, 0.1)$
Calibration false-positive level α	0.05
Response weights $(\lambda_{\mathrm{c}}, \lambda_n, \lambda_v)$	$(0.5, 0.5, 0.5)$
Semantic cutoffs $(\alpha_{\mathrm{lbl}}, \alpha_{\mathrm{sem}})$	$(0.05, 0.05)$
Semantic drift / minimum affinity change	$0.05/0.01$
Memory task support / merge similarity	$2/0.8$
Memory validation support m_{mem} / total cell capacity	$20/8$

E Additional Experimental Results

E.1 Datasets and Attack Configurations

Benchmarks and task construction. We evaluate class-disjoint Task-IL streams constructed from CIFAR-10 Krizhevsky & Hinton (2009), CIFAR-100, and TinyImageNet Le & Yang (2015). CIFAR-10 and CIFAR-100 each contain 50,000 training images and 10,000 test images at 32×32 resolution, covering 10 and 100 classes, respectively. TinyImageNet contains 200 classes with 500 training images per class; its 10,000 labeled validation images are used as the test set, and images have 64×64 resolution. Table 9 summarizes the continual partition. CIFAR-10 is divided into five 2-class tasks, whereas CIFAR-100 and TinyImageNet are divided into ten tasks containing 10 and 20 classes, respectively. The class order is fixed, the same stream is used by all methods within a

Table 9: Dataset and Task-IL stream configurations. "Train/test" gives the number of images before trusted training subsets are removed.

Dataset	Classes	Resolution	Train/test	Tasks	Classes/task
CIFAR-10	10	32×32	50k/10k	5	2
CIFAR-100	100	32×32	50k/10k	10	10
TinyImageNet	200	64×64	100k/10k	10	20

benchmark, and every task is visited once with multiple training epochs. Task identity is supplied at evaluation and predictions are restricted to the corresponding task classes.

Preprocessing. At the dataset interface, training uses a random crop with four-pixel padding, random horizontal flipping, conversion to a tensor, and dataset-specific normalization; test images use deterministic tensor conversion and the same normalization. In the architecture-controlled comparison, the native-resolution tensor is resized by bicubic interpolation to 224×224 before entering the frozen CLIP ViT-B/16. Native-protocol continual learners retain the input pipeline required by their original architecture. Trigger construction is performed in the native dataset coordinates so that its nominal spatial or warping parameters do not change with the CLIP input resize.

Trusted splits. Trusted examples are sampled only from the training split and removed from the untrusted stream. LoRA-I-BAU, LoRA-PDB, and LoRA-LMR reserve two disjoint sets of 20 examples per class. Under the two-split ISN protocol, each class provides 20 anchors and 20 calibration examples. Cross-fitted anchor scores support clean validation and score standardization; calibration examples determine only the final thresholds. Held-out antigen seeds are drawn separately from the untrusted stream. No trusted example is used for evaluation.

Attack overview. We evaluate clean training and three targeted attacks with complementary threat surfaces: the localized patch attack BadNets Gu et al. (2019), the distributed geometric attack WaNet Nguyen & Tran (2021), and the continual training-level Latent Task Backdoor (LTB) Guo et al. (2025). ISN receives no attack name, trigger parameters, target label, or poisoned-sample indicator. BadNets and WaNet are instantiated independently on every task. In each case, 10% of the untrusted current-task training examples are selected for poisoning and assigned to local class 0. The corresponding global target is therefore the first class of the current task. At evaluation, originally target-class examples are excluded from the attack cohort.

BadNets. For a native image of height H and width W, BadNets places a white square in the bottom-right corner with side length $\max\{3, \lfloor 0.1 \min(H, W) \rfloor\}$. The trigger is applied dynamically to the selected training examples, whose observed labels are replaced by the current task's target label. The same patch is applied to non-target test examples when computing ASR.

WaNet. WaNet generates a smooth elastic warping field from a 4×4 control grid. We use warping strength 0.5, grid rescaling 1.0, and cross ratio 2. The base grid and the poisoned/cross-sample membership are generated from a fixed attack seed, while stochastic data augmentation and cross-noise remain dynamic across training visits. Poisoned warped examples receive the current task's target label; cross examples retain their original labels. This construction keeps the trigger visually distributed and avoids reducing the evaluation to a fixed local patch.

Latent Task Backdoor. LTB is implemented by its attacker-specific training wrapper rather than by the data-layer poisoner used for BadNets and WaNet. The first task is attacked with global target class 0, and the first non-target class in that task is used as the default source. A stable 5% subset of that source-class pool supplies triggered examples. After two clean warm-up epochs, the wrapper uses the full attacked-task loader to estimate feature importance, selects the top 5% of classifier-input features by Fisher score, and repeatedly enforces the attacker's fixed embedding budget while optimizing its target loss. The trigger is a 5×5 black square in the bottom-right corner. We evaluate the attacked task after the complete task sequence and report persistence ASR; the 5% LTB rate refers to source-pool membership and is therefore not directly equivalent to the 10% per-task data-layer poison rate above.

