BACKDOOR SUCCESS OUTSIDE THE DATA MANIFOLD: A GEOMETRIC STUDY OF CLIP BACKDOOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Language–Image Pretraining (CLIP) has substantially advanced vision–language representation learning through self-supervised pretraining on web-scale image–text pairs. However, its heavy reliance on uncurated web data leaves it highly susceptible to backdoor attacks. Existing CLIP backdoor attacks can achieve high attack success rates, but they often leave a detectable geometric trace, where backdoor representations are pushed outside the clean data manifold, making them vulnerable to representation-based defenses. We mainly attribute this to a pervasive geometric phenomenon across CLIP backdoors, named *backdoor representation off-manifoldness*. In this work, we trace the phenomenon to the modality gap between CLIP's image and text embedding spaces because cross-modal backdoor objective pulls a triggered image(text) representation toward a textual(image) target, and a large fraction of the backdoor gradient lies outside the subspace supported by image(text) representations. Guided by this insight, we propose the *On-Manifold Backdoor Attack* (OMA), which retains the standard cross-modal backdoor objective while regularizing triggered representations against an intra-modal target feature bank via a principal-component normal loss and a local mean-squared-error loss. Across ImageNet-1K, ImageNet-R, and ImageNet-Sketch, OMA maintains a near-100% undefended attack success rate while effectively suppressing anomaly detection signals. Our study connects CLIP backdoor detectability directly to the modality gap and demonstrates that attack success can be decoupled from off-manifold detection signatures. Our code is in the supplementary material and will be open-sourced upon acceptance.

1 INTRODUCTION

Contrastive vision–language models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn transferable image and text representations from web-scale, weakly curated image–text pairs. By aligning matched pairs and separating mismatched ones in a shared representation space, these models support zero-shot recognition and cross-modal retrieval without task-specific labels (Radford et al., 2021; Jia et al., 2021). Beyond this training-free use, CLIP also provides a strong initialization for supervised downstream fine-tuning (Dong et al., 2022), while its shared semantic space has been extended beyond global recognition to region-level open-vocabulary perception (Zhong et al., 2021).

Despite the remarkable empirical success and zero-shot capabilities of contrastive vision–language pretraining, recent studies demonstrate that CLIP is highly vulnerable to backdoor and poisoning attacks (Carlini & Terzis, 2022; Carlini et al., 2024; Liang et al., 2024). Because contrastive pretraining relies on web-scale, uncurated data collection, an adversary can covertly inject a small fraction of malicious image–text pairs into the pretraining corpus (Carlini & Terzis, 2022; Liang et al., 2024). Specifically, a CLIP backdoor implants a cross-modal trigger–target association within the joint representation space: at inference time, trigger-bearing inputs (e.g., poisoned images) are misdirected toward an attacker-chosen target representation in the opposite modality (e.g., a specific text label), while clean inputs retain their expected semantic alignment (Song et al., 2026). As CLIP increasingly serves as a foundational upstream model, such backdoor vulnerabilities seamlessly propagate to diverse downstream tasks and real-world deployments through fine-tuning and zero-shot transfer, posing widespread security risks.

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.

To counter this threat, several test-time defenses, such as DetectCLIP (Huang et al., 2025), BDetCLIP (Niu et al., 2025), and Subspace Detection (Song et al., 2026), have been proposed to identify poisoned inputs by inspecting geometric and density anomalies in the representation space. Their empirical success raises a deeper question: **What shared geometric property of backdoor representations underlies this detectability?** Understanding this underlying geometry is crucial, as avoiding such geometric signatures could enable the design of more stealthy backdoors that evade current detection paradigms. In this paper, we examine the CLIP representation space through a geometric lens and uncover a pervasive phenomenon across existing attacks, which we term *backdoor representation off-manifoldness*. As conceptually illustrated in Figure 1, triggered backdoor representations deviate substantially from the clean data manifold, effectively behaving as off-manifold outliers that lie neither on the clean image manifold nor on the clean text manifold. We trace this off-manifold behavior to the well-known *modality gap* in vision–language models (Liang et al., 2022): image and text representations are not fully co-located in the embedding space but remain separated by a distinct geometric gap. Consequently, an opposite-modal target can contain substantial components weakly supported by the receiving modality, driving cross-modal backdoors to push triggered representations along unsupported normal directions.

Guided by this mechanism, we introduce the *On-Manifold Backdoor Attack* (OMA), which retains the standard cross-modal backdoor objective while constraining triggered representations with an intra-modal target bank. Because distribution matching is difficult with few poisoned samples, OMA uses two complementary surrogates: a PCA-based normal loss penalizes components orthogonal to the estimated target support, while a local mean-squared-error (MSE) loss anchors each representation to nearby target examples rather than a global class centroid. Together, these terms suppress unsupported cross-modal drift and steer triggered representations toward populated target regions without removing the target-alignment objective.

