REPAIR: INVERTING BACKDOOR ATTACKS IN SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite significant advances in computer vision, semantic segmentation models remain vulnerable to backdoor attacks, posing a security risk. These attacks redirect pixels of a chosen victim class to an attacker-specified target class when a trigger is present, while preserving clean predictions. Existing defenses are largely derived from image classification, while dedicated segmentation defenses remain scarce and typically require retraining or model modification. We introduce RePAIR, a post-hoc, label-free, attack-unaware defense based on the observation that backdoor-induced shifts can be stronger in the classifier head than in internal representations, leaving recoverable victim-class semantics. RePAIR infers the victim-target relation from an unlabeled, potentially contaminated image pool, constructs a safety-approved semantic reference, and localizes suspicious pixels by comparing head predictions with internal semantic evidence. It then applies feature-space and output corrections toward the inferred victim semantics, while conservative localization and output constraints limit changes to benign predictions. We model the attack as adversarial probability transport and repair as its approximate reversal, deriving bounds and sufficient conditions for localization, residual attack success, clean-side changes, defense-induced target predictions, and attack-unaware pair identification. Experiments across segmentation architectures and attack settings show that RePAIR substantially reduces attack success while largely preserving clean performance.

1 INTRODUCTION

Semantic segmentation assigns a semantic class to every pixel proving perception in safety-critical systems such as automated driving (Cakir et al., 2022) and medical imaging (Taghanaki et al., 2024). Semantic segmentation networks are vulnerable to backdoor attacks that preserve normal performance on benign inputs while inducing targeted failures when a trigger is present (Li et al., 2021b). For instance, a backdoor can cause a vehicle to disappear from the semantic map by relabeling its pixels as road, without affecting the prediction for the rest of the image. Semantic segmentation backdoors have developed from fixed trigger patterns and predefined victim–target mappings (Li et al., 2021b) toward increasingly flexible attack settings. Existing attacks allow the victim class to depend on trigger placement (Mao et al., 2023), exploit contextual relationships between semantic classes (Abbasi et al., 2025), or automatically optimize triggers and attack configurations (Lan et al., 2024). These attacks demonstrate that backdoors can induce targeted and spatially localized prediction changes under flexible conditions, making the development of robust defenses challenging. While attacks on dense prediction tasks have received increasing attention, effective defenses remain comparatively underexplored. Detection-oriented methods can identify suspicious inputs or models but do not necessarily remove the underlying backdoor (Liu et al., 2023). Mitigation approaches based on retraining (e.g. fine-tuning and pruning) often require additional clean data and may alter the model's behavior (Liu et al., 2018). For semantic segmentation, effective defenses should selectively correct attacked pixels while limiting changes to unrelated regions, motivating selective repair rather than global model modification.

In this paper, we introduce a **Re**presentation-guided pixel-wise attack Re**pair** (RePAIR) against backdoor attacks on semantic segmentation models. We focus on a key empirical observation: the backdoor-induced shift can be substantially stronger in the final task decision than in an appropriately chosen internal semantic representation, even when the encoder itself is affected by the

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: From left to right: poisoned image, prediction under the backdoor attack, prediction after applying our defense and ground truth segmentation.

backdoor. Victim semantics can therefore remain partially recoverable even when the head predicts the attacker-selected target. The representation that exposes this state is architecture-dependent. E.g. for DeepLabV3+ (Chen et al., 2018) with a ResNet-50 backbone, the final backbone stage immediately before atrous spatial pyramid pooling provides a useful reference. For SegFormer (Xie et al., 2021) with a MiT-B5 backbone, no single block is sufficient. Victim evidence is distributed across several internal blocks and is reconstructed by a learned multi-layer logistic read-out.

This observation has a natural geometric interpretation. For each pixel, the predicted class probabilities form a point in the probability simplex, where each coordinate represents the probability assigned to one semantic class. An idealized targeted backdoor shifts probability mass from a victim class to an attacker-selected target class, moving the prediction toward the target and away from the victim while remaining within the same simplex. Natural segmentation errors can also shift probability mass between classes (typically not concentrated in one direction), but a backdoor induces a systematic flow from one semantic class to another.

This directed view can be described by a local transport operator acting on the class-probability distribution. RePAIR approximates the corresponding reverse transport from surviving victim evidence and model-derived reference geometry, while treating feasible victim-target pairs as hypotheses and identifying the attacked relation without prior knowledge of either class.

A correct victim–target pair alone is not sufficient for a safe repair. The reference data used for repair must also be protected from poisoned samples, otherwise, the defense may learn the very distortion it is meant to remove. We therefore first identify candidate clean images and then check whether they provide reliable evidence of the victim class before using them for repair. An image is rejected if its internal victim evidence is severely degraded, or if a moderate degradation is accompanied by an unusually strong shift from the victim class to the target class. Rejected images are not replaced merely to meet a desired reference-set size. If too little reliable victim evidence remains, the defense does not construct a repair state. Together, these components form our method for selective and reliable backdoor repair. The source code of our method is publicly available at `https://github.com/`. Our contributions can be summarized as follows:

- We introduce a post-hoc repair method for backdoored semantic segmentation models that reconstructs victim-consistent semantics and selectively corrects attacked pixels, without retraining or globally modifying the deployed network.
- We develop a theoretical framework that models targeted segmentation backdoors as directed victim-to-target transport in the class-probability simplex. This view links surviving internal evidence to localization, residual attack success, attack-unaware pair identification, and clean-side safety.
- Our method for generating a reliable repair reference is attack-unaware and operates on potentially compromised data. It identifies consistent victim-to-target evidence and screens candidate clean-like images with a one-way victim-safety firewall.
- We develop architecture-specific semantic references for both convolutional- and transformer-based segmentation models, showing that the information required for repair may be concentrated in a single representation or distributed across multiple internal layers.

2 Related Work

Backdoors Attacks. The first work adapted the BadNets attack (Gu et al., 2019) from image classification to semantic segmentation, using either a black stripe at the image border as a non-semantic

trigger or a semantic class present in the image as a trigger (Li et al., 2021b). Once activated, the attack relabeled all pixels of a victim class as a target class. The Influencer Backdoor Attack (IBA, (Lan et al., 2024)) places a trigger on non-victim pixels while relabeling the victim class as a predefined target. It further introduces Nearest Neighbor Injection (NNI) and Pixel Random Labeling (PRL), which respectively place the trigger near the victim region and add random label changes to non-victim pixels to strengthen the attack's reliance on global context. The Object-Free Backdoor Attack (OFBA, (Mao et al., 2023)) places the trigger on the target object and varies victim classes and trigger positions during training, allowing the victim class to be selected by trigger placement at inference time while keeping the target class fixed. The Contextual Segmentation attack (ConSeg) uses a natural semantic trigger, such as a motorcycle, instead of an artificial pattern (Abbasi et al., 2025). It relabels the victim class as the target and adds a small number of co-occurring target-related classes within the victim region to strengthen the contextual association. BADSEG (Zhang et al., 2026) broadens the attack space to four attack types: Object-to-Object (O2O), Object-to-Background (O2B), Background-to-Object (B2O), and Background-to-Background (B2B), along with instance-level and conditional variants. A surrogate model optimizes trigger parameters and selects victim-target pairs using feature-space class centroids.

Defense Methods. Work on defenses against backdoor attacks in semantic segmentation remains limited, and existing studies therefore mainly adapt defenses originally developed for image classification. Test-time detection methods such as STRIP (Gao et al., 2019) and TeCo (Liu et al., 2023) detect suspicious inputs by measuring prediction entropy under image superposition or prediction consistency under image corruptions, respectively. Similarly, Zeng et al. (Zeng et al., 2021) detect anomalous trigger-related patterns in frequency domain (Ahmed et al., 1974). BEATRIX (Ma et al., 2023) detects anomalous internal representations using class-conditional higher-order Gram matrices (Zhang et al., 2026). These methods aim to identify suspicious inputs but do not repair the affected predictions or remove the underlying backdoor. Pruning (Liu et al., 2018) removes neurons with low activation on benign inputs and fine-tunes the remaining network on clean data to eliminate dormant backdoor components and recover clean performance. Fine-tuning (Sha et al., 2022) instead directly retrains the potentially compromised model on clean data to weaken the learned association between trigger and target. Both methods were also evaluated as defenses against segmentation backdoors, but both require access to clean data for mitigation (Lan et al., 2024; Zhang et al., 2026). Segment Cleanse (Mao et al., 2023) adapts Neural Cleanse (Wang et al., 2019) to semantic segmentation by reconstructing object-level trigger candidates and mitigating detected backdoors through retraining, likewise requiring model modification and clean data. Anti Backdoor Learning (ABL, (Li et al., 2021a)) exploits differences in training loss dynamics to identify suspicious samples and reduce their influence through gradient ascent. While these approaches identify or suppress backdoor behavior, they do not enable post-hoc, selective repair of affected pixels in a deployed model.

RePAIR differs from detection-based methods (e.g. STRIP, TeCo, frequency-domain defenses, and BEATRIX), which identify suspicious inputs, samples, or internal representations but do not repair the affected predictions. In contrast to model-modification approaches such as pruning, fine-tuning, and Segment Cleanse, our method does not retrain or otherwise alter the compromised model. It is also applied post-hoc, unlike ABL, which operates during training. Finally, RePAIR requires neither ground truth labels nor an external clean reference set, instead deriving the information needed for selective repair from the model and available data while leaving unaffected predictions unchanged.

3 Theoretical Framework

We view a targeted backdoor in semantic segmentation as a trigger-induced, spatially selective change in dense predictions. Semantic segmentation predicts a probability distribution over semantic classes for each pixel, allowing a backdoor to selectively alter the predictions of pixels belonging to a particular victim class. During training, the attacker introduces triggered images with modified pixel-wise labels, assigning victim-class pixels to an attacker-selected target class. When activated, the backdoor causes pixels belonging to a victim class to be relabeled as an attacker-selected target class, while leaving other pixels largely unaffected. We treat the images as samples of a stochastic process, we further assume the neural network to learn the induced distribution during training. In order to change the argmax of the attacked pixels' prediction the trigger has to lead to redistribution of probabilisitc mass from one atom to another (assumed to be mostly from victim to target class).

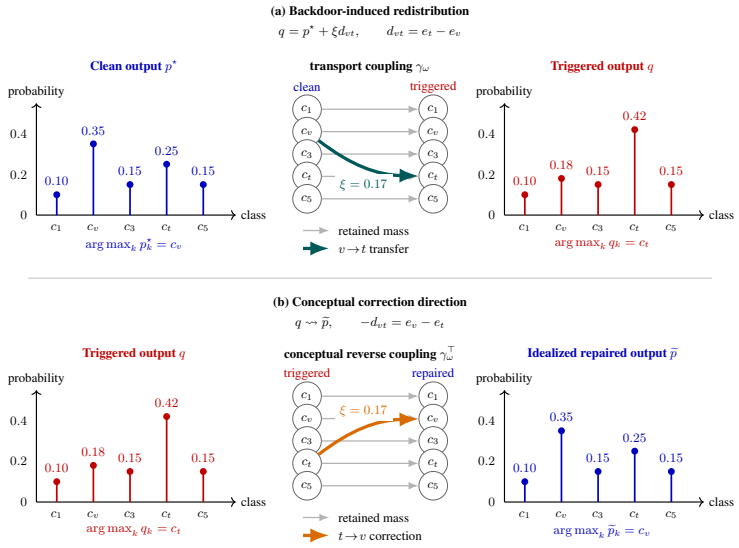


Figure 2: Targeted backdoor as directed victim-to-target transport. The prediction may move toward the target while internal representations retain victim evidence, which RePAIR uses for localization and repair. The forward operator and its coupling describe the same directed mass transfer at the marginal and joint-measure levels, respectively.