Comparisons. We conduct an architecture-controlled comparison in a unified frozen CLIP ViT-B/16 LoRA framework, including Vanilla LoRA and LoRA adaptations of AACL, ABL, I-BAU, PDB,

and LMR. We also compare representative continual learners under their native architectures and optimization protocols.

Architecture-controlled comparison. For Table 2, every method uses a frozen CLIP ViT-B/16, task-specific QKV LoRA with rank $r = 8$ and scale $\alpha = 16$, and a cosine classifier. Each task is trained for 10 epochs with AdamW, learning rate 10^{-2}, and batch size 32. We adapt AACL, ABL, I-BAU, PDB, and LMR to this common framework, denoting the resulting methods LoRA-AACL, LoRA-ABL, LoRA-I-BAU, LoRA-PDB, and LoRA-LMR. AACL was designed for continual learning; the remaining defenses are applied task-wise. Vanilla LoRA is the undefended counterpart.

Native-protocol comparison. For Table 3, each learner retains its original backbone and optimization setting. DER++ uses ResNet-18 and a 500-sample buffer. L2P, CODA-Prompt, InfLoRA, and EWC-LoRAZheng et al. (2026) use ImageNet-pretrained with ViT-B/16; and KeepLoRALuo et al. (2026) uses frozen CLIP ViT-B/16. All implementations use the Mammoth codebase, and each run uses a single GPU.

Clean accuracy and attack success. We report final Task-IL clean accuracy (CA) and clean-conditioned Task-IL ASR. For task t, let $\hat{y}_t(x) = \arg\max_{c \in \mathcal{Y}_t} [f_t(x)]_c$ denote the predicted class, and let $\mathcal{E}_t$ index paired clean and triggered inputs $(x_i, \tilde{x}_i)$ with true labels $y_i \neq y_t^\star$. The clean-conditioned ASR on this non-target cohort is

$$\mathrm{ASR}_{\mathrm{cc},t} = \frac{\sum_{i \in \mathcal{E}_t} \mathbb{I}[\hat{y}_t(\tilde{x}_i) = y_t^\star] \mathbb{I}[\hat{y}_t(x_i) = y_i]}{\sum_{i \in \mathcal{E}_t} \mathbb{I}[\hat{y}_t(x_i) = y_i]} \times 100\%, \tag{31}$$

where $y_t^\star$ is the target label in the ground-truth task. Both predictions are restricted to that task. Final BadNets/WaNet ASR pools the numerator and denominator over evaluated tasks before taking their ratio. LTB reports the attacked-task cohort after the complete sequence. A zero denominator is reported as unavailable, not zero ASR. We report the mean and standard deviation over three random seeds.

E.2 Additional Evidence for Antigen-to-Antibody Transfer

Evaluation question and protocol. We examine whether antigen evidence exposed by semantically conflicting, cross-source poisoned samples transfers to label-consistent same-source backdoors. We report three receptor-specific probes on the five-task CIFAR-10 stream: a spatial pixel receptor for BadNets, a model-conditioned activation–gradient contribution receptor for WaNet, and a YCrCb-DCT frequency receptor for FTrojan. Within each probe, semantic-conflict seeds are used to fit the antibody, while held-out antigen seeds are excluded from antibody-prototype fitting. Trusted clean examples provide class-conditioned self references, development evidence, and threshold calibration. Poison identities and original labels are used only after these choices are frozen, to form the evaluation cohorts and compute the reported metrics. These experiments test the transfer mechanism in three distinct observation spaces; they are not an evaluation of automatic attack-family classification.

Reading the visual evidence. For each attack, the antigen panel displays cohort-level signed self-residual maps for clean candidates, held-out antigens, the frozen antibody, and routed same-source candidates. Its color scale is shared within an attack, but residual magnitudes should not be compared across receptors. The affinity panel shows scores after subtraction of each target's frozen threshold; therefore, zero is the operational decision boundary even when cohorts from multiple tasks are pooled. Candidate-set AUROC measures post-hoc ranking of same-source backdoors against clean candidates. By contrast, empirical clean FPR and full-stream recall are measured at the independently frozen threshold; full-stream recall also reflects which poisoned examples entered the candidate set. The mechanism panel compares the same-source AUROC of the self-residual antibody with controls fitted from defender-visible data.

Localized spatial evidence under BadNets. Figure 4 shows that the held-out antigen and same-source cohorts have a strong lower-right signed response on the learned spatial support, whereas the clean cohort has little corresponding response. The fit-seed antibody retains this localized structure. More importantly, the frozen affinity score separates previously unseen same-source candidates from clean candidates, reaching an AUROC of 0.976. At the frozen operating point, the empirical clean FPR is 7.2% and full-stream same-source recall is 93.2%. The self-residual antibody substantially exceeds both the raw pixel template (0.558 AUROC) and a random antibody (0.489 AUROC).