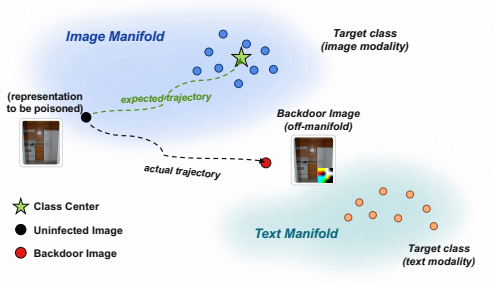


Figure 1: Off-manifold intuition: a backdoor representation is an outlier relative to the clean image support $\mathcal{M}_I$.

Our main contributions are summarized as:

- **Phenomenon.** We define backdoor representation off-manifoldness and show that it is prevalent across existing cross-modal CLIP backdoor attacks.
- **Mechanism.** We trace off-manifoldness to the training mechanism of cross-modal backdoors: gradients induced by cross-modal targets may contain components outside the support manifold of the attacked modality.
- **Method.** We introduce the *OMA* attack, which maintains attack success while substantially reducing observable off-manifold signatures. Extensive evaluations across diverse datasets and defenses validate its effectiveness.

2 Related Work

Backdoor attacks on CLIP. Backdoor attacks on CLIP exploit contrastive alignment to associate an adversarial trigger with attacker-chosen cross-modal semantics while preserving clean utility. Early work showed that poisoning image–text pairs can implant targeted behavior into contrastive encoders (Carlini & Terzis, 2022), while MM-Poison extended this threat to attacks that manipulate either side of a multimodal encoder (Yang et al., 2023b). Subsequent methods strengthen attacks via dual-embedding-guided visual trigger optimization (Liang et al., 2024), trigger-aware prompt learning (Bai et al., 2024), or text-based poisoning of CLIP pre-training corpora (Yao et al., 2025). Despite their different trigger modalities and optimization strategies, these attacks share a cross-modal objective: they pull a triggered representation toward an attacker-specified target in the opposite modality. Their evaluations thus emphasize attack success and clean utility, leaving open whether the triggered representation is supported by natural data from its own modality. We study this overlooked geometric property and propose to preserve same-modal support explicitly.

Backdoor defenses on CLIP. Existing defenses protect CLIP at different stages of its lifecycle. During pre-training, robust contrastive objectives reduce the influence of suspicious image–text pairs (Yang et al., 2023a; 2024). Given a potentially backdoored model, post-training defenses instead restore benign alignment through clean-data fine-tuning or representation recovery (Bansal et al., 2023; Singh et al., 2026). A complementary line of work detects triggered inputs at inference time by exposing backdoor-sensitive representation components (He et al., 2025), measuring reconstruction inconsistencies (Hou et al., 2025), or identifying local density anomalies in the CLIP embedding space (Huang et al., 2025). Although these methods show that successful backdoors often leave detectable representation artifacts, they do not establish why cross-modal poisoning produces such artifacts or whether an attack can avoid them while retaining high attack success. Our analysis attributes this behavior to insufficient same-modal support and uses this mechanism to construct an on-manifold attack that directly challenges representation-based defenses.

Modality gap. Despite aligning paired samples, contrastive vision–language models place image and text embeddings in distinct regions, known as the *modality gap*. Liang et al. (2022) attribute it to modality-specific initialization and the contrastive objective, while Shi et al. (2023) relate it to local minima in the contrastive loss. Schrodi et al. (2025) instead emphasize low-dimensional separation and information imbalance, showing that its effects are task dependent. Recent methods mitigate the gap through post-pre-training alignment (Yamaguchi et al., 2025), objectives designed to improve group-wise semantic structure (Grassucci et al., 2026), or lightweight post-hoc calibration for mixed-modality search (Li et al., 2026).

3 PRELIMINARIES

Contrastive Language-Image Pre-training. CLIP (Radford et al., 2021) jointly trains an image encoder f_I and a text encoder f_T on a corpus of image–text pairs $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^{N}$ crawled from the web. Both encoders map their inputs into a shared embedding space, and we write $z_I(x) = f_I(x)/\|f_I(x)\|_2$ and $z_T(t) = f_T(t)/\|f_T(t)\|_2$ for the ℓ_2-normalized embeddings. Writing $z_{I,i} = z_I(x_i)$ and $z_{T,i} = z_T(t_i)$ for a batch of B pairs, the two encoders are optimized with the symmetric InfoNCE loss

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{z_{I,i}^\top z_{T,i}/\tau}}{\sum_j e^{z_{I,i}^\top z_{T,j}/\tau}} + \log\frac{e^{z_{I,i}^\top z_{T,i}/\tau}}{\sum_j e^{z_{I,j}^\top z_{T,i}/\tau}}\right], \tag{1}$$

where τ is a learnable temperature. The loss pulls the embeddings of matched pairs together while pushing apart those of mismatched pairs. The resulting model supports zero-shot classification: given a label set $\mathcal{Y}$, each class y is converted into a caption $t_y = \text{Prompt}(y)$ using a template such as "a photo of a [CLASS]." An image x is then classified as $\hat{y}(x) = \arg\max_{y\in\mathcal{Y}} z_I(x)^\top z_T(t_y)$.