The possibility of recognizing and "inverting" this transport yields two effects, the identification of victim and target class and nullyfication of the triggers effect, thus rendering it use- and harmless.

We model a targeted segmentation backdoor as a spatially selective victim-to-target distortion, corresponding to a directed redistribution of class probability from a victim class v to a target class $t \neq v$. A successful backdoor may drive the final segmentation head toward the target class while an architecture-specific internal reference still retains evidence of the original victim semantics. This disagreement between the final prediction and the internal semantic evidence provides the signal used for localization and repair. RePAIR exploits this mismatch without assuming encoder invariance, semantic ground truth, trigger presence, or prior knowledge of the attacked pair.

Directed distortion. Let $p^\star \in \Delta^{K-1}$ denote the clean reference probability vector over K semantic classes and q its triggered counterpart predicted by a semantic segmentation network. The attack is modeled locally as a transfer of probability mass $\xi \geq 0$ from the victim class v toward the target class t

$$q = T_\xi(p^\star) = p^\star + \xi(e_t - e_v) \ , \qquad 0 \leq \xi \leq p^\star_v \ , \tag{1}$$

where e_c is the unit vector associated with class c. The geometric interpretation shown is in Fig. 2. Thus the victim probability decreases by ξ, the target probability increases by ξ, and the victim-over-target margin $\psi_{vt}(p) = p_v - p_t$ satisfies

$$\psi_{vt}(q) = \psi_{vt}(p^\star) - 2\xi \ . \tag{2}$$

Assuming $p^\star_v > p^\star_t$, pairwise preference changes once $\xi > (p^\star_v - p^\star_t)/2$. A complete class change additionally requires the target coordinate to exceed all remaining classes. This is an analytical description of the dominant directed failure, not a claim that the poisoned network is globally one-dimensional or that RePAIR observes the unknown clean $p^\star$ or the transported mass ξ.

Inverse operator and transport coupling. The local model in Equation equation 1 makes the notion of *inverting* precise. If ξ were known and q lay in the feasible image of T_ξ, the inverse marginal map would be

$$T_\xi^{-1}(q) = q + \xi(e_v - e_t) \ . \tag{3}$$

The same forward displacement can also be represented by a one-edge coupling γ_ξ. With $(\gamma_\xi)_{vv} = p^\star_v - \xi$, $(\gamma_\xi)_{vt} = \xi$, and $(\gamma_\xi)_{kk} = p^\star_k$ for $k \neq v$, its marginals satisfy

$$\gamma_\xi \mathbf{1} = p^\star \ , \qquad \gamma_\xi^\top \mathbf{1} = q \ . \tag{4}$$

These are related but distinct descriptions of reversal. The inverse operator T_ξ^{-1} maps the displaced marginal q back toward $p^\star$, whereas the transposed coupling $\gamma_\xi^\top$ exchanges the forward coupling's marginals and therefore represents the corresponding reverse mass flow as a joint measure. The transposed coupling is not itself the inverse operator, and no optimal-transport claim is implied without specifying a transport cost. RePAIR observes neither ξ nor the hidden attack operator or coupling. Its repair instead approximates the reverse victim-directed transport from the discovered pair, surviving internal victim evidence, and safety-approved reference geometry. This is the sense in which RePAIR *inverts* the backdoor effect.

Surviving internal evidence and localization. Let $A = \{\widehat{Y}^{\mathrm{head}} = t\}$ denote successful baseline attack on a triggered pixel of the victim class, where $\widehat{Y}^{\mathrm{head}}$ is the prediction of the undefended segmentation head. Let ρ_v indicate whether the internal representation still retains evidence for the victim class, and we define η_{ref} as the fraction of successful baseline attacks for which this direct victim evidence is absent. Thus, a small η_{ref} means that victim evidence survives in most successful attacks, although this does not guarantee correct recovery. When the released pair equals the true attacked pair, every successful attack with surviving victim evidence is selected by the direct semantic arm. Therefore,

$$\eta_{\mathrm{loc}} \le \eta_{\mathrm{ref}} \ . \tag{5}$$

Additional residual branches can only reduce the miss rate further.

Residual attack success. Among triggered victim pixels, let $\varepsilon_{\mathrm{rep}}$ be the fraction of localized baseline attacks that still predict t after repair, and let δ_{new} denote defended target predictions outside the original baseline attack event. Then the defended attack success rate is bounded by

$$\mathrm{ASR}_{\mathrm{def}} \le \mathrm{ASR}_{\mathrm{base}}\big[\eta_{\mathrm{loc}} + (1-\eta_{\mathrm{loc}})\varepsilon_{\mathrm{rep}}\big] + \delta_{\mathrm{new}} \ . \tag{6}$$

The bound separates three failure modes: missed attacked pixels, unsuccessful repair on localized pixels, and new target predictions introduced elsewhere. When the released pair equals the true attacked pair, combining $\eta_{\mathrm{loc}} \le \eta_{\mathrm{ref}}$ with the residual-ASR bound gives

$$\mathrm{ASR}_{\mathrm{def}} \le \mathrm{ASR}_{\mathrm{base}}\big[\eta_{\mathrm{ref}} + (1-\eta_{\mathrm{ref}})\varepsilon_{\mathrm{rep}}\big] + \delta_{\mathrm{new}} \ . \tag{7}$$

Thus, surviving internal victim evidence limits the fraction of successful attacks that can be missed by the localizer. Whether these localized attacks are actually recovered still depends on the subsequent repair step. The No New Target Veto prevents the defense from introducing unsupported target predictions outside the authorized repair support.

Attack-unaware pair identification. The victim and target classes are not assumed to be known. Instead, RePAIR evaluates all ordered class pairs using an attack-unaware image-level evidence score. For a pair (a, b), let $G_i(a, b)$ denote the score on image-level, and let $g_{ab} = \mathbb{E}[G_i(a, b)]$ be its expected value. For analysis, we decompose this score into natural evidence and an attack-induced contribution $g_{ab} = \nu_{ab} + \beta_{ab}$. Advantage over every competitor exceeds the strongest competing natural-evidence advantage plus δ_{est}. The attacked pair can be uniquely identified if its attack-induced advantage over every competing pair is larger than both the strongest natural-evidence advantage of a competitor and the estimation error δ_{est}. This provides a sufficient separation condition for pair identification, but does not constitute a distribution-free guarantee for the implemented multi-route selector.

Clean-side safety. Let κ_{cl} denote the probability that a benign pixel is selected for repair, and let $\delta_{\mathrm{spill}}^{\mathrm{cl}}$ denote the probability of a defended prediction change outside the selected mask. Then

$$\mathbb{P}_{\mathrm{cl}}(\widehat{Y}^{\mathrm{def}} \neq \widehat{Y}^{\mathrm{head}}) \le \kappa_{\mathrm{cl}} + \delta_{\mathrm{spill}}^{\mathrm{cl}} \ , \tag{8}$$

where $\mathbb{P}_{\mathrm{cl}}$ is the distribution over clean inputs and $\widehat{Y}^{\mathrm{def}}$ the prediction after repair. Thus, clean prediction changes are bounded by the probability of selecting benign pixels and the probability of changes outside the repair mask. Exact preservation is the special case where both terms are zero. The reference-calibration budgets are empirical constraints on the safety-approved population and do not provide population-level guarantees for unseen benign inputs.

The framework does not require reconstruction of an unobserved clean class distribution. Instead, it uses surviving victim evidence, attack-unaware pair identification, conservative localization, and a safety-approved reference population to selectively repair target predictions. The forward transport and its reversal provide the conceptual model, while the implemented defense approximates this reversal using model-derived evidence. Complete definitions and proofs are provided in Appendix A.1.

4 METHOD

RePAIR is a post-hoc, label-free, attack-unaware defense for targeted semantic segmentation backdoors. The compromised model remains frozen. The defense uses neither trigger reconstruction nor retraining, and uses no semantic ground truth, clean/poison provenance, attack mask, attack family, poisoning rate, or predefined victim-target pair. Its inputs are the model and an unlabeled image pool that may contain triggered samples. The method has two stages. First, an offline builder identifies a directed victim-target relation from model-derived evidence and constructs a conservative, safety-approved reference state. Throughout this section, (v, t) denotes the operational victim–target pair released by the attack-unaware builder. At inference, a pair-conditioned localizer identifies inconsistent pixels, feature transport moves them toward victim semantics, and output safety limits unintended changes. More details are provided in Appendix A.2.

This operational pipeline implements the reverse-transport interpretation of Section 3 without estimating the hidden attack operator T_ξ, its transported mass ξ, or the coupling γ_ξ. Pair discovery estimates the semantic direction, the safety-approved state supplies a victim-side reference geometry, and localization determines where that reverse direction may be applied. Repair magnitude is therefore pixel- and branch-dependent rather than an estimate of a single global ξ.

Architecture-specific semantic reference. We write the segmentation network as $f = h \circ g$, where g is the feature extractor and h the segmentation head. For each pixel x, the head produces class probabilities $q_k(x)$ and prediction $\widehat{Y}^{\mathrm{head}}(x) = \arg\max_k q_k(x)$. RePAIR compares this prediction with an architecture-specific internal semantic reference, which differs architectures.

For convolutional networks (CNNs), we use the normalized final backbone feature $z(x)$. For each class k, its reference centroid is computed as $\mu_k = |\mathcal{P}_k|^{-1} \sum_{x \in \mathcal{P}_k} z(x)$, where $\mathcal{P}_k$ contains pixels assigned to k by the frozen segmentation head. The internal prediction is then the class with the most similar centroid under cosine similarity $\widehat{Y}^{\mathrm{prot}}(x) = \arg\max_k \mathrm{sim}(z(x), \mu_k)$. Reference estimation is image-disjoint, so an image is never evaluated using a centroid estimated from that same image.

For transformers, semantic evidence is distributed across several transformer blocks rather than concentrated in a single feature space. For each feasible victim-target pair (v, t), we therefore use L selected blocks and compute the cosine similarity of each internal feature $z_\ell(x)$ to the corresponding victim and target reference centroids $\mu_{\ell,v}$ and $\mu_{\ell,t}$, obtaining $c_{\ell,v}(x)$ and $c_{\ell,t}(x)$. The $2L$ similarities are concatenated into $\phi_{v,t}(x)$ and standardized before a regularized joint logistic readout, $r_{v,t}(x) = \sigma(w^\top \widetilde{\phi}_{v,t}(x) + b)$. Here, $r_{v,t}(x)$ measures target compatibility, while $p_v^{\mathrm{int}}(x) = 1 - r_{v,t}(x)$ represents internal victim support. The readout is trained only from victim/target pseudo-labels produced by the frozen segmentation head on the reference population, i.e., no semantic or attack labels are used. The joint readout provides the direct victim-evidence signal, while the layer-wise similarities are retained as semantic coordinates for the residual geometry.

Attack-unaware pair discovery. Every feasible ordered pair (v, t) is initially treated as a candidate victim-target relation. For CNNs, the primary signal is internal-to-head contradiction: evidence for (v, t) increases when pixels retain victim-class evidence in the internal representation while the segmentation head predicts t. This evidence is aggregated over frozen, image-disjoint reference views and retained only when the support is sufficiently stable and separated from competing pairs. The resulting pair defines the victim-to-target direction used by the subsequent localization and repair stages. Thus, RePAIR can identify the attack direction directly from model-derived evidence, without requiring prior knowledge of the victim or target class.

