Erasing Self-Supervised Learning Backdoor by Cluster Activation Masking

2.6 / 5
AVERAGE RATING

Journal Article · Artificial Intelligence, Computer Science
Reviewed August 3, 2026

QUALITY RATINGS

Technical Soundness 3/5
Core ideas and experiments are plausible and coherent, but incomplete methodological/detail specification and lack of statistical rigor weaken the technical foundation.

Novelty 3/5
Builds meaningfully on PatchSearch with Cluster Activation Masking, but conceptual differentiation is moderate and could be articulated more sharply.

Organization & Clarity 3/5
Overall IMRaD structure is present, yet some sections mix implementation with method, and equations/figures occasionally suffer from formatting and ambiguity.

Writing Quality 3/5
Generally understandable, but marred by template artifacts, typos, and some unclear sentences and notational issues.

Statistics & Significance 2/5
Reports means over multiple runs but omits variance, confidence intervals, and significance tests for key comparisons.

Reproducibility 2/5
Key hyperparameters and algorithmic steps are underspecified; code/data are promised post-acceptance with no current artifacts.

Ethics/Privacy Compliance 2/5
Work is clearly defensive, but there is no explicit discussion of dual-use risks, responsible release, or privacy considerations.

ISSUES FOUND

3 major · 21 moderate · 20 minor · 2 polish

46 issues identified, graded 1 (polish) to 5 (critical).

AT A GLANCE

- Methodology and implementation details for core components (Cluster Activation Masking, heuristic search, mask generation, stopping criteria, sampling, preprocessing) are under-specified, preventing reproducibility (I01, I02, I03, I08, I09, I10, I11, I16, I17).

Executive Summary

- Methodology and implementation details for core components (Cluster Activation Masking, heuristic search, mask generation, stopping criteria, sampling, preprocessing) are under-specified, preventing reproducibility (I01, I02, I03, I08, I09, I10, I11, I16, I17).
- Key experimental protocols and hyperparameter choices, including reimplementation of PatchSearch, tuning procedures, data splits, and multi-target settings, are ambiguously documented or missing (I10, I11, I22, I23, I24, I25).
- Reported statistics and numerical results contain internal inconsistencies, arithmetic errors, and lack uncertainty estimates, undermining confidence in observed effect sizes and claims (I04, I15, S01, S02, S03, S04).
- Multiple cross-section inconsistencies (poisoning rates, accuracy percentages, definitions of ASR and detection metrics) create confusion about the actual evaluation setup and results (X01, X02, I05, I12).
- Figure content, captions, and visual quality are frequently mismatched, truncated, or illegible, obscuring the main ideas and experimental evidence (V01–V08, V09–V13, I18, V02, V04, V06, V10, V11, V12).
- The threat model and conceptual positioning relative to prior work (especially PatchSearch) are only partially articulated, leaving the novelty and applicability of the approach under-explained (I06, I20, I21).
- Ethics, broader impacts, and artifact availability are weakly addressed, with dual-use considerations and concrete release plans missing or contingent (I13, I14).
- Writing and formatting issues (template artifacts, truncated sentences, garbled captions) reduce clarity and perceived rigor throughout the manuscript (I19, I26).

Prioritized Issues

I01

MAJOR · 4/5

Section	Methodology
Location	Equation (1) and Section IV-B/IV-C description
Evidence (≤25w)	"we randomly initialize B masks M = {mj ⊠ {0, 1}H×W }B j=1 . In particular, we set the pixel value to zero"
Problem	Mask generation procedure (shape, sampling distribution, stride, overlap) is insufficiently specified to reproduce Cluster Activation Masking.
Standard Violated	Reproducible methods require precise description of algorithmic steps and randomness sources.
Impact	Other researchers cannot reliably implement PoisonCAM's core detection mechanism, undermining validation and extension.
Confidence	High

I08

MAJOR · 4/5

Section	Methodology
Location	Section IV-C Poisonous Image Search
Evidence (≤25w)	"we adopt a heuristic search strategy following [14] to iteratively compute the poison scores of clusters."
Problem	Heuristic search procedure (iteration count, cluster removal schedule, sampling scheme) is only loosely described; no pseudo-code or exact algorithm is provided.
Standard Violated	Heuristic algorithms should be described in enough detail (or pseudocode) to allow faithful replication.
Impact	Differences in heuristic implementation could materially change which samples are identified as poisonous.
Confidence	High