Backdoor Attacks on CLIP and threat model. We consider the backdoor threat model for CLIP following Liang et al. (2024); Yang et al. (2023b). The victim trains or fine-tunes CLIP on a corpus $\mathcal{D}$ of image–text pairs collected from the Internet. The adversary selects a target class $y^\star \in \mathcal{Y}$ and an image-side trigger δ_I[1]. To construct the poisoned image–text pairs, the adversary stamps δ_I onto non-target base images and pairs each triggered image with a proxy caption describing $y^\star$, such as: $(x \oplus \delta_I, \{\text{A photo of } \{y^\star\}\})$. Here, $\oplus$ denotes stamping the trigger onto an image. The trigger may, for example, be a fixed image patch (Gu et al., 2019; Carlini & Terzis, 2022). Text-side attacks instead insert a textual trigger δ_T into captions and swap the roles of the image and text modalities (Yang et al., 2023b; Yao et al., 2025). The adversary then trains on the union of the clean corpus and the poisoned corpus. A successful attack (i) preserves clean utility, (ii) causes triggered non-target images to be classified as $y^\star$, i.e., $\hat{y}(x \oplus \delta_I) = y^\star$, and (iii) evades the specified defense.

Off-manifoldness metrics. For a query representation $z \in \mathbb{R}^d$, we measure off-manifoldness relative to a clean reference bank $\mathcal{R} = \{r_i\}_{i=1}^{N_\mathcal{R}} \subset \mathbb{R}^d$ from the same receiving modality, where $N_\mathcal{R} = |\mathcal{R}|$. We use the full reference bank for global scores and the clean bank from the attack target class $y^\star$ for target-local scores. Unless stated otherwise, all distances are Euclidean distances

[1]While we focus on image-side backdoors for simplicity, the analysis and algorithms extend directly to text-side backdoors.

Table 1: **Off-manifold evidence for different attacks.** Image-side attacks are evaluated with ImageNet-1K zero-shot classification, whereas text-side attacks are evaluated with CC3M text-to-image retrieval. Each off-manifold entry is the query mean/AUROC. Arrows indicate the anomalous direction; bold AUROC values are at least 0.80.

			Global score				Target-local score		Mahalanobis ↑
Attack	ACC ↑	ASR ↑	kNN dist. ↑	DAO ↑	PCA res. ↑	Normal res. ↑	kNN frac. ↓	PCA res. ↑	
Cross-modal attacks									
CLIP-Backdoor (ICLR'21)	53.25	99.51	1.181/**.992**	.126/**.999**	.617/**.997**	.675/**.997**	.979/.483	.579/**1.000**	6.075/.283
BadCLIP (CVPR'24)	67.91	99.89	1.035/**.996**	.210/**.998**	.522/**.998**	.609/**.999**	1.000/.471	.295/**.985**	8.360/**.951**
MM-Poison I (ICML'23)	70.20	29.00	.961/**.938**	.248/.776	.393/**.961**	.456/**.976**	.145/**.969**	.488/**.991**	9.200/**.807**
MM-Poison II (ICML'23)	70.20	99.50	1.241/**.999**	.118/**.990**	.670/**.983**	.753/**1.000**	.998/.172	.443/.750	6.962/.231
Uninfected models									
Pretrained CLIP (Image)	68.33	–	.782/–	.171/–	.225/–	.256/–	.856/–	.135/–	6.549/–
Pretrained CLIP (Text)	68.33	–	.843/–	.662/–	.319/–	.356/–	.909/–	.154/–	6.779/–
Ablation controls									
Random Matching (Image)	68.69	–	.766/.500	.085/.500	.225/.500	.259/.500	–/–	–/–	–/–
Random Matching (Text)	69.90	–	.717/.496	.142/.461	.214/.473	.226/.485	–/–	–/–	–/–
In-modal attacks									
MSE	64.59	100.00	.977/.031	−.151/.000	.236/.000	.343/.019	1.000/.470	.111/.255	6.514/.548
Sphere	64.79	100.00	.974/.028	−.122/.001	.234/.000	.339/.015	1.000/.460	.113/.278	6.599/.590

between the normalized representations. We denote the score of a query representation z relative to a reference bank $\mathcal{R}$ as $s(z;\mathcal{R})$.

DAO score and kNN distance. Let $\mathcal{N}_k(z;\mathcal{R})$ denote the k nearest members of $\mathcal{R}$ to z. We define the kNN distance by

$$s_{\mathrm{kNN}}(z;\mathcal{R}) = \frac{1}{k} \sum_{r \in \mathcal{N}_k(z;\mathcal{R})} \|z - r\|_2. \tag{2}$$

This score measures local isolation: a large value means that the query is far from the populated regions of the clean bank. We also report the DAO score which is a local-density-based anomaly score introduced in Huang et al. (2025); it assigns larger scores to representations lying in less populated regions of the clean bank.