For transformers, candidate structure is first identified using lightweight head, transfer, centroid, and structural cues. Supported pairs are then evaluated using image-disjoint joint-LogReg out-of-fold (OOF) evidence, with a separate OOF state for each pair. Stable family proposals are formed from structural, mechanism, and family-verification evidence. The frozen rank prior is retained whenever it passes the candidate viability test. If the prior family is not viable, another family may be selected only through the conservative rescue path, which requires bilateral leave-one-out support and bilateral structural consistency.

After the family has been frozen, its two possible victim-target directions are evaluated using FullEdge, FamilyVerify, and eligible Mechanism evidence. A direction is accepted only when the required direction agreement is unique. Structural evidence is not a regular direction vote and is

used only to corroborate the rescue rule when the Mechanism direction channel is ineligible. If family viability or direction evidence is insufficient, the mapping remains inactive. The exact release contract is given in Appendix A.1.6.

Safety-approved reference state. After a victim-target pair has been selected, a correct mapping is not yet sufficient because the unlabeled image pool may itself contain poisoned samples. The builder therefore screens candidate reference images using image-disjoint cross-fitting and a victim-safety firewall. Images with unstable internal predictions are first rejected, followed by a safety filter that removes images showing severe victim-class collapse or moderate victim degradation accompanied by an unusually strong victim-to-target contradiction. These filters are applied sequentially and monotonically: rejected images are not reintroduced to satisfy a desired reference-set size, and thresholds are not relaxed to rescue an insufficient population.

The builder then checks whether the remaining images provide sufficient reliable support and semantic coverage for the selected victim-target pair. If not, the repair state remains inactive and no repair is applied at inference time. Otherwise, all runtime components, including semantic references, probability statistics, feature-space reference banks, and repair calibrations, are constructed exclusively from the final safety-approved population. We use this term rather than clean reference set because the defense has no access to the true provenance of the images and therefore cannot guarantee that they are trigger-free.

The resulting state provides the empirical victim-side reference used by the runtime defense. It is not treated as the unknown clean distribution $p^\star$ and is not assumed to be uncontaminated, i.e., it is instead a model-derived reference population that has passed the defense's one-way safety checks.

Pair-consistent localization. Localization combines direct semantic contradiction with residual tests for pixels whose internal victim evidence is weak or distributed across multiple features. A shared probability branch compares the observed target-over-victim probability excess $d_p(x) = q_t(x) - q_v(x)$ with a semantically matched reference excess. Victim and target reference neighborhoods provide matched reference values and a semantic interpolation coordinate, yielding the expected excess $\widehat{d}_{\mathrm{ref}}(x)$. RePAIR measures the one-sided residual $D_{\mathrm{prob}}(x) = [d_p(x) - \widehat{d}_{\mathrm{ref}}(x)]_+$, which is calibrated image-disjointly on the frozen safety-approved population using an empirical reference-pixel budget. This budget controls the probability-residual calibration and is distinct from the repair margin. Matching is performed after removing the victim-target semantic axis, so the expected margin is conditioned on the remaining semantic geometry rather than on the displacement being tested.

CNNs combine the preserved semantic contradiction with target-manifold, self-counterfactual, and probability-residual cues. Transformer use five sequential, mutually exclusive stages restricted to baseline head-target pixels: semantic contradiction, residual victim-manifold rupture, residual target-manifold rupture, residual counterfactual manifold trajectory, and residual semantic probability. Their union defines M_{TF} and satisfies $M_{\mathrm{TF}} \subseteq \{\widehat{Y}^{\mathrm{head}} = t\}$. The staged construction records why a pixel was selected, prevents later branches from relabeling an earlier decision, and enables branch-specific repair. Exact masks and image-disjoint calibration rules are given in Appendix A.2.4.

Feature-first repair. RePAIR first repairs the internal representation of localized pixels before applying any residual output correction. For each localized feature, the discovered victim-target direction is removed from the retrieval geometry, and safety-approved victim references are searched in the orthogonal complement. The retrieved neighborhood is aggregated and the feature is transported toward the victim reference with branch-dependent strength, while preserving its original norm. The modified representation is then passed through the unchanged remainder of the network. By matching in the orthogonal complement, the reference search is not dominated by the same victim-target displacement that the repair is intended to reverse.

If the feature repair does not fully restore the victim prediction, RePAIR applies a conservative output correction restricted to the discovered victim-target pair. The correction is calibrated from safety-approved victim references and can be partial depending on the repair branch. Thus, feature transport and output correction address the same victim-target discrepancy at different stages: feature transport moves the internal representation toward victim semantics, while the output step corrects residual target predictions by directly modifying only the victim and target logits.

The resulting feature transport approximates the reverse-transport direction without requiring the unknown clean distribution or transport magnitude. The discovered pair provides the direction, while the safety-approved victim references provide the local repair target.

Output safety and decision isolation. Let $\widehat{Y}^{\mathrm{rep}}$ denote the prediction after repair but before output safety. The No New Target Veto restores $\widehat{Y}^{\mathrm{head}}$ whenever the repair changes a pixel outside the authorized mask to the discovered target class t, provided that the original head prediction was not t. Thus, the veto prevents unsupported new target predictions outside the repair support while leaving existing target predictions unaffected. The veto is intentionally asymmetric and only constrains the creation of the discovered target.

5 Experiments

5.1 Experimental Setting

We evaluate RePAIR on Cityscapes (Cordts et al., 2016), PASCAL VOC 2012 (Everingham et al., 2015) and ADE20K (et al., 2017). We consider two representative segmentation architectures, DeepLabV3+ with a ResNet-50 backbone (Chen et al., 2018) and SegFormer with a MiT-B5 backbone (Xie et al., 2021).

We evaluate backdoor attacks at a 10% poison rate, including IBA with NNI and PRL strategies, ConSeg, OFBA, and BadSeg with all four attack configurations. For all object-to-background attacks (IBA NNI, IBA PRL, ConSeg, OFBA, and BadSeg-O2B), we use car→road as the victim-target pair. For the remaining attack configurations, we use car→person for object-to-object, sidewalk→road for background-to-background, and road→car for background-to-object attacks.

We compare RePAIR against all defense methods, i.e., pruning, fine-tuning, Segment Cleanse, and ABL. Performance is evaluated using Attack Success Rate (ASR), benign mean intersection-over-union (mIOU-B), and attacked mIoU (mIOU-A). For BadSeg-B2O, ASR is restricted to ground-truth victim pixels inside the designated attack region. More details on metric definitions and experimental configurations are provided in Appendix A.3.

5.2 Numerical Results

Attack success rate. Table 1 reports attack success rate before and after applying the evaluated defenses for the Cityscapes dataset. Lower values indicate stronger attack mitigation. ASR measures the fraction of victim-class pixels that are misclassified as the attacker-selected target class on triggered inputs.

For DeepLabV3+, the undefended model achieves high ASR across all attacks, ranging from 60.83% to 97.41%. Fine-tuning and pruning provide little mitigation and leave the ASR high for most attacks. In contrast, RePAIR consistently reduces the ASR to below 7.86% across all evaluated attacks. This consistent performance across different attack types is not observed for the other defense methods. Segment Cleanse achieves effective mitigation for IBA and OFBA, but performs substantially worse on the BadSeg attacks compared to? our defense. Overall, RePAIR achieves the best or second-best ASR for most attacks and maintains strong mitigation consistently across the full attack set.

For SegFormer, the undefended model shows lower ASR for several attacks, with values ranging from 26.43% to 95.46%. Fine-tuning and pruning again provide limited mitigation, while Segment Cleanse also fails to consistently reduce the ASR. ABL provides more consistent mitigation across the evaluated attacks. RePAIR achieves lower ASR than all other evaluated defenses on all backdoor attacks, reducing the ASR to below 0.54% and frequently to zero. Thus, for these cases, our defense prevents the attack from successfully redirecting any victim-class pixels to the target class.

Segmentation performance. Table 1 reports the mean mIoU-B and mIoU-A across all evaluated attacks in the two rightmost columns. The clean models, without a backdoor attack or defense, achieve an mIoU-B of 78.74% for DeepLabV3+ and 80.46% for SegFormer, representing the upper reference for benign segmentation performance. Under attack, the goal is to preserve this performance on benign classes while recovering the attacked classes.

Table 1: **Cityscapes results across backdoor attacks.** ASR is the primary metric, with lower values indicating better defense performance. mIoU-A and mIoU-B are averaged across the eight attacks and reported with their minimum and maximum deviations from the mean.

Model	Method	IBA (NNI)	IBA (PRL)	ConSeg	OFBA	BadSeg-O2B	BadSeg-O2O	BadSeg-B2O	BadSeg-B2B	mIoU-A	mIoU-B
DeepLabV3+	no defense	88.56	97.41	86.23	60.83	95.49	97.03	96.87	96.82	$71.14^{+6.49}_{-7.28}$	$73.28^{+6.00}_{-10.01}$
	fine-tuning	56.36	12.35	84.54	6.31	13.42	81.91	86.36	97.69	$35.86^{+14.73}_{-16.03}$	$37.65^{+18.41}_{-21.43}$
	pruning	89.60	98.13	90.07	71.36	96.63	96.55	97.58	97.25	$70.13^{+6.82}_{-6.58}$	$72.37^{+6.00}_{-9.91}$
	Seg.-Cleanse	**0.36**	**0.89**	86.23	**0.75**	3.31	94.20	96.85	36.95	$71.14^{+6.49}_{-7.36}$	$74.14^{+5.30}_{-7.31}$
	ABL	83.94	94.55	**1.12**	0.78	17.29	90.36	86.19	92.17	$36.66^{+4.39}_{-6.48}$	$40.97^{+5.46}_{-4.08}$
	RePAIR (ours)	1.18	1.01	1.51	1.10	**2.35**	**1.00**	**0.46**	**7.86**	$73.83^{+3.97}_{-7.77}$	$74.50^{+3.17}_{-5.69}$
SegFormer	no defense	26.43	28.71	89.15	45.32	87.39	86.24	94.11	95.46	$71.95^{+7.83}_{-21.68}$	$75.80^{+4.49}_{-27.52}$
	fine-tuning	26.57	35.44	93.90	46.92	86.46	85.12	89.73	95.10	$69.54^{+8.77}_{-25.96}$	$73.05^{+5.87}_{-33.14}$
	pruning	26.80	29.45	90.06	45.64	87.53	86.22	93.76	95.51	$71.74^{+7.94}_{-22.43}$	$75.54^{+4.64}_{-28.54}$
	Seg.-Cleanse	21.07	22.71	75.46	36.98	80.67	85.30	94.11	92.91	$71.91^{+7.87}_{-20.72}$	$75.77^{+4.52}_{-26.49}$
	ABL	10.89	15.09	0.71	9.11	62.48	2.32	73.35	17.60	$20.87^{+9.14}_{-9.36}$	$22.50^{+8.80}_{-9.46}$
	RePAIR (ours)	**0.00**	**0.00**	**0.54**	**0.00**	**0.00**	**0.00**	**0.00**	**0.31**	$74.52^{+4.78}_{-21.34}$	$72.30^{+4.80}_{-19.16}$

The undefended backdoored models retain an mIoU-B of 73.28% for DeepLabV3+ and 75.80% for SegFormer, showing that the attacks primarily affect the targeted classes. In contrast, ABL substantially degrades mIoU-B, indicating that its mitigation affects benign predictions beyond the attacked classes. RePAIR maintains mIoU-B at 74.50% for DeepLabV3+ and 72.30% for SegFormer, while achieving the highest mIoU-A among the evaluated defenses. Thus, RePAIR combines recovery of the attacked classes with preservation of the model's general segmentation performance.