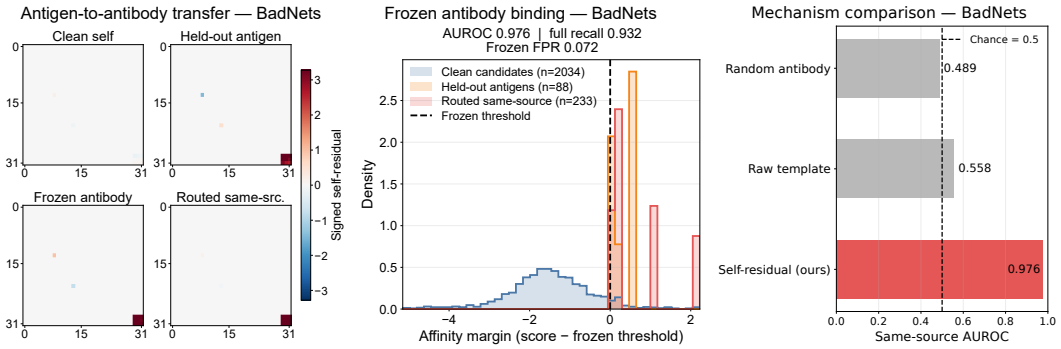


Figure 4: **Spatial antigen transfer under BadNets.** Left: support-restricted signed residual maps. Middle: threshold-aligned affinity distributions. Right: same-source AUROC against a random antibody and a raw-pixel template.

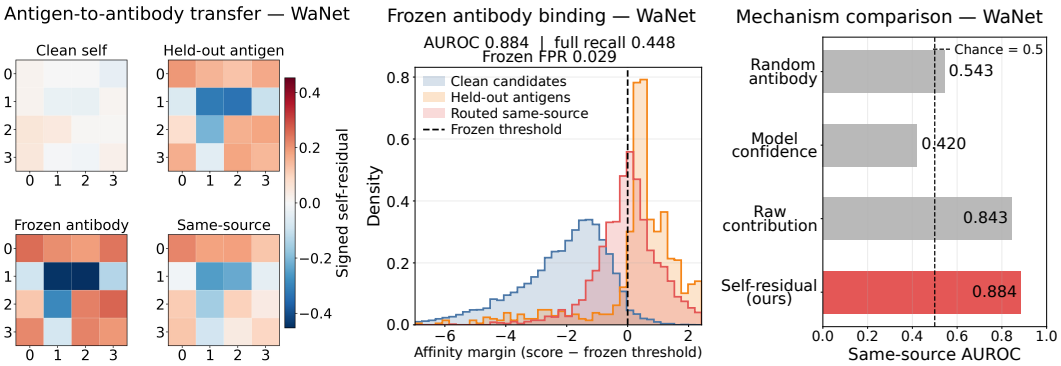


Figure 5: **Model-conditioned antigen transfer under WaNet.** Left: 4×4 activation–gradient contribution residuals, rather than an estimated input-space warp field. Middle: threshold-aligned affinity distributions. Right: mechanism comparisons.

Thus, the result is not explained merely by averaging poisoned images or by choosing an arbitrary direction in pixel space: removing expected class-conditioned content is central to the observed transfer.

Distributed model-conditioned evidence under WaNet. WaNet does not yield the localized pixel pattern seen in BadNets. Figure 5 therefore visualizes a different receptor: the spatial contribution of intermediate activations to the expert's target margin. After subtracting the content-class clean response, held-out antigens and routed same-source candidates exhibit related signed patterns on the 4×4 contribution grid, while clean candidates have a weaker residual. The displayed maps aggregate all five validated task antibodies using weights derived from held-out validation evidence, rather than selecting a task using same-source test performance. The resulting antibody reaches 0.884 same-source AUROC, compared with 0.843 for the raw contribution template, 0.543 for a random direction, and 0.420 for model confidence. At its frozen threshold, it operates at 2.9% empirical clean FPR and 44.8% full-stream recall. The affinity distributions accordingly support useful, though not complete, separation for this distributed distortion. The gain over the raw-contribution control isolates the benefit of removing normal class-conditioned model response rather than relying on the expert's activation pattern alone.

Sparse frequency evidence under FTrojan. Figure 6 places the comparison in DCT space. Held-out antigens and same-source candidates exhibit stronger signed responses than clean candidates at the sparse coordinates retained by the frozen antibody. Because all four maps are displayed on this learned support, their common locations should be interpreted as a controlled comparison of response magnitudes, not as an independent test of support recovery. The separate transfer evidence is the held-out affinity response and the post-freeze separation of same-source from clean candidates: candidate-set AUROC is 1.000, while the empirical clean FPR at the frozen threshold is 6.7% and

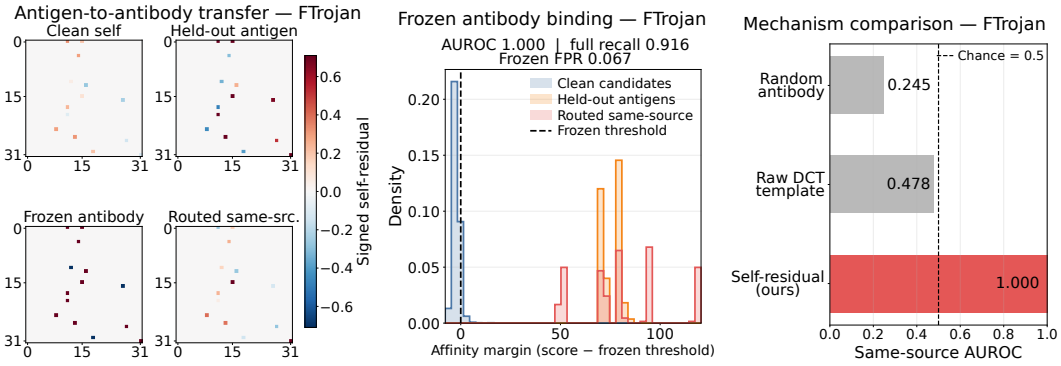


Figure 6: **Frequency antigen transfer under FTrojan.** Left: signed YCrCb-DCT residuals on the learned frequency support. Middle: threshold-aligned affinity distributions. Right: same-source AUROC against random and raw DCT-template controls.