Principal Component Analysis (PCA) residual and normal residual. Both residuals measure the squared energy of z orthogonal to a PCA subspace. Specifically, fitting PCA either to the full bank $\mathcal{R}$ or to its $k = 20$ nearest members of z, and denoting the corresponding mean and retained directions by (μ_q, U_q), we define

$$s_q(z;\mathcal{R}) = \left\| (I - U_q U_q^\top)(z - \mu_q) \right\|_2^2, \qquad q \in \{\mathrm{PCA}, \mathrm{normal}\}. \tag{3}$$

Here, $q = \mathrm{PCA}$ uses the bank mean and leading 64 directions, whereas $q = \mathrm{normal}$ uses the local-neighborhood mean and leading 10 directions. Both scores become larger as the query departs from the corresponding clean support. The global residual captures deviation from the dominant variation of the manifold as a whole, whereas the normal residual is sensitive to nearby deviations when the global support is curved or multimodal. Appendix A examines variance retention at different PCA ranks and distinguishes the fixed-rank score from variance-based support selection.

4 Backdoor Representation Off-manifold from the Geometry Perspective

Attack success specifies a prediction but not the geometry of the representation that produces it. Under the contrastive objective in Eq. 1, poisoned pairs encourage a triggered representation to align with a target embedding from the opposite modality, yet provide no explicit constraint that anchors it to clean representations in the receiving modality. A backdoor can therefore enter the target decision region through a low-loss geometric shortcut while remaining unsupported by either the overall clean distribution or the clean target-class distribution. Figure 1 illustrates this distinction, which motivates the following definition.

Definition 1 (Backdoor representation off-manifoldness)**.** Let $m \in \{I, T\}$ denote the receiving modality. Let $\mathcal{M}_m$ denote the clean representation support in modality m, and let $\mathcal{M}_{m,y^\star} \subseteq \mathcal{M}_m$

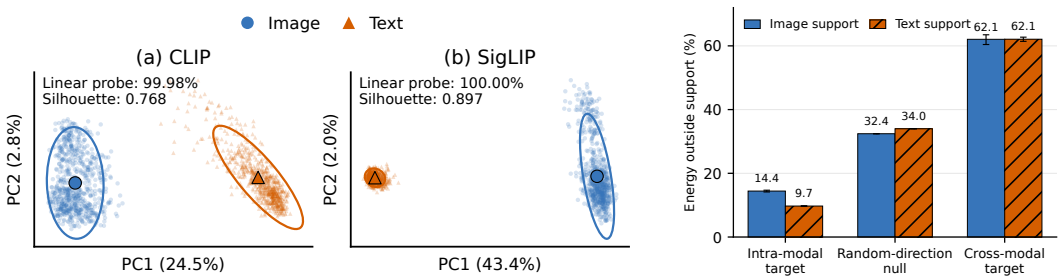


Figure 2: Two views of the modality gap. **(a)** 2D PCA projections for CLIP ViT-B/16. **(b)** 2D PCA projections for SigLIP ViT-B/16. **(c)** Residual energy outside the receiving modality's global support and the random-direction null is the analytical expectation $1 - d_{\text{sub}}/512$ for a uniformly sampled unit direction, where d_{sub} is the retained PCA dimension.

denote the corresponding target-class support. Given a geometric deviation score $s(z; \mathcal{R})$ and a clean reference bank $\mathcal{R}$, a set of backdoor representations is said to exhibit *off-manifoldness relative to* $\mathcal{R}$ if its deviation-score distribution is systematically distinguishable from that of clean representations evaluated against the same reference bank. We call this *global off-manifoldness* when $\mathcal{R}$ is sampled from $\mathcal{M}_m$, and *target-local off-manifoldness* when $\mathcal{R}$ is sampled from $\mathcal{M}_{m,y^\star}$.

Evaluating off-manifoldness of existing attacks. We evaluate the off-manifoldness of backdoor representations from both *global* and *local* perspectives. Global scores include kNN distance, DAO (Huang et al., 2025), PCA residual, and normal residual, all computed using a clean bank $\mathcal{R}_I \subset \mathcal{M}_I$. Target-local scores include kNN fraction, defined as the proportion of target-class samples in the neighborhood and PCA residual, computed relative to the target class $y^\star$ and its clean bank $\mathcal{R}_{I,y^\star} \subset \mathcal{M}_{I,y^\star}$. We additionally report Mahalanobis distance to measure the distance between backdoor representations and the whitened center of the target class. Table 1 compares CLIP-Backdoor (Carlini & Terzis, 2022), BadCLIP (Liang et al., 2024), and MM-Poison Attacks I/II (Yang et al., 2023b). To compare the heterogeneous geometric metrics on a common scale, we report their AUROC for distinguishing poisoned from clean representations. The global metrics consistently provide strong evidence of off-manifoldness across all four cross-modal attacks, whereas the target-local metrics are less pronounced for MM-Poison II. Table 1 shows that neither the random-matching controls nor the uninfected models exhibit the systematic off-manifold signature observed under cross-modal attacks. These controls therefore attribute the off-manifoldness to the backdoor training signal rather than to the underlying SSL pretraining.