I17

MODERATE · 3/5

Section	Methodology
Location	Section IV-D Poison Classifier Training
Evidence (≤25w)	"These synthesized images as the positive samples and original images in X as the negative samples form the poison classification set X'."
Problem	The labeling of synthesized versus original images and how truly poisoned originals are handled is under-explained, especially given noisy labels.
Standard Violated	Learning setup (label definition, noise model) must be described clearly for novel classifiers.
Impact	Unclear training labels and noise handling hinder understanding of why the poison classifier works and how robust it is.
Confidence	Medium

I15

MODERATE · 3/5

Section	Results
Location	Section V-A/V-B discussion
Evidence (≤25w)	"PoisonCAM achieves 2.8%, 1.8%, and 1.2% average ACC improvements... ASR reductions of 4.1%, 0.9%, and 1.0%"
Problem	Effect size is given without any uncertainty estimates; small ASR reductions (e.g., 0.9%) may be within noise.
Standard Violated	Claims of superiority should be supported by statistical analysis, especially for modest improvements.
Impact	Readers may overestimate robustness gains relative to experimental variability.
Confidence	Medium

I23

MODERATE · 3/5

Section	Evaluation
Location	Section VI Implementation Details
Evidence (≤25w)	"We employed the following parameters for PoisonCAM on ImageNet-100 as same as PatchSearch: the cluster count of l = 1000, the number of samples per cluster s = 2"
Problem	It is unclear whether these hyperparameters are optimal for both methods or tuned preferentially for PoisonCAM; tuning protocol is not specified.
Standard Violated	Fair comparison requires balanced hyperparameter tuning and explicit description of tuning procedures.
Impact	Could bias comparisons in favor of PoisonCAM if baseline is not equally optimized.
Confidence	Medium

I11

MODERATE · 3/5

Section	Evaluation
Location	Baseline description
Evidence (≤25w)	"we reimplement PatchSearch using the official code. 1 . We will release our poisoned datasets, code, and trained models once this paper is accepted."
Problem	Reimplementation details and any deviations from PatchSearch's original settings (hyperparameters, tuning, data) are not specified.
Standard Violated	Comparative evaluation requires disclosing baseline configuration to ensure fairness.
Impact	Reported superiority over PatchSearch could partly stem from unreported implementation differences.
Confidence	Medium

I10

MODERATE · 3/5

Section	Evaluation
Location	Section V-A Datasets
Evidence (≤25w)	"Following previous work [14], we adopt ImageNet-100 [49]... We also adopt STL-10 [51]"
Problem	Train/validation/test splits, use of unlabeled STL-10 data, and any preprocessing or resizing operations are not fully documented.
Standard Violated	Dataset usage and preprocessing must be thoroughly described for experimental transparency.
Impact	Differences in splits or preprocessing could explain performance variations and hinder fair comparison.
Confidence	High

I07

MODERATE · 3/5

Section	Methodology
Location	Equation (2) definition of a_j
Evidence (≤25w)	"we compute the clustering outlier score aj of the mask mj as follows: (dnj ŋj = ŋ ⊠ aj = 1 ŋj ⊠= ŋ ⊠ ."
Problem	Mathematical definition of a_j is ambiguous due to missing line breaks/formatting; mapping of d_nj to cases is unclear.
Standard Violated	Equations must be precisely typeset to avoid ambiguity in algorithmic logic.
Impact	Readers may misimplement how mask importance scores are computed, altering attention maps and performance.
Confidence	High

I06

MODERATE · 3/5

Section	Threat Model
Location	Section III
Evidence (≤25w)	"we introduce the threat model under the SSL backdoor attack proposed in [12]. The primary objective of the attacker is to manipulate the output"
Problem	Threat model omits some key dimensions (attacker's knowledge of SSL algorithm, defender's access to compute, online/offline setting, presence of clean data) or states them only implicitly.
Standard Violated	Security papers should articulate threat model assumptions comprehensively and explicitly.
Impact	Unclear applicability boundaries; readers may overgeneralize or misinterpret the defense's guarantees.
Confidence	High