full-stream recall is 91.6%. The raw DCT template reaches only 0.478 AUROC, compared with 1.000 for the self-residual antibody. The contrast indicates that sparse frequency energy alone is insufficient; its deviation from the expected clean class response carries the discriminative signal in this setting.

Cross-attack interpretation. The three probes expose different observable forms of the same mechanism: a localized pixel residual, a distributed model-conditioned contribution residual, and a sparse frequency residual. In each case, an antibody fitted from semantic-conflict evidence produces an affinity response on hidden, label-consistent same-source candidates. The negative controls further show why antigen extraction must account for clean self response rather than memorize raw inputs or raw receptor outputs. These are receptor-matched mechanism experiments, so the results support cross-source-to-same-source transfer across the three evaluated attack constructions without implying that the figures alone establish an automatic three-way receptor-selection policy for arbitrary unknown attacks.

E.3 Fine-grained Ablation Study

The module-level ablation in the main text examines the overall contributions of immune perception, response, and memory. Here, we further investigate which receptor pathways support different attacks and how individual response operators contribute to backdoor suppression.

Evaluation protocol. We evaluate the five-task CIFAR-10 stream and report final Task-IL clean accuracy (CA) and clean-conditioned attack success rate (ASR). All available entries in Table 10 are means and sample standard deviations over two random seeds, 42 and 2042. The three receptor ablations are evaluated under both BadNets and WaNet, whereas response and memory ablations are evaluated under BadNets. Unavailable attack–variant combinations are marked as `n/a`. All comparisons below use the Full result for the corresponding attack in this table.

Ablation definitions. The *w/o Spatial*, *w/o Act.–Grad.*, and *w/o Frequency* variants disable the spatial, model-conditioned activation–gradient, and frequency receptor pathways, respectively, while retaining semantic screening and the remaining receptor families. These interventions remove the corresponding pathway from the defense pipeline, so their effects can propagate to antibody selection, response construction, and memory reuse. The *w/o Neutralization* and *w/o Vaccination* variants disable their respective response operators while retaining the other training interventions. The *w/o Response* variant bypasses all defensive training interventions, including correction, quarantine, neutralization, and vaccination, and trains on the original observed labels. Finally, *w/o Memory* disables cross-task antibody retention and recall while retaining current-task antibody construction and response.

Attack-dependent receptor contributions. The spatial and activation–gradient pathways exhibit different contributions across the two attacks. Under BadNets, removing the spatial pathway increases ASR from 0.07% to 24.02%, a rise of 23.95 percentage points, while CA changes by only $+0.04$ percentage points. This result is consistent with the spatial receptor's role in capturing the

Table 10: Fine-grained ablation on the five-task CIFAR-10 stream. Results are final Task-IL CA and clean-conditioned ASR (%), reported as mean ± sample standard deviation over seeds 42 and 2042. Bold indicates the best mean in each column. `n/a` denotes an unevaluated condition.

Variant	BadNets		WaNet	
	CA ↑	ASR ↓	CA ↑	ASR ↓
Full	99.59±0.08	0.07±0.04	**99.49**±0.02	**0.65**±0.52
w/o Spatial	99.63±0.04	24.02±7.28	99.50±0.01	**1.10**±0.11
w/o Act.–Grad.	99.57±0.12	0.03±0.01	99.30±0.00	13.18±1.66
w/o Frequency	99.60±0.04	5.50±7.75	99.36±0.02	8.22±2.54
w/o Neutralization	99.59±0.01	14.02±3.56	n/a	n/a
w/o Vaccination	99.61±0.03	1.77±0.31	n/a	n/a
w/o Response	**99.71**±0.00	94.00±0.57	n/a	n/a
w/o Memory	99.65±0.06	9.30±12.83	n/a	n/a

recurring localized residual associated with the patch trigger. Under WaNet, the largest receptor-ablation degradation instead occurs when the activation–gradient pathway is removed: ASR increases from 0.65% to 13.18%, while CA decreases from 99.49% to 99.30%. This supports the usefulness of model-conditioned contribution evidence for the distributed warping attack.

The reverse comparisons show that these contributions are not uniform across attacks. Removing the activation–gradient pathway yields a BadNets ASR of 0.03%, compared with 0.07% for Full. Likewise, removing the spatial pathway yields a WaNet ASR of 1.10%, compared with 0.65% for Full. Thus, Full does not achieve the lowest mean ASR in every individual column, and the results do not establish that every receptor is independently necessary for every attack. Rather, they show that the defense relies on different observation mechanisms under different attack constructions. The small differences between near-zero BadNets ASRs should also be interpreted cautiously given the two-seed evaluation.