Other datasets. We additionally evaluate RePAIR on PASCAL VOC 2012 and ADE20K, considering only Object-to-Background attack configurations for both datasets. On ADE20K, the evaluated attack configurations do not produce a sufficiently strong backdoor effect in the undefended models to support a meaningful mitigation study. This behavior is further analyzed theoretically in Appendix A.1, and the corresponding results are reported in Table 5.

PASCAL VOC differs from Cityscapes primarily in the number of semantic classes present within each image. Images in PASCAL VOC typically contain a single foreground object class together with a background label that aggregates all other classes. This different class composition affects the learned class representations and, consequently, the information available to the defense. Across the evaluated attacks, the undefended models achieve ASR ranging from 12.21% for NNI to 99.71% for ConSeg, while RePAIR reduces the remaining ASR to between 5.41% and 19.67%. The corresponding results are reported in Table 6 and 7.

6 Conclusion

We introduced RePAIR, a post-hoc, label-free, and attack-unaware defense that exploits surviving internal victim-class evidence to localize and selectively repair backdoor-affected pixels without retraining the compromised segmentation model. Across eight attack configurations on Cityscapes, RePAIR reduces the defended ASR to 0.46%-7.86% for DeepLabV3+ and to at most 0.54% for SegFormer, with complete suppression of the measured attack success in six of the eight SegFormer settings. At the same time, RePAIR preserves substantial segmentation utility, achieving an average mIoU-B of 74.50% for DeepLabV3+ and 72.30% for SegFormer, compared with 73.28% and 75.80% for the corresponding undefended models.

Beyond the overall mitigation performance, the experiments show that effective backdoor repair can be achieved without explicitly reconstructing the trigger or modifying the compromised model. The directed-transport formulation provides a view of the attack as a structured victim-to-target shift, while the internal representation provides a model-derived reference for reversing this shift selectively. Across convolutional and transformer-based architectures, these results indicate that surviving internal semantic evidence can serve as a basis for attack-unaware localization and constrained pixel-wise repair.

AI USE STATEMENT

In this work, we used generative AI tools to assist coding, debugging, mathematical-notation consistency, and LaTeX drafting. We did not use generative AI tools to generate experimental measurements or to replace human evaluation of the theoretical claims. AI-assisted outputs were reviewed for mathematical consistency, and generated LaTeX was compiled and visually inspected. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies defenses against data-poisoning backdoors in semantic segmentation, with emphasis on safety-critical perception. Published attacks are used only as controlled evaluation mechanisms for the proposed defense. The paper does not propose a new attack or expand the attacker's capabilities. Any release of code, models, or configuration files will follow the licenses of the underlying datasets and software and will be framed for defensive research and reproducible evaluation. The study does not involve human-subject experiments.

REPRODUCIBILITY STATEMENT

We provide the implementation of RePAIR, including the model-specific feature extraction, victim-target pair selection, reference construction, localization, feature transport, and output-safety procedures. We use the publicly available Cityscapes, PASCAL VOC 2012, and ADE20K datasets and evaluate the method on DeepLabV3+ with ResNet-50 and SegFormer with MiT-B5. The attack and defense configurations, evaluation metrics, and implementation details are described in the main paper and the appendix. The code will be uploaded on Github after acceptance.

REFERENCES

Bilal Hussain Abbasi, Zirui Gong, Yanjun Zhang, Shang Gao, Antonio Robles-Kelly, and Leo Zhang. Conseg: Contextual backdoor attack against semantic segmentation, 2025. URL https://arxiv.org/abs/2507.19905.

N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784.

Senay Cakir, Marcel Gauß, Kai Häppeler, Yassine Ounajjar, Fabian Heinle, and Reiner Marchthaler. Semantic segmentation for autonomous driving: Model evaluation, dataset generation, perspective comparison, and real-time capability, 2022. URL https://arxiv.org/abs/2207.12939.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 833–851, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Bolei Zhou et al. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, pp. 113–125, New York, NY, USA,

2019. Association for Computing Machinery. ISBN 9781450376280. doi: 10.1145/3359789.3359790. URL https://doi.org.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Haoheng Lan, Jindong Gu, Philip Torr, and Hengshuang Zhao. Influencer backdoor attack on semantic segmentation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/01977a08d8e76cf4df40b5b57407c9eb-Abstract-Conference.html.

Yige Li, Xilie Xu, Ning Yu, Teng Li, and Ji Liu. Anti-backdoor learning: Training clean models on poisoned data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14930–14941. Curran Associates, Inc., 2021a.

Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, and Shu-Tao Xia. Hidden backdoor attack against semantic segmentation models. In *ICLR Workshop on Security and Safety in Machine Learning Systems*, 2021b. URL https://arxiv.org/abs/2103.04038.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses (RAID)*, pp. 273–294. Springer, 2018.

Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20960–20969, June 2023.

Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The "beatrix" resurrections: Robust backdoor detection via gram matrices. 01 2023. doi: 10.14722/ndss.2023.23069.

Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Object-free backdoor attack and defense on semantic segmentation. *Computers & Security*, 132:103365, 2023.

Zeyang Sha, Xinlei He, Pascal Berrang, Mathias Humbert, and Yang Zhang. Fine-tuning is all you need to mitigate backdoor attacks, 12 2022.

Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: A review, 2024. URL https://arxiv.org/abs/1910.07655.

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 *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pp. 707–723, 2019. doi: 10.1109/SP.2019.00031.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16473–16481, October 2021.

Guangsheng Zhang, Huan Tian, Leo Zhang, Tianqing Zhu, Ming Ding, Wanlei Zhou, and Bo Liu. Poisoning the pixels: Revisiting backdoor attacks on semantic segmentation, 2026. URL https://arxiv.org/abs/2603.16405.

A Appendix

A.1 Theory: Full Statements and Proofs

In the following, we formalize the local transport operator, distinguish its inverse from the transposed transport coupling, and describe its operational interpretation for RePAIR. The emphasis here is on the formal domain, feasibility conditions, and proof obligations underlying the theoretical claims.

A.1.1 Probability space and standing conventions

Let Ω denote the evaluated pixel population and

$$Y : \Omega \to [K] \cup \{\varnothing\} \ , \qquad \widehat{Y}^{\mathrm{head}}, \widehat{Y}^{\mathrm{rep}}, \widehat{Y}^{\mathrm{def}} : \Omega \to [K] \ ,$$

where K is the ground-truth class, $\widehat{Y}^{\mathrm{head}}$ the undefended head prediction, $\widehat{Y}^{\mathrm{rep}}$ the repaired prediction before output safety, and $\widehat{Y}^{\mathrm{def}}$ the final defended prediction. Pixels with $Y = \varnothing$ are excluded from semantic metrics.

Let T denote the event that a trigger is present, and assume $\mathbb{P}(T) > 0$ and $\mathbb{P}(T^c) > 0$. Define

$$\mathbb{P}_{\mathrm{trig}}(\cdot) = \mathbb{P}(\cdot \mid T) \tag{9}$$

as the conditional distribution over triggered inputs, and

$$\mathbb{P}_{\mathrm{cl}}(\cdot) = \mathbb{P}(\cdot \mid T^c) \tag{10}$$

as the conditional distribution over clean inputs.

For victim class v and target class $t \neq v$, assume $\mathbb{P}_{\mathrm{trig}}(Y = v) > 0$ and define

$$\mathbb{P}_v(B) := \mathbb{P}_{\mathrm{trig}}(B \mid Y = v) \tag{11}$$

as the conditional probability over triggered pixels whose ground-truth class is v. Let

$$A := \{\widehat{Y}^{\mathrm{head}} = t\} \tag{12}$$

denote the event that the undefended model predicts the target class. We then define the baseline and defended attack success rates as

$$\mathrm{ASR}_{\mathrm{base}} = \mathbb{P}_v(A) \ , \qquad \mathrm{ASR}_{\mathrm{def}} = \mathbb{P}_v(\widehat{Y}^{\mathrm{def}} = t) \ . \tag{13}$$

For theoretical evaluation, (v, t) denotes the true attacked victim-target pair, which unknown to the defense. The attack-unaware builder either releases an active runtime state with discovered pair $\widehat{P} = (\hat{v}, \hat{t})$ or abstains and releases an inactive fail-closed state. Let $R \in \{0, 1\}$ indicate whether an active state is released. We use the following conditions throughout the analysis:

$$C_{\mathrm{pair}} : \quad R = 1 \text{ and } (\hat{v}, \hat{t}) = (v, t) \ , \tag{14}$$

$$C_{\mathrm{tgt}} : \quad R = 1 \text{ and } \hat{t} = t \ . \tag{15}$$

Clearly, C_{pair} implies C_{tgt}. All pair-conditioned runtime quantities are constructed from the released state and therefore use $(\hat{v}, \hat{t})$ rather than the unknown pair (v, t). The residual-ASR decomposition below applies to the actually executed defense state without assuming C_{pair}. In contrast, statements that relate localization error to internal victim evidence for the true pair require C_{pair}. Guarantees concerning the creation of the true target class require only C_{tgt}. If $R = 0$, fail-closed inference leaves the undefended prediction unchanged.

Let $\rho_v \in \{0, 1\}$ denote the evaluation-side internal victim-evidence event for the true pair. For CNNs, $\rho_v = \mathbf{1}\{\widehat{Y}^{\mathrm{prot}} = v\}$. For transformers, $\rho_v = \mathbf{1}\{p_v^{\mathrm{int}} > \tau_{\mathrm{sem}}\}$, where τ_{sem} is the frozen, pair-specific clean-null threshold for internal victim support. In the official transformer state it is calibrated as the 0.995 empirical quantile of image-disjoint out-of-fold victim-support scores on the final safety-approved reference population. For an active released state, the implemented direct semantic localization arm uses the discovered pair,

$$M_{\mathrm{semantic}} = \mathbf{1}\{\widehat{Y}^{\mathrm{head}} = \hat{t}\}\rho_{\hat{v}} \ . \tag{16}$$

Here $\rho_{\hat{v}}$ is constructed by the same architecture-specific rule using the released pair. Let $M : \Omega \to \{0, 1\}$ denote the final localization indicator. The implemented masks preserve the semantic arm additively, so

$$M_{\text{semantic}} \leq M \ . \tag{17}$$

Under C_{pair}, the runtime quantities satisfy $\rho_{\hat{v}} = \rho_v$ and $M_{\text{semantic}} = \mathbf{1}\{\widehat{Y}^{\text{head}} = t\}\rho_v$. If $R = 0$, we set $M_{\text{semantic}} = M = 0$ consistently with fail-closed inference.