I04

MODERATE · 3/5

Section	Evaluation
Location	Table I and related text
Evidence (≤25w)	"All results are averaged over 5 independent runs with different seeds."
Problem	Only mean ACC and ASR are reported; there are no standard deviations, confidence intervals, or significance tests.
Standard Violated	Quantitative claims should include variability estimates and, where appropriate, significance analysis.
Impact	The strength of claimed improvements over baselines cannot be assessed for statistical reliability.
Confidence	High

I03

MODERATE · 3/5

Section	Methodology
Location	Section IV-D Poison Classifier Training
Evidence (≤25w)	"we eliminate a proportion p of images with the highest poison scores in X and utilize strong augmentations and early stopping to alleviate the noisy problem."
Problem	No numerical value or selection rationale for proportion p, early stopping criteria, or augmentation details is provided.
Standard Violated	Parameter choices affecting training and selection must be reported to support reproducibility and fair comparison.
Impact	Performance of the poison classifier is sensitive to these hyperparameters; results may not be replicable or comparable.
Confidence	High

I02

MODERATE · 3/5

Section	Methodology
Location	Figure 2 and Section IV overall
Evidence (≤25w)	"The overall architecture of our method is shown in Figure 2, which mainly consists of four steps"
Problem	High-level pipeline is provided, but critical implementation details (e.g., stopping criteria, exact sampling protocol, data preprocessing) are not fully specified.
Standard Violated	Experimental pipelines in CS papers should be sufficiently described to allow independent reimplementation.
Impact	Limits the community's ability to confirm results or compare fairly with alternative defenses.
Confidence	High

I24

MINOR · 2/5

Section	Evaluation
Location	Multi-target attack experiments
Evidence (≤25w)	"we conduct experiments against multi-target attacks... combine target categories of "rottweiler" and "tabby cat"... further add "ambulance""
Problem	Multi-target setting details (poison rates per class, triggers overlap, evaluation metrics under multi-target) are not fully explained.
Standard Violated	Complex attack scenarios must be carefully specified to interpret robustness claims.
Impact	Limits confidence in generalization claims to multi-target backdoor attacks.
Confidence	Medium

I25

MINOR · 2/5

Section	Evaluation
Location	Figure 11 hyper-parameter sensitivity
Evidence (≤25w)	"we have the following observations: (1) While increasing the number B of masks, IoU, CR of the detected candidate triggers and F1-score... first increase and then become relatively stable"
Problem	Hyper-parameter studies are descriptive only; no quantitative thresholds, optimal ranges, or constraints (e.g., computational budget) are formalized.
Standard Violated	Parameter sensitivity analysis should quantify trade-offs to guide practical configurations.
Impact	Practitioners may struggle to choose B, w, l appropriately for new datasets or constraints.
Confidence	Medium

I22

MINOR · 2/5

Section	Evaluation
Location	Table II
Evidence (≤25w)	"Total Rem. 12807.5 1570.8 ... Recall↑ 84.4 98.7 ... Precision↑ 5.4 49.3"
Problem	No explanation of why Total Rem. is non-integer (e.g., 12807.5) and how aggregation over runs is computed.
Standard Violated	Reported counts should be interpretable; averaging over runs needs clarification for count metrics.
Impact	Ambiguous values hinder accurate understanding of how aggressively each method removes data.
Confidence	High

I21

MINOR · 2/5

Section	Novelty/Related Work
Location	Section II Related Work and positioning
Evidence (≤25w)	"We propose a novel PoisonCAM method based on Cluster Activation Masking to detect and delete the poisonous samples"
Problem	While PatchSearch is acknowledged, the precise conceptual difference between Grad-CAM-based and Cluster Activation Masking approaches is not fully dissected.
Standard Violated	Novelty claims should clearly articulate how the method advances beyond closest prior work at a technical level.
Impact	Readers may struggle to assess the incremental contribution relative to established SSL backdoor defenses.
Confidence	Medium

I20

MINOR · 2/5

Section	Organization
Location	Masking strategies description
Evidence (≤25w)	"we have empirically tested three masking strategies, as shown in Figure 3. Although it seems that the 0-1 Interval Masking and Random Masking can break the pattern of the trigger"
Problem	Motivation and intuition for why Full Coverage Masking performs best are only vaguely stated as empirical, without further analysis.
Standard Violated	Method design choices should be conceptually motivated, not only justified post-hoc by experiments.
Impact	Makes it harder to generalize the approach or understand when alternative masking strategies could be preferable.
Confidence	Medium