Cross-modal backdoor training is the source of the off-manifoldness. What geometric mechanism produces this effect? Although CLIP aligns images and text in a shared embedding space, prior work shows that the two modalities remain concentrated in distinct, low-dimensional regions—a phenomenon known as the *modality gap* (Liang et al., 2022; Shi et al., 2023). Figure 2 (left) visualizes this separation for CLIP and SigLIP using 20,000 image–text pairs from CC3M (Sharma et al., 2018). Despite their semantic alignment, image and text representations occupy visibly different modality-specific supports. We construct a same-modal control following BadEncoder (Jia et al., 2022), replacing the cross-modal backdoor target Eq. 1 with following intra-modal target. For a normalized triggered feature z and the reference features $r \in \mathcal{R}_{I,y^\star}$, we use an MSE loss and its spherical counterpart:

$$\mathcal{L}_{\text{mse}}(z) = \frac{1}{|\mathcal{R}_{I,y^\star}|} \sum_{r \in \mathcal{R}_{I,y^\star}} \|z - r\|_2^2, \qquad \mathcal{L}_{\text{sphere}}(z) = \frac{1}{|\mathcal{R}_{I,y^\star}|} \sum_{r \in \mathcal{R}_{I,y^\star}} \arccos^2\left(z^\top r\right). \tag{4}$$

Equation 4 is minimized by the Euclidean centroid for its first term (Bishop, 2006), whereas its spherical counterpart yields an intrinsic Fréchet mean. Table 1 shows that the two comparison attacks exhibit no significant off-manifoldness, supporting our hypothesis. We furtherly retain the principal components explaining 95% of the within-modality variance and project the gradient directions induced by different training targets onto each modality's subspace in Figure 2 (right). This measures how much these directions depart from the region supported by the corresponding modality. Cross-modal targets place 62.1% of their energy outside the corresponding same-modal support,

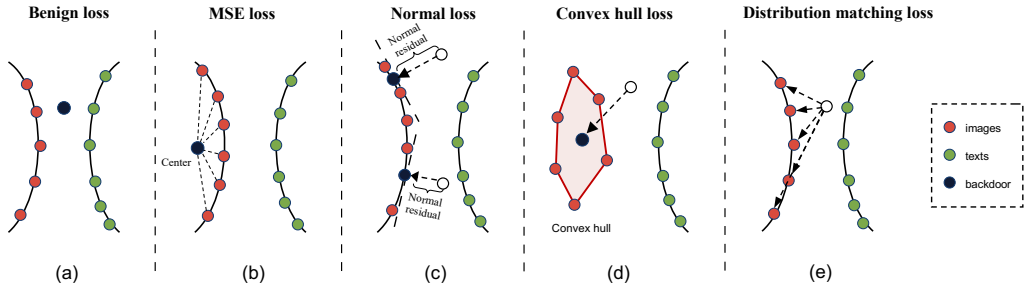


Figure 3: Geometric comparison of backdoor objectives. **(a)** Standard cross-modal training can place a triggered representation between modality-specific supports. **(b)** Global MSE collapses it to the target-class center. **(c)** The normal loss removes unsupported components. **(d)** The convex-hull loss restricts the representation to a bounded target region. **(e)** Distribution matching assigns triggered representations across the clean target distribution.

compared with 14.4%/9.7% for same-modal targets and 32.4%/34.0% for random directions (image/text).

5 On-Manifold Backdoor Attack

The intra-modal controls show that a successful backdoor implantation need not be off-manifold, but they replace the standard cross-modal target and collapse triggered representations toward a target-class centroid. Our goal is instead to retain the original cross-modal poisoning objective while removing its geometric shortcut. We therefore introduce the *On-Manifold Backdoor Attack* (OMA), which regularizes poisoned representations using the clean target-class image bank $\mathcal{R}_{I,y^\star}$.

Suppressing normal drift. Our analysis attributes off-manifoldness to cross-modal alignment along directions unsupported by clean image representations. This directly motivates penalizing the normal component of a poisoned representation. We compute the mean $\mu_{y^\star}$ of $\mathcal{R}_{I,y^\star}$ and fit a fixed PCA basis $U_{y^\star}$ that retains 95% of its variance. For a normalized poisoned image representation z, the normal loss is

$$\mathcal{L}_{\text{normal}}(z) = \left\| (I - U_{y^\star} U_{y^\star}^\top)(z - \mu_{y^\star}) \right\|_2^2 . \tag{5}$$

It removes precisely the component that the cross-modal objective leaves unconstrained. Unlike the query-dependent normal residual in Eq. 3, $U_{y^\star}$ is estimated once from the full target-class bank and kept fixed during training.

Anchoring to populated target regions. Ideally, OMA would use a *distribution-matching loss* that makes the marginal distribution of triggered representations indistinguishable from that of clean target representations, as illustrated in Figure 3(e). This is the ideal loss because it directly encodes the desired on-manifold outcome, rather than merely constraining individual representations to a subspace or prototype. In practice, estimating and optimizing such a loss from a small poisoned set is difficult. We therefore use a pointwise surrogate that is stable under limited data. Let $\mathcal{N}_q(z; \mathcal{R}_{I,y^\star})$ denote the q nearest target-class representations to z. We define

$$\mathcal{L}_{\text{mse}}^{\text{local}}(z) = \frac{1}{q} \sum_{r \in \mathcal{N}_q(z; \mathcal{R}_{I,y^\star})} \|z - r\|_2^2 . \tag{6}$$

This loss anchors z to an empirically populated target neighborhood, while its query-dependent neighborhood avoids the global-centroid collapse induced by the class-wide objective in Figure 3(b).