To make all statements well-defined at zero prevalence, we define

$$\eta_{\text{ref}} = \begin{cases} \mathbb{P}_v(\rho_v = 0 \mid A), & \mathbb{P}_v(A) > 0 \ , \\ 0, & \mathbb{P}_v(A) = 0 \ , \end{cases} \tag{18}$$

$$\eta_{\text{loc}} = \begin{cases} \mathbb{P}_v(M = 0 \mid A), & \mathbb{P}_v(A) > 0 \ , \\ 0, & \mathbb{P}_v(A) = 0 \ , \end{cases} \tag{19}$$

and

$$\varepsilon_{\text{rep}} = \begin{cases} \mathbb{P}_v(\widehat{Y}^{\text{def}} = t \mid M = 1, A), & \mathbb{P}_v(M = 1, A) > 0 \ , \\ 0, & \mathbb{P}_v(M = 1, A) = 0 \ . \end{cases} \tag{20}$$

Finally,

$$\delta_{\text{new}} = \mathbb{P}_v(\widehat{Y}^{\text{def}} = t, A^c) \ . \tag{21}$$

These zero-prevalence conventions change no positive-mass conditional and make all products in the subsequent bounds unambiguous.

A.1.2 Fixed-simplex geometry

The class-probability simplex and its tangent space are

$$\Delta^{K-1} = \{p \in \mathbb{R}^K_{\geq 0} : \mathbf{1}^\top p = 1\} \ , \tag{22}$$

$$\mathcal{T}_\Delta = \{d \in \mathbb{R}^K : \mathbf{1}^\top d = 0\} \ . \tag{23}$$

For pair (v, t) define

$$d_{vt} = e_t - e_v \in \mathcal{T}_\Delta \ , \tag{24}$$

where e_c denotes the one-hot basis vector for class c.

Lemma 1 (Directed victim–target displacement)**.** *Let $p^\star \in \Delta^{K-1}$ and let*

$$q = p^\star + \xi d_{vt} \ , \qquad 0 \leq \xi \leq p^\star_v \ , \tag{25}$$

with $q \in \Delta^{K-1}$. Here, ξ denotes the amount of probability mass transported from the victim coordinate v to the target coordinate t, and $d_{vt} = e_t - e_v$ is the corresponding directed victim-to-target displacement. The constraint $0 \leq \xi \leq p^\star_v$ ensures that the victim coordinate remains nonnegative. For $\psi_{vt}(p) = p_v - p_t$,

$$\psi_{vt}(q) = \psi_{vt}(p^\star) - 2\xi \ . \tag{26}$$

Moreover, with the sign convention positive on the victim side, the signed Euclidean distance to $\mathcal{H}_{vt} = \{p : p_v = p_t\}$ is

$$\text{sdist}(p, \mathcal{H}_{vt}) = \frac{p_v - p_t}{\sqrt{2}} \ . \tag{27}$$

Proof. Equation 25 gives $q_v = p^\star_v - \xi$ and $q_t = p^\star_t + \xi$, proving the margin identity. The hyperplane normal is $e_v - e_t$, whose Euclidean norm is $\sqrt{2}$. Thus, the stated signed-distance formula follows. □

Corollary 2 (Pairwise preference flip)**.** *If $p^\star_v > p^\star_t$, then the strict pairwise ordering $q_t > q_v$ holds whenever*

$$\xi > \frac{p^\star_v - p^\star_t}{2} \ . \tag{28}$$

At equality the pair is tied.

The corollary concerns only the victim-target pairwise boundary. A complete change of $\arg\max$ additionally requires q_t to exceed every other class coordinate. The construction is a local analytical model, i.e., RePAIR does not observe $p^\star$ at runtime.

A.1.3 FORMAL TRANSPORT OPERATOR AND COUPLING

Section 3 introduces the local operator $T_\xi(p) = p + \xi(e_t - e_v)$. Here we record its feasible inverse domain and the coupling construction used in the proofs.

Define

$$\mathcal{D}_\xi = \{p \in \Delta^{K-1} : p_v \geq \xi\} \ , \qquad \mathcal{R}_\xi = \{q \in \Delta^{K-1} : q_t \geq \xi\} \ . \tag{29}$$

Lemma 3 (Local inverse operator)**.** *For fixed $\xi \geq 0$, the map*

$$T_\xi : \mathcal{D}_\xi \to \mathcal{R}_\xi \ , \qquad T_\xi(p) = p + \xi(e_t - e_v) \ , \tag{30}$$

is bijective with

$$T_\xi^{-1}(q) = q + \xi(e_v - e_t) \ . \tag{31}$$

Proof. For $p \in \mathcal{D}_\xi$, the victim coordinate remains nonnegative after subtracting ξ, while all coordinates still sum to one, so $T_\xi(p) \in \Delta^{K-1}$ and its target coordinate is at least ξ. Conversely, for $q \in \mathcal{R}_\xi$, subtracting ξ from q_t and adding it to q_v yields a valid simplex point in $\mathcal{D}_\xi$. Direct substitution shows that the two maps are mutual inverses. □

Lemma 4 (One-edge coupling)**.** *Let $p \in \mathcal{D}_\xi$ and $q = T_\xi(p)$. Define $\gamma_\xi \in \mathbb{R}_{\geq 0}^{K \times K}$ by*

$$(\gamma_\xi)_{vv} = p_v - \xi \ , \qquad (\gamma_\xi)_{vt} = \xi \ , \qquad (\gamma_\xi)_{kk} = p_k \quad (k \neq v) \ ,$$

with all remaining entries zero. Then

$$\gamma_\xi \mathbf{1} = p \ , \qquad \gamma_\xi^\top \mathbf{1} = q \ . \tag{32}$$

Hence γ_ξ is a feasible coupling between the forward marginals p and q, while $\gamma_\xi^\top$ is a feasible coupling from q back to p.

Proof. The victim row sums to $(p_v - \xi) + \xi = p_v$ and every other row to p_k. The victim column sums to $p_v - \xi = q_v$, the target column to $p_t + \xi = q_t$, and every other column to $p_k = q_k$. Transposition exchanges these marginals. □

The last statement concerns a joint measure, not a function on probability vectors. Thus $\gamma_\xi^\top$ represents reversed mass flow but is not the operator T_ξ^{-1}. No optimality claim is made for either coupling without specifying a transport cost.

A.1.4 INTERNAL-REFERENCE SURVIVAL AND LOCALIZATION

Lemma 5 (Surviving victim evidence)**.** *Under the conventions in Section A.1.1,*

$$\mathbb{P}_v(A, \rho_v = 1) = (1 - \eta_{\text{ref}})\text{ASR}_{\text{base}} \ . \tag{33}$$

Proof. If $\mathbb{P}_v(A) > 0$, the multiplication rule gives $\mathbb{P}_v(A, \rho_v = 1) = \mathbb{P}_v(\rho_v = 1 \mid A)\mathbb{P}_v(A)$, which is the claimed expression. If $\mathbb{P}_v(A) = 0$, both sides are zero by definition. □

The relevant condition is therefore not complete encoder invariance. η_{ref} measures only the fraction of successful baseline attacks for which the direct internal victim-evidence event is absent.

Proposition 6 (Internal evidence bounds localization misses)**.** *Assume C_{pair} and Equation 17. Then*

$$\eta_{\text{loc}} \leq \eta_{\text{ref}} \ . \tag{34}$$

Proof. Under C_{pair}, the released pair satisfies $(\hat{v}, \hat{t}) = (v, t)$, so the implemented semantic arm is exactly $M_{\text{semantic}} = \mathbf{1}\{\widehat{Y}^{\text{head}} = t\}\rho_v$. If $\mathbb{P}_v(A) = 0$, both quantities are zero by convention. Otherwise, on A, $\rho_v = 1$ implies $M_{\text{semantic}} = 1$, and Equation 17 implies $M = 1$. Therefore $\{M = 0\} \cap A \subseteq \{\rho_v = 0\} \cap A$. Conditioning on A proves the inequality. □

Additional residual branches may reduce localization misses beyond the direct semantic arm but cannot invalidate this inclusion.

A.1.5 DIRECTED PAIR EVIDENCE AND POPULATION SEPARATION

Pair discovery is architecture specific, so we first state a generic scalar separation result. For each ordered pair (a, b), let $G_i(a, b)$ be an attack-unaware image-level evidence statistic and

$$g_{ab} := \mathbb{E}[G_i(a, b)] \ . \tag{35}$$

Given n builder images for which the statistic is defined, let

$$\widehat{g}_{ab} := \frac{1}{n} \sum_{i=1}^{n} G_i(a, b) \tag{36}$$

denote the corresponding empirical statistic. For analysis only, let ν_{ab} denote the expected value of the same statistic under a corresponding no-backdoor reference population and define

$$\beta_{ab} := g_{ab} - \nu_{ab} \ , \qquad g_{ab} = \nu_{ab} + \beta_{ab} \ . \tag{37}$$

Neither ν_{ab} nor β_{ab} is observed by the defense.

For CNNs, one primitive statistic is

$$G_i^{\mathrm{CNN}}(a, b) = \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} \mathbf{1}\{\widehat{Y}^{\mathrm{prot}}(x) = a \wedge \widehat{Y}^{\mathrm{head}}(x) = b\} \ . \tag{38}$$

For transformers, one primitive OOF statistic is a fixed upper-quantile summary of internal victim support on head-target pixels,

$$G_i^{\mathrm{TF}}(a, b) = Q_\tau\Big(p_a^{\mathrm{int}}(x) : \widehat{Y}^{\mathrm{head}}(x) = b\Big) , \tag{39}$$

when the eligible set is nonempty. Pairs for which this primitive statistic is undefined are outside the scope of the scalar proposition below. The full implementation also uses support, structural, and held-out mechanism evidence.

Proposition 7 (Sufficient scalar pair separation)**.** *Suppose (v, t) is the true attacked pair and, for every competitor $(a, b) \neq (v, t)$,*

$$|(\widehat{g}_{vt} - \widehat{g}_{ab}) - (g_{vt} - g_{ab})| \leq \delta_{\mathrm{est}} \ . \tag{40}$$

If

$$\beta_{vt} - \max_{(a,b)\neq(v,t)} \beta_{ab} > \max_{(a,b)\neq(v,t)} (\nu_{ab} - \nu_{vt}) + \delta_{\mathrm{est}} \ , \tag{41}$$

then $\widehat{g}_{vt} > \widehat{g}_{ab}$ for every competitor.

Proof. For any competitor,

$$g_{vt} - g_{ab} = (\beta_{vt} - \beta_{ab}) - (\nu_{ab} - \nu_{vt}) \ .$$

The stated condition implies $g_{vt} - g_{ab} > \delta_{\mathrm{est}}$ for every competitor. Equation 40 then gives $\widehat{g}_{vt} - \widehat{g}_{ab} > 0$. □

The result is sufficient, not necessary, and is conditional on the stated uniform score-difference error bound. It should not be read as a finite-sample guarantee for the complete transformer mapping selector, whose family and direction decisions use multiple evidence views rather than a single scalar pair ranking.

A.1.6 TRANSFORMER MAPPING RELEASE CONTRACT

Transformer mapping is modeled as a partial release rule over attack-unaware builder evidence. Let $\mathcal{F}$ denote the set of feasible undirected class families. Pair-owned out-of-fold evidence provides structural, mechanism, and family verification views for each candidate family. These views define a stable rank prior together with a proposal set

$$\mathcal{F}_{\mathrm{prop}} = \mathcal{C}_S \cup \mathcal{C}_M \cup \mathcal{C}_V \ , \tag{42}$$

where $\mathcal{C}_S$, $\mathcal{C}_M$, and $\mathcal{C}_V$ are the stable candidate cores induced by the corresponding evidence views.