I18

MINOR · 2/5

Section	Figures & Tables
Location	Figure 1 reference
Evidence (≤25w)	"As shown in Figure 1, our PoisonCAM significantly improves the top JOURNAL OF LATEX CLASS FILES"
Problem	Figure references are occasionally broken by formatting artifacts (e.g., page headers) and lack concise in-text descriptions of axes/metrics.
Standard Violated	Figures should be clearly referenced with unambiguous descriptions to aid interpretation.
Impact	Readers may struggle to interpret improvements (e.g., “top 20, 50, 100 accuracy”) without explicit metric definitions in text.
Confidence	High

I16

MINOR · 2/5

Section	Methodology
Location	Section VI-A Implementation Details
Evidence (≤25w)	"we sample top-20 poisonous images (i.e., |X p | = 20) and remove the top 10% poisonous samples in X to reduce noise in the poison classification set X."
Problem	Earlier description of poison set X^p and elimination proportion p does not specify that these are fixed at 20 images and 10%; consistency is only revealed later.
Standard Violated	Key design choices should be introduced coherently where the method is first defined, not only in implementation notes.
Impact	Earlier sections appear underspecified, complicating understanding of core algorithm design versus implementation choices.
Confidence	High

I14

MINOR · 2/5

Section	Reproducibility
Location	Data/code availability
Evidence (≤25w)	"Our code, data, and trained models will be open once this paper is accepted."
Problem	Artifact availability is contingent on acceptance; there is no current link, versioning, or description of what exactly will be released.
Standard Violated	Reproducibility standards increasingly expect concrete, verifiable artifact plans independent of publication outcome.
Impact	Reviewers and readers cannot verify results or inspect implementation details at submission time.
Confidence	High

I13

MINOR · 2/5

Section	Ethics/Privacy
Location	Global
Evidence (≤25w)	"No explicit ethics, risk, or responsible disclosure discussion is present in the manuscript."
Problem	The paper lacks an ethics or broader impact discussion about dual-use risks of improving backdoor methodologies and releasing code/datasets.
Standard Violated	Security/ML papers are expected to discuss potential misuse, dual-use, and mitigations.
Impact	Omitting ethical considerations may hinder responsible use and deployment of the research outputs.
Confidence	High

I12

MINOR · 2/5

Section	Evaluation
Location	Table II caption
Evidence (≤25w)	"Total Rem. DENOTES THE NUMBER OF TOTAL REMOVED SAMPLES ."
Problem	Precision and recall definitions for poison detection are not explicitly stated (e.g., poison defined at image level, handling multi-target cases).
Standard Violated	Derived metrics must be formally defined, especially when evaluating detection algorithms.
Impact	Potential confusion interpreting large differences in precision (e.g., 5.4 vs 49.3) and Total Rem.
Confidence	High

I09

MINOR · 2/5

Section	Methodology
Location	Section IV-C end
Evidence (≤25w)	"Finally, we take the top-k scored images with the highest poison scores to form a poison set X p . This poison set will be utilized to train a poison classifier."
Problem	Exact value of k and its dependence on dataset size or poison rate are unclear; later text mentions |X^p|=20 without reconciling with this general description.
Standard Violated	Key hyperparameters influencing training sets must be clearly stated and consistently referenced.
Impact	Ambiguity about poison set size makes it hard to interpret Table II and replicate classifier training.
Confidence	High

I05

MINOR · 2/5

Section	Evaluation
Location	Definition of Attack Success Rate
Evidence (≤25w)	"Attack Success Rate (ASR) refers to the proportion of non- target classes misclassified as the targeted class."
Problem	ASR definition lacks details (e.g., conditioned on presence of trigger, per-class averaging) and differs from some backdoor conventions without clarification.
Standard Violated	Metrics must be rigorously defined to avoid ambiguity and enable comparison to prior work.
Impact	Readers may misinterpret ASR values and comparisons to existing defenses/attacks.
Confidence	High