Contribution of the frequency pathway. Removing the frequency pathway increases mean ASR to 5.50% under BadNets and 8.22% under WaNet, corresponding to increases of 5.43 and 7.57 percentage points over Full. CA remains close to the corresponding Full result, at 99.60% and 99.36%, respectively. These results indicate a beneficial aggregate effect of retaining the frequency pathway in the evaluated pipeline. However, the BadNets standard deviation of 7.75 percentage points indicates substantial variation across seeds. Moreover, final ASR measures the outcome of the complete training process: it does not identify whether the improvement arises from direct frequency-based detections, changes in receptor selection, or downstream response and memory effects. The receptor-specific transfer experiments therefore provide a separate form of evidence from this end-to-end ablation.

Neutralization and vaccination. Both response operators contribute to BadNets suppression, with a larger observed effect for neutralization. Removing neutralization increases ASR from 0.07% to 14.02%, while mean CA remains unchanged at 99.59%. Removing vaccination increases ASR to 1.77% with 99.61% CA. These changes are consistent with their different roles: neutralization reduces antigen expression in detected label-consistent samples, whereas vaccination introduces label-preserving counterfactuals that weaken the association between the learned antigen and its apparent target. The results support retaining both operators in this setting, although their individual effects cannot be added to estimate a joint effect because they operate within the same detection and training pipeline.

From detection to training intervention. Disabling the complete response produces the largest degradation, increasing ASR to 94.00% while preserving 99.71% CA. This contrasts sharply with the low ASR of Full despite similar clean accuracy. The result shows that high clean accuracy alone does not prevent acquisition of the backdoor and that defensive evidence must influence training to suppress the shortcut. The full-response ablation is broader than either operator ablation because it also removes correction and quarantine. Consequently, its degradation should not be attributed solely to neutralization or vaccination, nor interpreted as a direct measurement of their interaction.

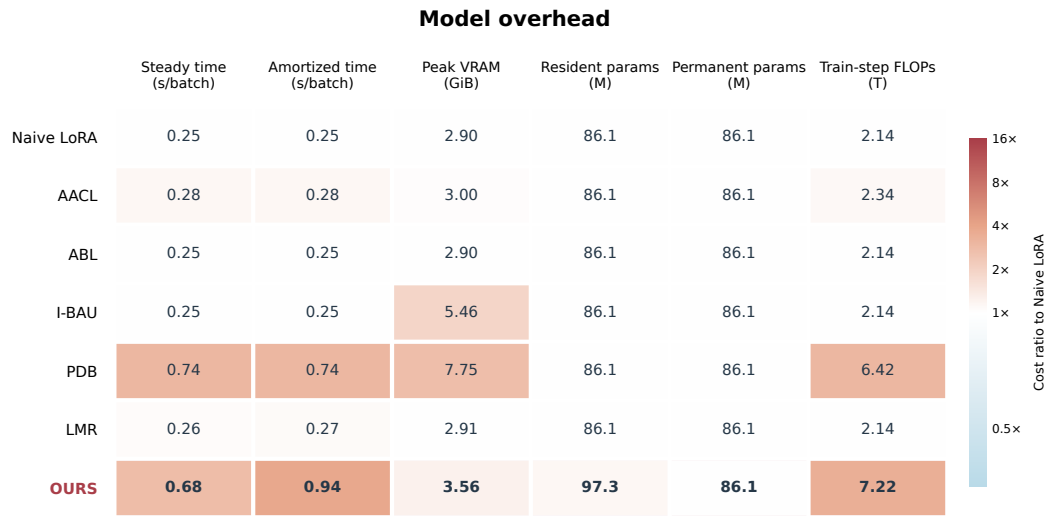


Figure 7: Task-0 overhead in the architecture-controlled CLIP–LoRA comparison on CIFAR-10/BadNets (seed 42, batch size 32). Numbers are absolute measurements; colors show the log-scaled cost ratio to Vanilla LoRA within each metric (blue: lower; red: higher).

Cross-task immune memory. Without memory, mean BadNets ASR increases from 0.07% to 9.30%, whereas CA changes from 99.59% to 99.65%. This indicates that retaining and recalling antibodies can improve final robustness beyond current-task processing alone. However, the ASR standard deviation of 12.83 percentage points reveals substantial seed sensitivity. The evidence therefore supports an average benefit under the evaluated runs, rather than a uniform improvement across random initializations. Final aggregate ASR also does not reveal when recalled antibodies become useful; the task-wise memory analysis in the main text addresses that temporal behavior.

Clean accuracy and scope of the conclusions. Across the reported variants, mean CA ranges from 99.57% to 99.71% under BadNets and from 99.30% to 99.52% under WaNet. The much larger changes in ASR therefore occur alongside relatively small changes in clean classification performance. Together, these ablations support attack-dependent receptor contributions and the importance of converting antibody evidence into training interventions. They quantify the effects of removing components from the complete pipeline, rather than independent or additive component effects. All conclusions are restricted to the evaluated attack–variant combinations; response and memory effects under WaNet are not assessed here.