If $\mathcal{F}_{\mathrm{prop}} = \varnothing$, the builder releases no mapping and returns $(R, \widehat{P}) = (0, \perp)$. Otherwise, let $F_{\mathrm{prior}} \in \mathcal{F}_{\mathrm{prop}}$ denote the leading family under the frozen stable rank prior. Let

$$\mathsf{V}_{\mathrm{base}}(F) \in \{0, 1\} \tag{43}$$

denote the frozen candidate viability test. The builder preserves F_{prior} whenever $\mathsf{V}_{\mathrm{base}}(F_{\mathrm{prior}}) = 1$.

If the prior family is not viable, a different family can be considered only through the rescue path. Let

$$\mathsf{V}_{\mathrm{res}}(F) \in \{0, 1\} \tag{44}$$

indicate that F has the required bilateral leave-one-out support together with bilateral structural consistency. The frozen family is therefore

$$\widehat{F} = \begin{cases} F_{\mathrm{prior}}, & \mathsf{V}_{\mathrm{base}}(F_{\mathrm{prior}}) = 1 \ , \\ \mathsf{Select}_{\mathrm{rank}}\left(\{F \in \mathcal{F}_{\mathrm{prop}} : \mathsf{V}_{\mathrm{res}}(F) = 1\}\right), & \mathsf{V}_{\mathrm{base}}(F_{\mathrm{prior}}) = 0 \text{ and the set is nonempty} \ , \\ \perp, & \text{otherwise} \ . \end{cases} \tag{45}$$

Here $\mathsf{Select}_{\mathrm{rank}}$ denotes selection according to the frozen family ranking. No family is released when the required viability evidence is absent.

Assume now that $\widehat{F} \neq \perp$. The two orientations of $\widehat{F}$ form the directed candidate set $\mathcal{D}(\widehat{F})$. For each $e \in \mathcal{D}(\widehat{F})$, let

$$D_{\mathrm{FE}}(e) \ , \quad D_{\mathrm{FV}}(e) \ , \quad D_{\mathrm{M}}(e) \in \{0, 1\} \tag{46}$$

denote the FullEdge, FamilyVerify, and Mechanism direction votes. Let $E_{\mathrm{M}} \in \{0, 1\}$ indicate whether the Mechanism direction evidence passes its frozen response and endpoint balance requirements.

The regular direction score is

$$V_{\mathrm{dir}}(e) = D_{\mathrm{FE}}(e) + D_{\mathrm{FV}}(e) + E_{\mathrm{M}} D_{\mathrm{M}}(e) \ . \tag{47}$$

A regular direction is released only when one orientation $e^\star \in \mathcal{D}(\widehat{F})$ is the unique orientation satisfying

$$V_{\mathrm{dir}}(e^\star) \geq 2 \ . \tag{48}$$

A separate rescue rule is available only when the Mechanism direction channel is ineligible. Let $D_{\mathrm{S}}(e) \in \{0, 1\}$ denote structural corroboration. Structural evidence is not a regular direction vote. If $E_{\mathrm{M}} = 0$, FullEdge and FamilyVerify disagree, and structural evidence corroborates the FamilyVerify orientation, the FamilyVerify orientation may be released through the rescue rule provided that the frozen family itself satisfies the rescue viability condition.

Let $e^\star$ denote the unique directed orientation accepted by either the regular rule or the rescue rule. The released transformer mapping state is

$$(R, \widehat{P}) = \begin{cases} (1, e^\star), & \widehat{F} \neq \perp \text{ and a unique direction is accepted} \ , \\ (0, \perp), & \text{otherwise} \ . \end{cases} \tag{49}$$

Proposition 8 (Transformer fail-closed mapping contract)**.** *The transformer builder releases a directed pair only if a viable family is frozen and one of its two orientations satisfies the regular direction rule or the rescue rule above. If family viability is insufficient, if direction evidence is insufficient, or if the accepted orientation is not unique, then $R = 0$ and $\widehat{P} = \perp$. Once the direction firewall has rejected the frozen family, no later family fallback can create an active mapping.*

Proof. Equation 45 returns either a viable frozen family or $\perp$. Conditional on a viable family, Equation 49 sets $R = 1$ only when exactly one orientation satisfies an authorized direction rule. Every other case returns $(R, \widehat{P}) = (0, \perp)$. Structural evidence can corroborate the rescue rule but does not contribute to the regular direction vote in Equation 47. The absence of a later family fallback completes the fail-closed alternative. □

This proposition is a deterministic release contract for the observable builder state. It does not claim that the released pair equals the unknown true attacked pair. Ground truth, attack provenance, and the oracle attacked pair are not inputs to the release rule.

A.1.7 REPAIR-STATE SAFETY

Let $\mathcal{C}_0$ be the provisional reference population after architecture-specific cross-fitting and let $V_i \in \{0, 1\}$ denote the victim-safety veto. Define

$$\mathcal{C}_{\text{safe}} = \{i \in \mathcal{C}_0 : V_i = 0\} \ . \tag{50}$$

Proposition 9 (One-way safety)**.** *Every later capacity-controlled repair bank satisfies*

$$\mathcal{C}_{\text{bank}} \subseteq \mathcal{C}_{\text{safe}} \subseteq \mathcal{C}_0 \ . \tag{51}$$

Consequently, a vetoed reference cannot be reintroduced after the safety decision.

Proof. The first inclusion is the construction rule for all later banks and fixed points. The second follows directly from the definition of $\mathcal{C}_{\text{safe}}$. □

This guarantees monotone selection, not true clean provenance. If required reference count or semantic coverage cannot be maintained, the method abstains.

A.1.8 ORTHOGONAL SEMANTIC MATCHING

Let $u(x)$ be the architecture-specific semantic coordinate used by the probability reference. Assume distinct victim and target centers $\mu_v^u \neq \mu_t^u$ and define

$$a_{vt} = \frac{\mu_t^u - \mu_v^u}{\|\mu_t^u - \mu_v^u\|_2} \ , \qquad \Pi_{vt}^{\perp} = I - a_{vt} a_{vt}^{\top} \ . \tag{52}$$

Lemma 10 (Orthogonal matching ignores pair-axis displacement)**.** *For semantic coordinates z, r and scalar c,*

$$\Pi_{vt}^{\perp}[(z + c a_{vt}) - r] = \Pi_{vt}^{\perp}(z - r) \ . \tag{53}$$

Proof. Since $\|a_{vt}\|_2 = 1$, $\Pi_{vt}^{\perp} a_{vt} = (I - a_{vt} a_{vt}^{\top}) a_{vt} = 0$. □

Thus orthogonal retrieval is exactly invariant to displacement along the chosen victim–target axis. It does not assume that a real attack lies entirely on that axis.

A.1.9 MINIMAL VICTIM-TARGET LOGIT CORRECTION.

Consider a residual pair-logit substep applied after any feature transport. For pre-correction logits $\ell \in \mathbb{R}^K$, let $q = \mathrm{softmax}(\ell)$ and define

$$m_\ell = \ell_v - \ell_t = \log \frac{q_v}{q_t} \ . \tag{54}$$

Let $m_{vt}^{\text{rep}} > 0$ be the frozen required victim-over-target margin, calibrated from safety-approved victim reference pixels.

Proposition 11 (Minimum-ℓ_2 full pair correction)**.** *Among corrections $\Delta \in \mathbb{R}^K$ satisfying*

$$\Delta_k = 0 \ , \quad (k \notin \{v, t\}) \ , \tag{55}$$

$$\Delta_v + \Delta_t = 0 \ , \tag{56}$$

$$(\ell_v + \Delta_v) - (\ell_t + \Delta_t) \geq m_{vt}^{\text{rep}} \ , \tag{57}$$

the unique minimum-ℓ_2 correction is

$$\Delta_v = \frac{d}{2} \ , \qquad \Delta_t = -\frac{d}{2} \ , \qquad d = [m_{vt}^{\text{rep}} - (\ell_v - \ell_t)]_+ \ . \tag{58}$$

Proof. Write $r = m_{vt}^{\text{rep}} - (\ell_v - \ell_t)$. The zero-sum constraint gives $\Delta_t = -\Delta_v$, so feasibility is $2\Delta_v \geq r$ and $\|\Delta\|_2^2 = 2\Delta_v^2$. If $r \leq 0$, $\Delta_v = 0$ is feasible and uniquely minimizes the norm. If $r > 0$, the smallest feasible value is $\Delta_v = r/2$. Combining both cases yields Equation 58. □

Corollary 12 (Effect of the full pair correction on odds)**.** *Let q and q' be the softmax distributions immediately before and after the full correction in Proposition 11. Then*

$$\frac{q'_v/q'_t}{q_v/q_t} = e^d \ . \tag{59}$$

The actual runtime may use a branch-dependent fraction of this residual correction and may subsequently apply a fail-closed projection. Feature transport can also change logits of classes outside $\{v, t\}$ indirectly, i.e., the "pair-only" statement applies only to this explicit residual-logit substep.

A.1.10 Residual attack success

Proposition 13 (Residual-ASR bound)**.** *Under the definitions in Section A.1.1,*

$$\mathrm{ASR}_{\mathrm{def}} \le \mathrm{ASR}_{\mathrm{base}} \left[\eta_{\mathrm{loc}} + (1 - \eta_{\mathrm{loc}})\varepsilon_{\mathrm{rep}}\right] + \delta_{\mathrm{new}} \ . \tag{60}$$

Proof. Partition $\{\widehat{Y}^{\mathrm{def}} = t\}$ into A and A^c. The A^c contribution is exactly δ_{new}. If $\mathbb{P}_v(A) = 0$, the A contribution is zero and the claim follows. Otherwise, partition A by M. On $A \cap \{M = 0\}$, the defended-target event has probability at most $\mathrm{ASR}_{\mathrm{base}}\eta_{\mathrm{loc}}$. On $A \cap \{M = 1\}$, the exact contribution is $\mathrm{ASR}_{\mathrm{base}}(1 - \eta_{\mathrm{loc}})\varepsilon_{\mathrm{rep}}$, including the zero-mass convention of Equation 20. Adding the disjoint contributions proves the bound. □

Corollary 14 (Reference-aware residual-ASR bound)**.** *Assume C_{pair} and Equation 17. Then*

$$\mathrm{ASR}_{\mathrm{def}} \le \mathrm{ASR}_{\mathrm{base}} \left[\eta_{\mathrm{ref}} + (1 - \eta_{\mathrm{ref}})\varepsilon_{\mathrm{rep}}\right] + \delta_{\mathrm{new}} \ . \tag{61}$$

Proof. Proposition 6 gives $\eta_{\mathrm{loc}} \le \eta_{\mathrm{ref}}$. For fixed $\varepsilon_{\mathrm{rep}} \in [0, 1]$, the function $f(\eta) = \eta + (1 - \eta)\varepsilon_{\mathrm{rep}}$ is nondecreasing because $f'(\eta) = 1 - \varepsilon_{\mathrm{rep}} \ge 0$. Substitute this inequality into Proposition 13. □

A.1.11 Effect of the No New Target Veto

Let $\mathcal{S} = \{x : M(x) = 1\}$ denote the supported repair region. The No New Target Veto is applied to the discovered target $\hat{t}$ and enforces

$$x \notin \mathcal{S} \ , \quad \widehat{Y}^{\mathrm{head}}(x) \ne \hat{t} \quad \Longrightarrow \quad \widehat{Y}^{\mathrm{def}}(x) \ne \hat{t} \ . \tag{62}$$

Proposition 15 (New-target support restriction)**.** *Assume C_{tgt}, i.e., an active state is released and the discovered target satisfies $\hat{t} = t$. Within $\mathbb{P}_v$,*

$$\delta_{\mathrm{new}} \le \mathbb{P}_v(x \in \mathcal{S}, A^c) \ . \tag{63}$$

If $\mathcal{S} \subseteq A$, then $\delta_{\mathrm{new}} = 0$.