I19

POLISH · 1/5

Section	Organization
Location	Abstract
Evidence (≤25w)	"SSL is an effective paradigm for learning representations from unlabeled data, such the security and robustness of SSL models, of which the proceas text, images, and videos."
Problem	Abstract contains typographical and grammatical errors (truncated sentence, “proceas”), reducing clarity.
Standard Violated	Abstract should present a polished, grammatically correct overview of the work.
Impact	Weakens initial impression and may obscure the main contribution for first-time readers.
Confidence	High

I26

POLISH · 1/5

Section	Writing
Location	First page header/body
Evidence (≤25w)	"JOURNAL OF LATEX CLASS FILES, VOL. O, NO. O, FEBRUARY 2024 Erasing Self-Supervised Learning Backdoor"
Problem	LaTeX template headers and footers are interleaved with body text in several places, occasionally breaking sentences.
Standard Violated	Manuscript should be cleanly typeset without template artifacts in the main text.
Impact	Reduces readability and may obscure some sentence boundaries, though technical content remains inferable.
Confidence	High

Consistency Audit

I16

Type	other
Evidence A	"we can compute the poison scores of all images and simply take the top-k image..." (P42)
Evidence B	"we sample top-20 poisonous images (i.e., |X p | = 20) and remove the top 10% poisonous samples in X" (P54)
Inconsistency	General description of selecting top-k poisonous images is later instantiated as specifically top-20 and 10% removal without early clarification, causing parameter ambiguity.

I07

Type	notation_units
Evidence A	"aj of the mask mj as follows: (dnj ŋj = ŋ ⊠ aj = 1 ŋj ⊠= ŋ ⊠ ." (P38)
Evidence B	"To scale cluster distances, we utilize Max-Min Scaling d -mink dk dnj = maxjk dk -min k dk ⊠ [0, 1]." (P33)
Inconsistency	Equation formatting obscures whether a_j equals d_nj when ŋ_j = ŋ* or vice versa; inconsistent line breaks make notation ambiguous.

I22

Type	numeric
Evidence A	"Total Rem. 12807.5 1570.8" (P52)
Evidence B	"The training set has about 127K samples and the validation set has 5K samples." (P46)
Inconsistency	Total removed sample counts appear as fractional values despite underlying discrete datasets, indicating averaging over runs without explicit explanation.

Action Plan

I01, I02, I03, I08, I09, I10, I11, I16, I17

Action	Provide a complete, implementation-level description of Cluster Activation Masking and the overall pipeline: (a) formalize mask generation (shape, sampling distribution, stride, overlap) and the heuristic search algorithm (iteration counts, cluster removal schedule, sampling scheme) with pseudo-code; (b) specify all hyperparameters (including p, k, early stopping, augmentations, poisoning set size |X^p|, fixed values like 20 images and 10% where used) and how they scale with dataset size/poison rate; (c) document all data preprocessing, resizing, and the handling/labeling of synthesized vs original (including truly poisoned) images; and (d) detail any deviations from PatchSearch's original implementation (hyperparameters, tuning, data).
Effort (S/M/L)	L
Priority (P0–P2)	P0
Success Criteria	An independent reader can reimplement the full method from the Methods section alone and reproduce results within expected variance, without needing to guess any algorithmic steps, hyperparameters, or data handling choices.

I10, I11, I22, I23, I24, I25

Action	Standardize and fully document the experimental protocol: (a) explicitly describe train/validation/test splits and use of unlabeled STL-10 data; (b) state all preprocessing operations; (c) define the hyperparameter tuning protocol for both PoisonCAM and PatchSearch (search space, selection criterion, fairness); (d) explain how aggregate metrics like non-integer 'Total Rem.' are computed over runs; (e) clearly specify multi-target settings (poison rates per class, trigger overlaps, metrics in this regime); and (f) summarize quantitative hyperparameter studies with identified optimal ranges and practical constraints (e.g., compute budget).
Effort (S/M/L)	M
Priority (P0–P2)	P0
Success Criteria	Experimental setup can be exactly replicated by an external group, and all table entries (including aggregates like Total Rem.) can be recomputed from the described protocol and hyperparameters.