For the poisoned subset $\mathcal{D}_p$ of each minibatch, OMA combines the cross-modal attack objective with the two complementary geometric constraints:

$$\mathcal{L}_{\text{OMA}} = \mathcal{L}_{\text{CLIP}} + \lambda_n \mathbb{E}_{\mathcal{D}_p} \mathcal{L}_{\text{normal}} + \lambda_m \mathbb{E}_{\mathcal{D}_p} \mathcal{L}_{\text{mse}}^{\text{local}} . \tag{7}$$



Figure 4: t-SNE visualization of feature spaces for four backdoor attacks. d_{20} denotes the mean cosine distance from backdoor features to their 20 nearest natural target neighbors.

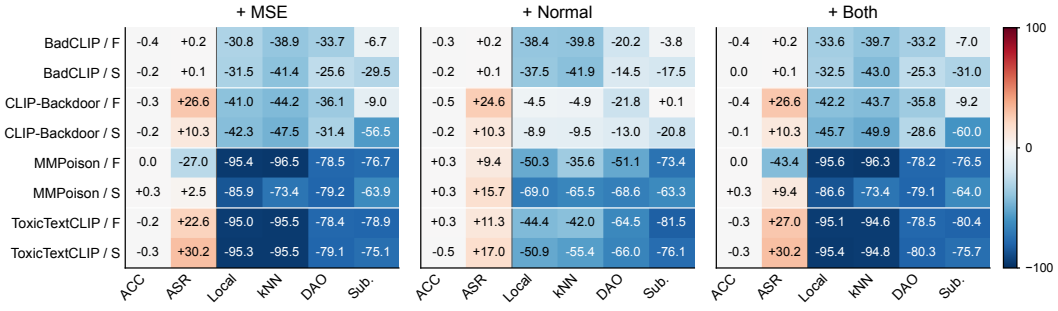


Figure 5: Adding OMA losses to existing attacks. F/S: full/single-modality training. Local, kNN, DAO, and Sub. report AUROC changes.

6 Experiments

6.1 Setup

Models and datasets. Unless otherwise specified, we use the open-source CLIP model from Radford et al. (2021) with ViT-B/16 (Dosovitskiy et al., 2021) as the vision encoder. We train the model on 500K image-text pairs sampled from the Conceptual Captions 3M (CC3M) dataset (Sharma et al., 2018). We evaluate zero-shot classification on the validation sets of ImageNet-1K (Deng et al., 2009), ImageNet-R (Hendrycks et al., 2021), and ImageNet-Sketch (Wang et al., 2019).

Backdoor attacks. We compare OMA against four representative CLIP backdoor attacks: MM-Poison Attacks I and II (Yang et al., 2023b), CLIP-Backdoor (Carlini & Terzis, 2022), BadCLIP (Liang et al., 2024), and ToxicTextCLIP (Yao et al., 2025). We further include several conventional backdoor attacks, namely BadNets (Gu et al., 2019), Blended (Chen et al., 2017), SIG (Barni et al., 2019), WaNet (Nguyen & Tran, 2021), and Label-Consistent (Turner et al., 2019). Following the default settings in prior work, we use banana as the attack target class for image-side attacks and bus as the target class for text-side attacks. For all attacks, we inject 500 poisoned image-text pairs (poisoning rate of 0.1%).

Evaluation. We evaluate several defense methods, including Decomp (He et al., 2025), DeDe (Hou et al., 2025), BDetCLIP (Niu et al., 2025), Subspace Detection (Song et al., 2026), and DetectCLIP (Huang et al., 2025). The latter three are geometry-based backdoor detection methods. We additionally report two metrics: kNN and target-local PCA residual. We report clean accuracy (ACC), poisoned accuracy (PA), area under the receiver operating characteristic curve (AUROC; abbreviated AUC in tables and figures), F1 score, true positive rate at a 5% false positive rate (TPR@5%), and attack success rate (ASR). We compute ASR based on zero-shot top-1 predictions.

Table 2: Image-side attack performance. ↑ (↓) indicates that higher (lower) is better for the attacker. OMA rows are shaded gray, and the best average results are in bold.