E.4 Model Overhead

Profiling protocol. Figure 7 compares the methods on the same frozen CLIP ViT-B/16 task-LoRA backbone, using one GPU and the first task of the BadNets CIFAR-10 stream. For each method we execute the complete task-begin procedure, two warm-up batches, ten timed training batches, and then the algorithmic task-end procedure. The steady-time entry is the median duration of the timed batches. To expose costs that a per-batch profile would miss, we also report an estimated amortized time: the measured task-begin and task-end durations are divided by the number of batches in the method's configured ten-epoch Task-0 schedule and added to the steady-time median. The task-end hook is measured after the sampled updates, not after a complete task, so this estimate should not be read as observed end-to-end training time.

Time overhead. OURS requires 0.68 s per steady-state batch, compared with 0.25 s for Vanilla LoRA (approximately $2.7\times$) and 0.74 s for LoRA-PDB. The additional computation within an ordinary update is therefore substantial, but the recurring batch time remains below that of LoRA-PDB in this setting. The full Task-0 setup for OURS takes 783.45 s (13.1 min). Distributing this one-off cost over the configured training schedule contributes approximately 0.25 s per batch, increasing the estimated amortized time to 0.94 s per batch, about $3.8\times$ Vanilla LoRA. Reporting only the

Table 11: CIFAR-10 results under BadNets and WaNet. The upper block uses the frozen CLIP ViT-B/16 task-LoRA backbone (including Ours), whereas the lower block uses native backbones. ASR is clean-conditioned Task-IL attack success rate; CA is clean accuracy. Values are percentages. None reports clean-training CA from a single run; all other entries report mean ± standard deviation over three independent runs with random seeds 42, 256, and 2042. Lower ASR and higher CA are better; bold marks the best mean for each attack metric within each block.

Method	None	BadNets		WaNet	
	CA	ASR	CA	ASR	CA
Vanilla LoRA	99.20	95.34±0.76	99.22±0.13	63.52±1.99	97.72±0.24
LoRA-AACL	98.96	**0.00**±0.00	97.62±0.32	54.82±10.07	98.42±0.41
LoRA-ABL	77.53	47.52±1.14	81.59±2.65	10.70±4.65	76.26±2.67
LoRA-I-BAU	98.87	95.68±0.42	95.67±0.45	67.37±3.50	89.38±0.97
LoRA-PDB	93.40	0.06±0.03	93.21±0.47	0.99±0.08	93.84±0.09
LoRA-LMR	99.31	84.73±9.62	99.03±0.23	3.94±2.57	99.36±0.12
Ours	99.61	0.07±0.04	**99.59**±0.08	**0.65**±0.52	**99.49**±0.02
DER++	84.28	**70.69**±1.50	85.54±1.57	**14.50**±3.53	85.80±0.58
L2P	99.60	88.56±8.23	99.13±0.14	36.65±6.13	96.11±2.14
CODA-Prompt	99.71	89.81±0.84	99.07±0.12	41.62±0.23	97.14±0.09
InfLoRA	99.66	76.88±3.25	99.06±0.13	41.88±1.10	98.16±0.03
EWC-LoRA	99.67	95.12±0.36	**99.49**±0.05	58.78±2.11	**98.23**±0.19
KeepLoRA	99.05	95.44±1.00	**99.49**±0.16	57.64±2.75	97.88±0.21

steady-time column would consequently understate the practical training cost of immune-response construction, while reporting setup alone would obscure the recurring cost.

Memory and model size. OURS reaches 3.56 GiB peak allocated GPU memory, approximately 23% above Vanilla LoRA's 2.90 GiB, but below LoRA-I-BAU's 5.46 GiB and LoRA-PDB's 7.75 GiB. Its resident parameter count after task setup is 97.3 M, versus 86.1 M for Vanilla LoRA, reflecting temporary task modules used during profiling. After task end, the permanent parameter count is 86.1 M for both methods. This equality concerns parameters only: it does not by itself establish identical optimizer-state, buffer, or later-task immune-memory footprints.

Computation and scope. The profiled training step of OURS uses 7.22 TFLOPs, versus 2.14 TFLOPs for Vanilla LoRA and 6.42 TFLOPs for LoRA-PDB. FLOPs count only operators covered by the PyTorch profiler and omit the task-boundary procedures; they are a lower bound rather than a complete computational budget. The different ordering of FLOPs and elapsed batch time for OURS and LoRA-PDB also cautions against using FLOPs alone as a runtime proxy. Finally, this is a single-seed Task-0 profile, not a full-stream average: future tasks can have different setup and memory costs as experts and immune memory accumulate.

E.5 Additional Results on CIFAR-10

Table 11 extends the BadNets and WaNet evaluation to the five-task CIFAR-10 stream. The upper block compares defenses under the same frozen CLIP ViT-B/16 task-LoRA backbone, while the lower block provides context from continual learners using their native backbones. All attacked results are reported as mean ± standard deviation over seeds 42, 256, and 2042. The None column contains single-run clean-training controls.