I04, I15, S01, S02, S03, S04

Action	Correct all numerical/statistical inconsistencies and add uncertainty quantification: (a) verify all arithmetic in the text and tables (e.g., sums, percentages like 5.4/100, totals such as 70.3) and correct errors; (b) for all key metrics (ACC, ASR, precision, recall, effect sizes), report standard deviations or confidence intervals based on multiple runs; (c) note when small changes (e.g., 0.9% ASR reduction) are within or outside the noise level; and (d) briefly describe the number of runs and statistical summarization method.
Effort (S/M/L)	M
Priority (P0–P2)	P0
Success Criteria	All numerical claims are internally consistent; each reported performance figure on main tables/plots includes a measure of variability, and small effect sizes are explicitly contextualized relative to uncertainty.

X01, X02, I05, I12

Action	Resolve cross-section inconsistencies and formalize metric definitions: (a) ensure poisoning rates and any derived descriptions (e.g., '5.0% meaning 50% of target-category images poisoned') are self-consistent across abstract, methods, and results; (b) reconcile any conflicting accuracy percentages; (c) provide precise mathematical definitions for ASR (conditioning on trigger presence, per-class or overall averaging) and poison detection precision/recall (unit of poison, multi-target handling); and (d) update all text/tables to use these definitions consistently.
Effort (S/M/L)	M
Priority (P0–P2)	P0
Success Criteria	There are no conflicting values for poisoning rates or accuracies in different sections, and formal metric definitions allow a reader to reproduce every reported number without ambiguity.

V01, V02, V03, V04, V05, V06, V07, V08, V09, V10, V11, V12, V13, V14, I18

Action	Overhaul figures and captions for clarity and legibility: (a) remake low-resolution plots with larger fonts, clear axis labels (with units and metric names), and visually distinct colors/linestyles; (b) ensure each caption accurately and completely describes the corresponding figure content (no truncation, boilerplate, or mismatched descriptions like masking strategies vs heatmaps); (c) add legends and textual labels to diagrams explaining each sub-panel and operation; (d) include the missing visual content (e.g., the Grad-CAM example for Figure 4); and (e) fix broken in-text references and briefly describe axes/metrics in the main text when first referencing each figure.
Effort (S/M/L)	L
Priority (P0–P2)	P1
Success Criteria	All figures are readable when printed or viewed at normal zoom, caption text is complete and correctly aligned with figure content, and a reader unfamiliar with the work can understand each figure's purpose from its caption and in-text reference.

I06, I20, I21

Action	Clarify the threat model and deepen the conceptual analysis: (a) explicitly state attacker and defender assumptions (knowledge of SSL algorithm, availability of clean data, online vs offline setting, compute/resources); (b) clearly articulate how Grad-CAM-based and Cluster Activation Masking approaches differ conceptually and practically; and (c) provide substantive intuition or analysis explaining why Full Coverage Masking performs best (e.g., qualitative examples, connections to feature coverage or cluster structure).
Effort (S/M/L)	M
Priority (P0–P2)	P1
Success Criteria	Threat model bullets can be summarized in a short list of explicit assumptions, and the paper contains a distinct subsection that clearly contrasts PoisonCAM with PatchSearch and justifies the behavior of masking strategies beyond empirical observation.

l13, l14

Action	Add an ethics/broader impact and artifact availability section: (a) discuss dual-use risks of improving backdoor methodologies and releasing code/datasets, including how misuse might be mitigated (e.g., restricted release, usage guidelines); (b) state clearly what artifacts (code, configs, trained models, data splits) will be released, where, and under what license; and (c) if availability is contingent, indicate timelines and any conditions, or preferably provide an anonymized or public repository link.
Effort (S/M/L)	S
Priority (P0–P2)	P2
Success Criteria	The manuscript contains a dedicated paragraph or section on ethical considerations and a concrete artifact plan with a stable URL or explicit, time-bound availability commitment.

l19, l26

Action	Perform a thorough writing and formatting pass: (a) remove all stray LaTeX template headers/footers from the body and captions; (b) fix truncated and garbled sentences in the abstract and elsewhere; (c) standardize terminology and grammar; and (d) run automated spell-check plus manual proofreading to catch remaining typographical issues.
Effort (S/M/L)	S
Priority (P0–P2)	P2
Success Criteria	No template artifacts remain in the compiled PDF, sentences are complete and grammatical in the abstract and main text, and reviewers can read the manuscript without being distracted by formatting glitches.