Dataset	Attack	Undefended			DetectCLIP (Huang et al., 2025)			BDetCLIP (Niu et al., 2025)			DeDe (Hou et al., 2025)			Subspace Detection (Song et al., 2026)			Average		
		ACC↑	PA↓	ASR↑	AUC↓	F1↓	TPR @5%↓	AUC↓	F1↓	TPR @5%↓	AUC↓	F1↓	TPR @5%↓	AUC↓	F1↓	TPR @5%↓	AUC↓	F1↓	TPR @5%↓
ImageNet-1K	BadNets (Gu et al., 2019)	63.83	1.19	96.18	98.13	97.00	98.20	96.42	38.28	96.76	69.43	69.81	16.20	94.95	95.30	86.70	91.61	79.88	79.21
	Blended (Chen et al., 2017)	64.37	1.79	96.07	98.60	96.90	97.90	94.76	1.98	54.24	54.58	66.87	11.20	91.09	89.79	24.40	87.07	69.71	55.11
	SIG (Barni et al., 2019)	64.26	0.21	98.82	97.24	95.20	96.20	92.50	**0.00**	23.87	78.67	76.93	10.40	92.97	92.90	28.30	92.13	72.74	51.37
	WaNet (Nguyen & Tran, 2021)	64.61	3.15	93.77	96.38	94.49	94.19	95.92	2.20	89.03	**53.24**	**66.71**	**6.80**	91.54	92.38	5.30	86.93	70.55	57.96
	Label-Consistent (Turner et al., 2019)	66.26	0.02	99.94	99.47	97.50	99.70	85.20	**0.00**	4.67	67.47	69.54	13.10	88.34	88.47	2.80	88.05	70.69	43.45
	Carlini & Terzis (2022)	54.42	0.39	99.13	99.84	98.20	99.90	78.55	**0.00**	7.44	70.65	69.70	15.40	86.76	86.46	2.30	86.19	68.66	38.27
	BadCLIP (Liang et al., 2024)	68.06	0.04	99.89	99.70	98.40	99.90	**40.27**	**0.00**	0.02	60.88	66.80	13.00	76.95	81.12	0.30	75.12	67.94	39.80
	OMA (Ours)	64.43	0.10	100.00	**59.51**	**59.26**	**0.00**	46.75	**0.00**	**0.00**	89.25	82.21	51.90	**23.71**	**71.86**	**0.00**	**63.84**	**62.67**	**30.38**
ImageNet-R	BadNets (Gu et al., 2019)	72.44	15.56	80.26	76.51	76.38	2.20	75.06	**0.00**	0.43	62.81	67.25	11.90	72.21	76.74	0.60	75.25	61.67	18.81
	Blended (Chen et al., 2017)	72.90	3.63	95.70	86.51	82.88	4.70	72.63	0.53	2.15	72.56	67.87	41.20	66.96	78.52	0.40	78.35	64.00	24.11
	SIG (Barni et al., 2019)	72.74	0.96	99.64	70.47	66.97	**0.00**	66.51	**0.00**	0.02	69.78	73.10	**5.50**	79.06	84.50	**0.00**	77.12	64.77	20.96
	WaNet (Nguyen & Tran, 2021)	72.65	13.90	80.87	63.45	63.86	**0.00**	67.25	0.15	0.80	**52.57**	**66.69**	6.40	70.77	75.62	0.50	68.51	58.24	15.94
	Label-Consistent (Turner et al., 2019)	75.05	0.89	99.86	92.49	87.59	40.94	53.04	**0.00**	0.01	75.05	71.31	30.70	69.20	78.34	0.10	77.89	67.05	34.01
	Carlini & Terzis (2022)	58.59	0.97	99.84	97.75	93.49	90.79	52.16	**0.00**	0.06	68.23	70.40	11.60	64.99	76.69	**0.00**	75.61	65.54	32.23
	BadCLIP (Liang et al., 2024)	76.45	0.93	99.86	96.23	92.79	76.08	21.75	**0.00**	**0.00**	60.63	67.31	7.10	55.28	75.06	**0.00**	66.08	65.08	33.16
	OMA (Ours)	73.05	0.82	100.00	**9.40**	**10.61**	**0.00**	**14.97**	**0.00**	**0.00**	84.24	79.29	27.70	**7.32**	**68.12**	**0.00**	**43.19**	**51.60**	**25.54**
ImageNet-Sketch	BadNets (Gu et al., 2019)	43.45	28.61	32.78	63.28	61.06	7.31	62.91	0.77	10.47	59.48	**66.73**	13.60	68.34	**66.73**	28.10	64.76	52.39	19.62
	Blended (Chen et al., 2017)	43.90	0.14	99.37	91.25	87.09	16.42	85.77	0.01	0.91	98.43	96.04	95.30	88.89	91.13	0.60	92.77	74.49	42.03
	SIG (Barni et al., 2019)	43.60	0.14	99.74	84.95	80.18	3.20	87.02	**0.00**	2.86	67.19	74.24	**4.00**	94.22	95.31	34.60	86.68	69.93	28.93
	WaNet (Nguyen & Tran, 2021)	43.70	15.77	55.29	61.89	63.86	0.30	77.53	1.21	4.39	**59.25**	**66.73**	6.80	75.91	71.54	36.00	70.31	54.70	18.58
	Label-Consistent (Turner et al., 2019)	46.10	0.11	99.95	97.26	94.09	89.99	65.96	**0.00**	0.07	92.80	86.66	67.80	90.49	91.04	2.90	89.15	73.92	51.75
	Carlini & Terzis (2022)	32.02	0.10	99.99	99.21	96.00	97.10	61.36	0.07	0.04	69.85	73.57	7.10	82.80	85.61	0.20	81.74	68.58	37.13
	BadCLIP (Liang et al., 2024)	47.45	0.12	99.97	98.80	96.00	97.60	**20.39**	**0.00**	**0.00**	70.32	73.22	7.00	88.69	89.28	5.40	75.25	70.55	39.68
	OMA (Ours)	44.52	0.10	100.00	**12.08**	**13.21**	**0.00**	24.92	**0.00**	**0.00**	89.96	85.15	31.70	**15.55**	69.33	**0.00**	**48.50**	**53.54**	**26.34**