Architecture-controlled comparison. ISN combines low mean ASR with high clean accuracy under both attacks. For BadNets, it achieves $0.07 \pm 0.04\%$ ASR and $99.59 \pm 0.08\%$ CA, compared with $95.34 \pm 0.76\%$ ASR and $99.22 \pm 0.13\%$ CA for Vanilla LoRA. For WaNet, ISN achieves $0.65 \pm 0.52\%$ ASR and $99.49 \pm 0.02\%$ CA, compared with $63.52 \pm 1.99\%$ ASR and $97.72 \pm 0.24\%$ CA for Vanilla LoRA. These correspond to mean ASR reductions of 95.27 and 62.87 percentage points, respectively. ISN has the highest mean CA under both attacks and the lowest mean WaNet ASR within the matched-backbone block.

Attack suppression and clean-accuracy trade-offs. LoRA-AACL reports the lowest BadNets ASR ($0.00 \pm 0.00\%$), whereas ISN's mean ASR is 0.07%. This small difference accompanies a 1.97-percentage-point CA advantage for ISN (99.59% vs. 97.62%). LoRA-AACL retains $54.82 \pm 10.07\%$ ASR under WaNet, indicating that its strong BadNets result does not extend to this attack in the evaluated setting. LoRA-PDB also attains low ASR under both attacks (0.06% and 0.99%). Its mean BadNets ASR is slightly lower than ISN's, but its CA is lower by 6.38 and 5.65 percentage points under BadNets and WaNet, respectively. LoRA-LMR preserves high CA (99.03% and 99.36%), yet retains 84.73% BadNets ASR and 3.94% WaNet ASR. These comparisons support ISN's favorable observed balance between attack suppression and clean accuracy across the two attack constructions.

Comparison with native continual learners. The native-backbone methods retain mean ASR ranging from 70.69% to 95.44% under BadNets and from 14.50% to 58.78% under WaNet. DER++ has the lowest ASR in this block for both attacks, but its CA is 85.54% and 85.80%, respectively. The remaining native learners attain at least 99.06% CA under BadNets while retaining at least 76.88% ASR, illustrating that high Task-IL clean accuracy can coexist with substantial backdoor vulnerability. ISN has lower mean ASR and higher mean CA than all methods in this block under both attacks. Because these learners use different backbones and optimization protocols, this comparison provides benchmark context; the upper block supplies the architecture-controlled comparison of defense performance.

Interpretation of the reported statistics. The None entries provide descriptive clean-training references rather than three-run estimates, so small differences between None and the attacked CA means should not be interpreted as statistically established gains or losses. Likewise, the reported means and standard deviations summarize three runs and do not by themselves establish significance for close rankings. In particular, ISN's WaNet ASR varies more across runs than its BadNets ASR (0.52 vs. 0.04 percentage-point standard deviation), despite a low mean in both cases. The results therefore support low observed ASR with high CA for these two evaluated attacks, while retaining the uncertainty visible in the per-attack statistics.

E.6 Hyperparameter Sensitivity

Evaluation question and protocol. We examine how the operating choices in immune perception, response, and memory affect the final accuracy–security trade-off. We conduct a one-at-a-time sweep on the five-task CIFAR-10 stream under BadNets and WaNet, with a poisoning rate of 10%. This controlled implementation study uses fixed training and trusted-data settings throughout the sweep. We vary four quantities: a shared level α, the antibody-validation recall cutoff η_{rec}, the per-batch vaccination budget B_{vac}, and the minimum number of distinct supporting tasks $n_{\min}$ for memory maturation. In this sweep, the semantic cutoffs are tied to the calibration level, $\alpha_{\mathrm{lbl}} = \alpha_{\mathrm{sem}} = \alpha$; thus, the first sweep changes both semantic screening and affinity calibration, rather than only the final screening threshold. The tested values are $\alpha \in \{0.01, 0.025, 0.05, 0.10\}$, $\eta_{\mathrm{rec}} \in \{0.05, 0.10, 0.20, 0.30\}$, $B_{\mathrm{vac}} \in \{4, 8, 16, 32\}$, and $n_{\min} \in \{2, 3, 4\}$. Unless varied, these quantities are fixed to $(0.05, 0.10, 32, 2)$. We report final Task-IL CA and clean-conditioned Task-IL ASR, pooling the ASR numerator and denominator across tasks. Each configuration uses training seed 42; the WaNet attack seed is also fixed to 42. The default run is shared across the four sweeps, not counted as an independent replicate.