Proof. Under C_{tgt}, $\hat{t} = t$, so Equation 62 excludes $\{\widehat{Y}^{\mathrm{def}} = t, A^c\}$ outside $\mathcal{S}$. Hence any event counted by δ_{new} must lie in $\mathcal{S} \cap A^c$. If $\mathcal{S} \subseteq A$, this set is empty. □

Corollary 16 (No new true-target attacks under target-correct transformer localization)**.** *Assume C_{tgt}. Then the transformer runtime satisfies*

$$M_{\mathrm{TF}} \subseteq \{\widehat{Y}^{\mathrm{head}} = \hat{t}\} = \{\widehat{Y}^{\mathrm{head}} = t\} = A \ .$$

Consequently,

$$\delta_{\mathrm{new}} = 0 \ . \tag{64}$$

Proof. Every transformer localization branch is restricted to baseline pixels predicted as the discovered target $\hat{t}$, and the final localization mask is their union. Hence $M_{\mathrm{TF}} \subseteq \{\widehat{Y}^{\mathrm{head}} = \hat{t}\}$. Under C_{tgt} we have $\hat{t} = t$, so $M_{\mathrm{TF}} \subseteq A$. Proposition 15 therefore gives $\delta_{\mathrm{new}} = 0$. □

A.1.12 CLEAN-SIDE INTERVENTION

Define

$$\kappa_{\mathrm{cl}} = \mathbb{P}_{\mathrm{cl}}(M = 1) \tag{65}$$

and

$$\delta^{\mathrm{cl}}_{\mathrm{spill}} = \mathbb{P}_{\mathrm{cl}}(\widehat{Y}^{\mathrm{def}} \neq \widehat{Y}^{\mathrm{head}}, M = 0) \ . \tag{66}$$

Proposition 17 (Clean prediction-change bound)**.**

$$\mathbb{P}_{\mathrm{cl}}(\widehat{Y}^{\mathrm{def}} \neq \widehat{Y}^{\mathrm{head}}) \leq \kappa_{\mathrm{cl}} + \delta^{\mathrm{cl}}_{\mathrm{spill}} \ . \tag{67}$$

Proof. Partition the prediction-change event by M. The localized contribution is at most $\mathbb{P}_{\mathrm{cl}}(M = 1) = \kappa_{\mathrm{cl}}$ and the non-localized contribution is exactly $\delta^{\mathrm{cl}}_{\mathrm{spill}}$. □

Exact preservation follows when both terms are zero, but this is not assumed. The probability-residual branch is calibrated under a finite empirical activation budget on the safety-approved reference population. This budget does not constitute a population-level guarantee on future clean data.

A.1.13 DEPENDENCY SUMMARY

Table 2: Role and scope of the theoretical results.

Result	Main condition	Consequence
Lemma 1	local directed transport	pair margin decreases
Cor. 2	positive initial pair margin	pairwise preference flips
Lemma 3	fixed feasible transport amount	local inverse exists
Lemma 4	feasible one-edge shift	coupling representation
Lemma 5	evidence definition	surviving victim evidence
Prop. 6	pair-correct state and additive semantic arm	$\eta_{\mathrm{loc}} \leq \eta_{\mathrm{ref}}$
Prop. 7	scalar evidence gap	sufficient scalar separation
Prop. 8	family viability and direction agreement	fail-closed mapping release
Prop. 9	monotone filtering	vetoed references cannot return
Lemma 10	nondegenerate pair axis	axis-invariant matching
Prop. 11	full pair-only correction	minimum-ℓ_2 substep
Prop. 13	localization/repair terms	residual-ASR bound
Cor. 16	target-correct state and transformer support $\subseteq A$	$\delta_{\mathrm{new}} = 0$
Prop. 17	selective intervention	clean-change bound

Remark 18 (Scope of the theory). The analysis does not claim that a real backdoor changes only two class probabilities, that its internal effect is one-dimensional, or that the unknown attack operator is observed exactly. The inverse operator and the transposed coupling are distinct mathematical objects, as stated in the main text and formalized above. It also does not claim that the safety-approved reference population is genuinely clean, or that a scalar population pair score reproduces the complete transformers mapping selector. The transport model is a local abstraction and the coupling is feasible rather than optimal. Pair-separation results are sufficient, not necessary. Proposition 8 states a deterministic release contract conditional on observable builder evidence and is not a distribution-free theorem that the released pair equals the true attacked pair. Calibration budgets are empirical reference-population constraints. Clean preservation depends on localization and spillover rather than on an oracle clean-input gate.

A.2 ADDITIONAL METHOD DETAILS

This subsection specifies the implementation details required to reproduce the operational pipeline in Section 4. Throughout this section, (v, t) denotes the operational victim-target pair released by the attack-unaware builder. In the preceding theoretical analysis, (v, t) instead denotes the true attacked pair, while $(\hat{v}, \hat{t})$ denotes the released pair.

We use *safety-approved* or *clean-like* reference rather than *clean reference*, since the defense has no access to true clean/poison provenance.

A.2.1 INFORMATION BOUNDARY AND STATE CONSTRUCTION

RePAIR receives a frozen segmentation model and an unlabeled image pool that may contain triggered samples. Semantic ground truth, clean/poison provenance, attack masks, clean twins, attack identity, poisoning rate, and the true victim–target pair are excluded from all defense decisions. Annotation fields are removed before every forward pass contributing to pair discovery, reference construction, calibration, localization, or repair. Oracle information is accessed only after the official state has been frozen and serialized and is used exclusively for evaluation and diagnostics.

State construction proceeds through three fail-closed stages: attack-unaware mapping inference, construction of an image-disjoint safety-approved reference population, and reconstruction of the frozen runtime state from that population. Failure to establish sufficient mapping confidence, reference capacity, or semantic coverage leaves the defense inactive.

A.2.2 ARCHITECTURE-SPECIFIC SEMANTIC EVIDENCE

For CNNs, normalized features from the final backbone representation define the CNN semantic reference. For head-pseudo-labeled reference pixels $\mathcal{P}_k$, class centroids are

$$\mu_k = \frac{1}{|\mathcal{P}_k|} \sum_{x \in \mathcal{P}_k} z(x) \ , \qquad \widehat{Y}^{\mathrm{prot}}(x) = \arg\max_k \mathrm{sim}(z(x), \mu_k) \ . \tag{68}$$

Directed internal-to-head disagreement for candidate pair (v, t) is measured by

$$G_i^{\mathrm{CNN}}(v,t) = \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} \mathbf{1}\widehat{Y}^{\mathrm{prot}}(x) = v \wedge \widehat{Y}^{\mathrm{head}}(x) = t \ . \tag{69}$$

The final builder resolves the target before selecting the victim. Once the target is confidence-resolved, victim proposals are restricted to that target and combine the established pair-evidence score with complementary victim-retention and victim-interior cues. A proposal is accepted only when it is sufficiently separated from competing candidates, otherwise mapping fails closed. No class identifier or attack-specific pair is hard-coded. Every selected pair must subsequently pass the unchanged reference-safety pipeline.

Transformers combine semantic evidence from L selected internal representations. For representation ℓ, victim and target centroid similarities are

$$c_{\ell,v}(x) = \mathrm{sim}(z_\ell(x), \mu_{\ell,v}) \ , \qquad c_{\ell,t}(x) = \mathrm{sim}(z_\ell(x), \mu_{\ell,t}) \ . \tag{70}$$

The selected layers are concatenated into

$$\phi_{v,t}(x) = \left[c_{1,v}(x), \ldots, c_{L,v}(x), c_{1,t}(x), \ldots, c_{L,t}(x)\right]^\top \ , \tag{71}$$

and a regularized logistic readout

$$r_{v,t}(x) = \sigma\left(w^\top \widetilde{\phi}_{v,t}(x) + b\right) \tag{72}$$

provides the internal target compatibility. where w and b denote the learned weight vector and bias of the regularized logistic readout, respectively, and $\widetilde{\phi}_{v,t}(x)$ is the standardized joint multi-layer feature representation for pair (v, t).The corresponding victim support is

$$p_v^{\mathrm{int}}(x) = 1 - r_{v,t}(x) \ . \tag{73}$$

Training labels for this readout are generated exclusively from frozen head pseudo-labels.

Transformers mapping first forms stable undirected family proposals from structural, mechanism, and family-verification evidence. A frozen rank prior is retained whenever its candidate passes the base viability test. If it is not viable, replacement is permitted only through the rescue path, which requires bilateral leave-one-out support together with bilateral structural consistency. Direction is then authorized within the frozen family using FullEdge, FamilyVerify, and an eligible Mechanism vote. Structural evidence does not act as a regular direction vote and is used only to corroborate the rescue rule. Insufficient family viability or direction evidence leaves the mapping inactive. The formal release contract is given in Proposition 8.

A.2.3 REFERENCE-STATE SAFETY

Let $\mathcal{C}_0$ denote the provisional reference population and $V_i \in \{0,1\}$ the victim-safety veto. The safety-approved population is

$$\mathcal{C}_{\text{safe}} = \{i \in \mathcal{C}_0 : V_i = 0\} \ . \tag{74}$$

All later repair banks satisfy

$$\mathcal{C}_{\text{bank}} \subseteq \mathcal{C}_{\text{safe}} \subseteq \mathcal{C}_0 \ . \tag{75}$$

Thus, a vetoed reference cannot be reintroduced by refinement, capacity restriction, fallback, or backfilling.

Safety statistics are estimated image-disjointly. References with severe victim-evidence collapse, or moderate collapse accompanied by unusually strong victim-to-target rupture, are rejected. State construction requires at least 24 safety-approved images and at least 75% semantic-class coverage.

Capacity restriction is also fail-closed. If a compressed reference core loses sufficient victim or target prototype support, the implementation may retry the same capacity check on the larger already safety-approved reservoir. This fallback does not relax thresholds, re-admit vetoed samples, or introduce backfill.

A.2.4 LOCALIZATION AND REPAIR

For CNNs, runtime localization combines direct semantic contradiction, reference-calibrated geometric residuals, and a semantic probability residual:

$$M_{\text{CNN}} = M_{\text{semantic}} \vee M_{\text{geom}} \vee M_{\text{prob}} \ . \tag{76}$$

For transformers, the final support is the sequential union

$$M_{\text{TF}} = M_{\text{semantic}} \vee M_{\text{src}} \vee M_{\text{tgt}} \vee M_{\text{traj}} \vee M_{\text{prob}} \ , \tag{77}$$

with

$$M_{\text{TF}} \subseteq \{\widehat{Y}^{\text{head}} = t\} \ . \tag{78}$$

Both architectures use the same final probability-residual principle. For discovered pair (v,t), define the target-over-victim probability excess

$$d_p(x) = q_t(x) - q_v(x) \ . \tag{79}$$

Relative to the expected excess $\widehat{d}_{\text{ref}}(x)$ obtained from the safety-approved semantic reference,

$$D_{\text{prob}}(x) = \Big[d_p(x) - \widehat{d}_{\text{ref}}(x)\Big]_+ \ . \tag{80}$$

The residual, δ_{prob}, is calibrated in four image-disjoint folds under the fixed empirical reference-pixel budget.

CNNs repair localized features from the final backbone layer toward matched safety-approved victim references and rerun the unchanged remainder of the network. Transformers transports localized pixels toward its frozen multi-layer victim manifold. Both architectures may subsequently apply a conservative pair-specific output correction. No network parameter is retrained or permanently modified.

Finally, feature repair may influence predictions outside directly localized cells. Let $\widehat{Y}^{\text{rep}}$ denote the prediction immediately before the output veto and $\mathcal{S} = \{x : M(x) = 1\}$. RePAIR reverts to the undefended prediction whenever

$$x \notin \mathcal{S} \ , \qquad \widehat{Y}^{\text{head}}(x) \neq t \ , \qquad \widehat{Y}^{\text{rep}}(x) = t \ . \tag{81}$$

Thus, the repair cannot introduce a new unsupported target prediction outside the localized support.

A.3 Details on experimental setting

Datasets. We consider three datasets for evaluation, namely Cityscapes Cordts et al. (2016), ADE20K (et al., 2017), and PASCAL VOC 2012 (Everingham et al., 2015). Cityscapes consists of 2,975 training and 500 validation images with a fixed resolution of 2048×1024 pixels and defines 19 semantic classes for evaluation. The dataset focuses on densely annotated urban street scenes and contains a diverse set of simultaneously occurring object and stuff classes, such as roads, buildings, vegetation, pedestrians, and vehicles.

ADE20K is a substantially larger and more scene-centric benchmark, comprising 20,210 training and 2,000 validation images with variable native resolutions and dense pixel-level annotations across 150 semantic classes, including both object and stuff categories.

We further evaluate RePAIR on PASCAL VOC 2012, whose semantic segmentation benchmark comprises 1,464 training and 1,449 validation images with variable image resolutions, typically around 500×375 pixels, and defines 20 foreground classes plus a background class. In contrast to the more scene-centric Cityscapes and ADE20K datasets, PASCAL VOC is predominantly object-centric: images generally focus on one or more salient foreground objects belonging to the predefined object classes, while pixels not assigned to these classes are represented by a single background label.

Since both, PASCAL VOC and ADE20K, contain images with varying native resolutions, we resize all images to 512×512 pixels. We use ADE20K as an attack-effectiveness negative control. since the evaluated attacks do not produce a sufficiently strong backdoor effect on this dataset, RePAIR is not applied.

Segmentation models. We evaluate two representative segmentation architectures: DeepLabV3+ with a ResNet-50 backbone (Chen et al., 2018) and SegFormer with a MiT-B5 backbone (Xie et al., 2021). DeepLabV3+ is trained for 80,000 iterations on Cityscapes and ADE20K and for 40,000 iterations on PASCAL VOC 2012, while SegFormer is trained for 160,000 iterations on all datasets. The other training parameters are as specified in the repository[1].

Backdoor attacks. We consider IBA with NNI and PRL (Lan et al., 2024), ConSeg (Abbasi et al., 2025), OFBA (Mao et al., 2023), and BadSeg (Zhang et al., 2026). All attacks are trained with a 10% poisoning rate, and trigger configurations follow the corresponding attack papers.

For IBA, we use the Hello Kitty trigger and the victim-to-target relabeling described in the original method. IBA-NNI places the trigger on non-victim pixels close to the victim region, using a 15×15 trigger on PASCAL VOC and a 55×55 trigger on Cityscapes. IBA-PRL additionally applies Pixel Random Labeling (PRL), where 50,000 additional pixels are assigned randomly sampled labels from the same segmentation mask.

OFBA uses a 29×29 black-and-white checkerboard trigger with visibility coefficient $\lambda = 0.4$, placed entirely within an object region. The corresponding object is relabeled to the target class, while the attacked object and trigger position are varied during training.

ConSeg uses a naturally occurring object as a semantic trigger and relabels the victim class as the target class. It additionally modifies victim-region labels using target-related co-occurring classes, with five such classes and four replaced pixels per class in the default configuration.

BadSeg jointly optimizes the trigger and victim-target pair using a surrogate segmentation model. We evaluate its object-to-object (O2O), object-to-background (O2B), background-to-object (B2O), and background-to-background (B2B) configurations.

Baseline defenses. For the defense benchmarks, we use the method-specific parameters listed in Table 3, following their original configurations.

Evaluation metrics. We evaluate attack effectiveness and segmentation utility using Attack Success Rate (ASR), benign mean intersection-over-union (mIoU-B), and attacked mIoU (mIoU-A). ASR measures the fraction of ground-truth victim-class pixels on triggered inputs that are predicted

[1] https://github.com/open-mmlab/mmsegmentation

Table 3: Hyperparameters of the benchmark defenses.

Defense	Parameter	Value
Fine-tuning	Clean data fraction	10%
	Epochs	10
Pruning	Pruning ratio	10%
	Criterion	Least-active final-layer channels
ABL	Isolation epochs	20
	Isolation ratio p	0.01
	γ	0.5
	Batch / scoring batch	$8/8$
	Unlearning epochs	5
	Unlearning optimizer	SGD
	Unlearning learning rate	5×10^{-4}
	Recovery epochs	0
	Clean-loss guard	32 samples, threshold $3\times$
Segment Cleanse	Clean images	1000
	Working resolution	512×1024
	Target knowledge	Unknown; all alternatives scanned
	Reverse optimizer	Adam
	Reverse learning rate	5×10^{-3}
	Regularization α	10^{-4}
	Maximum reverse epochs	100
	Early-stop warm-up / patience	$10/5$
	Relative threshold	10^{-4}
	Parallel target candidates	2
	Mitigation	Decoder-only
	Mitigation optimizer	SGD
	Mitigation learning rate	10^{-4}
	Mitigation epochs	20

as the attacker-selected target class. For BadSeg-B2O, where the attack is restricted to a designated spatial region, ASR is computed only within the corresponding evaluation region.

mIoU-A measures the mean IoU on triggered inputs after excluding the pixels used to evaluate the attack. For class-wide attacks, all ground-truth victim-class pixels are excluded, whereas for BadSeg-B2O only the designated attack region is excluded. mIoU-B measures the mean IoU on clean inputs. In the literature, mIoU-A and mIoU-B are also referred to as Poisoned Benign Accuracy (PBA) and Clean Benign Accuracy (CBA), respectively (Lan et al., 2024; Abbasi et al., 2025).

Lower ASR indicates stronger attack mitigation, while higher mIoU-A and mIoU-B indicate better segmentation utility.

Hyperparameter selection. All RePAIR hyperparameters are treated as global method settings and fixed independently of the evaluated attack instance. We do not select parameters based on attack identity, the oracle victim-target pair, clean/poison provenance, attack masks, ground-truth segmentation labels, or evaluation ASR. Structural parameters, such as minimum reference capacity, semantic coverage requirements, retrieval sizes, and calibration budgets, are fixed across attacks. Data-dependent thresholds are reconstructed only from the frozen safety-approved reference population using image-disjoint calibration and remain fixed during evaluation. Thus, attack-specific test performance is never used to tune the repair state.

Run-specific preprocessing, checkpoints, random seeds, and execution parameters are provided in the experiment configurations summarized in Table 4.

A.4 Additional Experimental Results

ADE20K. ADE20K contains 150 semantic classes and constitutes a substantially different semantic setting from Cityscapes and PASCAL VOC. In our experiments, the evaluated attack con-

Table 4: Principal global state-construction and inference parameters for Cityscapes.

Parameter	Variable	Value
Shared		
Image-disjoint safety/calibration folds	K_{fold}	4
Minimum safety-approved images	$N_{\text{safe}}^{\min}$	24
Minimum semantic class coverage	γ_{cov}	75%
Semantic probability budget	δ_{prob}	0.05
DeeplabV3+		
Target-authority winner ratio	τ_{tgt}	4.00
Legacy victim winner ratio	τ_{vic}	1.20
Legacy victim robust MAD-z	$z_{\text{vic}}^{\min}$	1.00
Geometry-fallback robust MAD-z	$z_{\text{geo}}^{\min}$	2.00
Reference-bank cap (per relevant class)	$N_{\text{ref}}^{\max}$	4096
Self-counterfactual strength	λ_{cf}	0.25
Base residual allowance	δ_{base}	0.02
Auxiliary residual allowance	δ_{aux}	0.005
SegFormer		
Semantic victim-support quantile	q_{sem}	0.995
Repair-margin quantile	q_{rep}	0.10
Repair-margin quantile	q_{rep}	0.10
Minimum suspicious OOF group (images)	$N_{\text{OOF}}^{\min}$	12
Manifold retrieval neighbors	k_{man}	8
Trajectory residual budget	δ_{traj}	0.005

Table 5: **ADE20K Attack Results** Undefended attack success rate (ASR, %). RePAIR is not applied because the evaluated attacks do not establish a meaningful backdoor effect.

Model	Method	IBA (NNI)	IBA (PRL)	ConSeg	OFBA	BadSeg-O2B
DeepLabV3+	no defense	3.02	3.22	2.88	2.40	2.58
SegFormer	no defense	2.51	2.38	4.93	2.48	2.43

figurations do not establish a meaningful backdoor effect in the undefended ADE20K models. The victim-target pair bein used on ADE20K is car-road.

We therefore do not treat ADE20K as an additional defense benchmark. A defense result is only informative when the underlying attack first achieves a substantive victim-to-target effect; reducing an already negligible ASR would not demonstrate successful backdoor removal.

Accordingly, the official ADE20K evaluation is baseline-only. RePAIR state construction and repair are not applied. We report the undefended ASR to document the attack effectiveness of the evaluated configurations.

The ADE20K results should therefore be interpreted as an attack-effectiveness control rather than as evidence of defense performance. They show that, under the evaluated configurations, the prerequisite backdoor behavior does not emerge strongly enough to support a meaningful mitigation experiment.

PASCAL VOC 2012. PASCAL VOC provides a challenging transfer setting because the background class occupies a large fraction of many images and represents a heterogeneous collection of visual content rather than a compact semantic object class. Consequently, natural victim-to-background relations can compete with the directed evidence produced by an O2B backdoor.

Unlike the Cityscapes experiments, all attacks considered on VOC use the same O2B-style semantic relation. The victim–target pair is used only to compute the reported ASR and is never supplied to RePAIR. The chosen victim-target pair for VOC is person-background.

Table 6 evaluates whether the attack-unaware state builder can recover a usable repair state despite the strongly background-dominated semantic distribution of VOC. The reported defended values

Table 6: **PASCAL VOC 2012.** Undefended and defended attack success rate (ASR, %).

Model	Attack	No defense	RePAIR
DeeplabV3+	IBA (NNI)	12.21	5.41
	IBA (PRL)	15.03	5.45
	ConSeg	16.12	7.05
	OFBA	99.71	8.51
	BADSEG-O2B	92.14	19.67
SegFormer	IBA (NNI)	4.83	0.12
	IBA (PRL)	7.26	0.19
	ConSeg	6.67	0.15
	OFBA	98.42	0.50
	BADSEG-O2B	95.78	0.34

Table 7: **PASCAL VOC 2012 segmentation utility.** Mean IoU on clean inputs (mIoU-B) and triggered inputs (mIoU-A).

Model	mIoU-A	mIoU-B
DeeplabV3+ / ResNet-50	$66.24^{+2.71}_{-4.99}$	$58.21^{+1.48}_{-2.05}$
SegFormer / MiT-B5	$84.45^{+1.28}_{-3.34}$	$76.17^{+0.22}_{-0.62}$

correspond only to states released by the fail-closed construction procedure. No oracle victim or target information is used during state construction.

To quantify the utility retained after repair, we additionally report clean and attacked segmentation performance in Table 7.