Shared semantic and calibration level. Figure 8(a) reveals a finite-reference boundary rather than uniform insensitivity to α. At $\alpha = 0.01$ and 0.025, ASR remains 94.40–94.42% under BadNets and 62.01–62.02% under WaNet. With 20 anchors per class, the empirical semantic compatibility score satisfies

$$p_q(x) \geq \frac{1}{|\mathcal{A}_q^{\mathrm{anchor}}| + 1} = \frac{1}{21} \approx 0.0476. \tag{32}$$

Consequently, neither smaller level can satisfy the observed-label rejection test $p_q(x) \leq \alpha_{\mathrm{lbl}}$, preventing semantic-conflict seed discovery. Consistently, these runs record no semantic-conflict seeds in any of the five tasks. The finite-sample calibration operator likewise returns $+\infty$ at these levels when given 20 scores. Increasing the shared level to 0.05 yields ASRs of 0.02% and 1.51% for BadNets and WaNet; at 0.10, the corresponding values are 0.08% and 0.14%. These observations show that the nominal level must be compatible with the available reference-set resolution. They

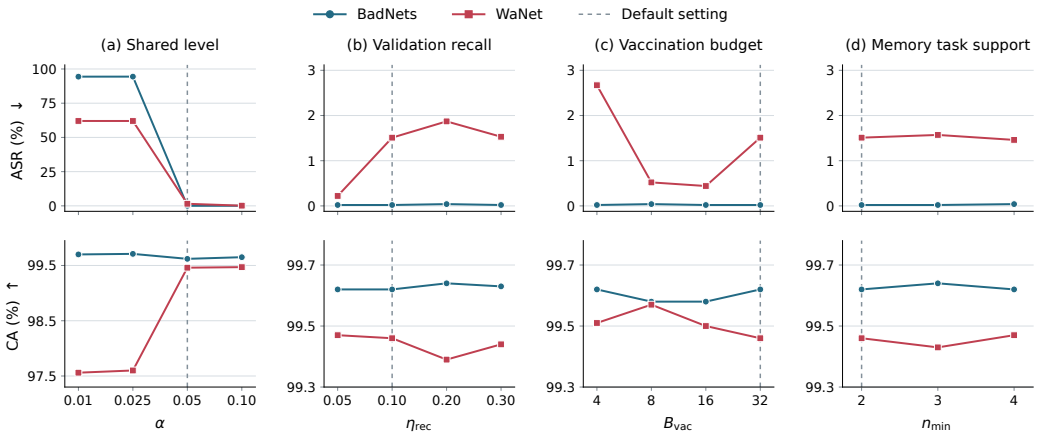


Figure 8: **Hyperparameter sensitivity on CIFAR-10.** Columns vary the shared semantic/calibration level α, antibody-validation recall cutoff η_{rec}, vaccination budget B_{vac}, and memory task-support threshold $n_{\min}$, respectively. The top row reports clean-conditioned Task-IL ASR (↓); the bottom row reports Task-IL CA (↑). Vertical dashed lines mark the default settings. Each point is a single completed run with seed 42, so no error bars are shown. Tested values are equally spaced for readability; connecting lines are visual guides. Y-axis ranges differ across panels, and CA axes are truncated.

do not identify an attack-independent optimum or establish a formal false-positive guarantee for the semantic score.

Antibody-validation recall. The cutoff η_{rec} controls how much held-out antigen evidence must remain above the clean-reference threshold before an antibody passes validation. We vary this cutoff while holding the ranking and effect-size requirements fixed. Across the tested values, BadNets ASR stays within 0.02–0.04%, with CA between 99.62% and 99.64%. WaNet ASR varies from 0.22% to 1.87%, while CA remains within 99.39–99.47%. A stricter recall requirement therefore does not monotonically improve final protection: it restricts antibody admission but does not directly optimize downstream ASR. The tested settings preserve clean accuracy, although WaNet exhibits greater sensitivity in residual attack success than BadNets.

Counterfactual vaccination budget. The budget B_{vac} limits vaccination views per batch without changing their semantic-admissibility tests or loss weight. BadNets ASR remains 0.02–0.04% throughout the sweep. For WaNet, increasing the budget from 4 to 8 and 16 changes ASR from 2.67% to 0.52% and 0.44%, respectively, whereas a budget of 32 gives 1.51%. CA varies by at most 0.11 percentage points within either attack. Thus, the results support useful protection with a modest vaccination budget, but not a monotonic benefit from enlarging it. The smallest observed ASR at 16 is a single-seed outcome, not sufficient evidence for selecting a new default or attributing the increase at 32 to overfitting.

Distinct-task support for immune memory. We replace the default requirement of two supporting tasks by $n_j \geq n_{\min}$, retaining the agreement and clean-separation requirements for memory maturation. BadNets ASR is 0.02%, 0.02%, and 0.04% at $n_{\min} = 2, 3, 4$, respectively, with CA within 99.62–99.64%. The WaNet runs give ASRs of 1.51%, 1.57%, and 1.46% at $n_{\min} = 2, 3, 4$, respectively, with CAs of 99.46%, 99.43%, and 99.47%. Because memory updates are committed at task end, increasing $n_{\min}$ delays the earliest possible recall and reduces its opportunities within a five-task stream. Across the tested support thresholds, WaNet ASR varies by only 0.11 percentage points and CA by 0.04 percentage points. These results indicate limited sensitivity to the maturation requirement in this stream, but do not replace a memory-removal ablation.

Scope of the evidence. Within this controlled single-seed sweep, the validation, vaccination, and memory settings retain high clean accuracy and low residual ASR over the completed configura-

tions. The shared semantic/calibration level is the exception: settings below the empirical reference resolution prevent the seed-driven perception process from starting. These findings describe the observed operating range and its boundary, rather than statistical robustness across seeds or universal insensitivity across attacks and task streams.