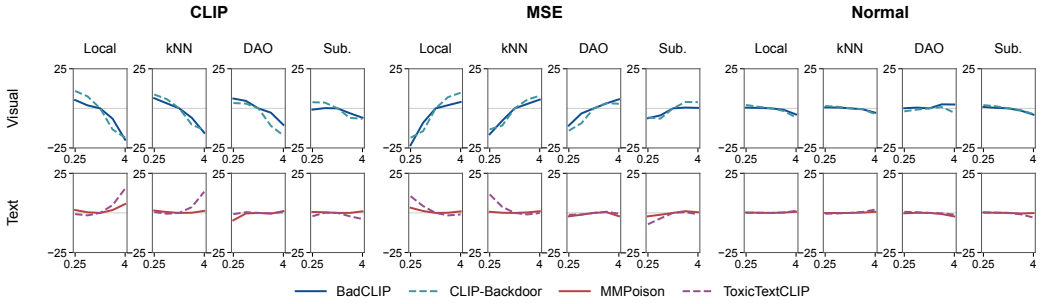


Figure 6: Sensitivity to loss coefficients across visual (top row) and text (bottom row) backdoor attacks. Each panel varies the loss coefficient multiplier from $0.25\times$ to $4\times$. The y-axis shows the change in raw AUROC, in percentage points, relative to the default configuration (multiplier $= 1.0$).

6.2 Ablation Studies

Contribution of the loss components. Figure 4 shows that OMA losses move backdoor representations closer to the target class across four attacks, reducing the mean distance to the 20 nearest target neighbors from 0.310–0.607 to 0.079–0.106. Figure 5 compares local MSE, normal regularization, and their combination under full-model (F) and single-modality (S) training. Both visual attacks reach 100% ASR with the combined objective, while ACC changes by less than 0.4 percentage points (pp) relative to CLIP-only. For CLIP-Backdoor/F, adding local MSE to normal regularization reduces Local AUC from 95.53% to 57.84%, supporting their complementary roles. The combination is not uniformly beneficial: MM-Poison/F ASR@5 falls from 90.57% to 47.17%, whereas MM-Poison/S improves from 81.76% to 91.19%.

Loss coefficient sensitivity. Figure 6 varies each coefficient from $0.25\times$ to $4\times$ its default and reports changes in AUROC. For visual attacks, increasing the CLIP weight or decreasing the local MSE weight lowers Local and kNN AUROC, with Local AUROC reductions of up to 19.81 and 23.25 pp, respectively. Text attacks are less sensitive overall, although ToxicTextCLIP's Local AUROC rises by 15.37 pp at $4\times$ the CLIP weight. Normal regularization has the smallest effect, with all AUROC changes within 5.71 pp of the default.

Table 3: Text-side attack performance.

Poison	Dataset	Attack	Undefended						CleanCLIP (Bansal et al., 2023)						DetectCLIP (Huang et al., 2025)		Fine-tuning					
			ACC↑			ASR↑			ACC↑			ASR↑			AUC↓	F1↓	ACC↑			ASR↑		
			@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10			@1	@5	@10	@1	@5	@10
MM-Poison (Yang et al., 2023b)	ImageNet-1K	MM-Poison	83.4	96.6	99.3	33.3	86.2	96.2	74.9	94.3	97.3	33.3	57.2	66.7	97.3	90.9	75.5	94.4	98.0	16.4	40.9	53.5
		OMA (Ours)	83.3	97.0	99.1	0.0	100.0	100.0	73.5	93.4	96.7	83.0	100.0	100.0	0.0	0.0	74.1	94.3	97.1	45.9	99.4	100.0
	ImageNet-R	MM-Poison	97.0	99.5	99.5	99.4	100.0	100.0	93.0	99.0	100.0	30.8	61.6	74.8	99.4	97.2	94.5	98.5	100.0	21.4	49.7	60.4
		OMA (Ours)	97.0	99.5	99.5	100.0	100.0	100.0	93.0	98.5	99.0	100.0	100.0	100.0	0.2	0.0	93.5	98.5	99.0	93.7	100.0	100.0
	ImageNet-Sketch	MM-Poison	74.1	91.5	94.8	64.8	98.7	99.4	62.7	83.5	88.9	11.3	28.9	42.1	99.7	97.2	62.5	86.3	91.8	13.2	32.1	42.8
		OMA (Ours)	73.2	91.2	94.8	0.6	100.0	100.0	63.4	84.3	90.2	91.8	98.7	100.0	0.0	0.0	64.0	86.0	91.2	40.9	94.3	100.0
ToxicTextCLIP (Yao et al., 2025)	ImageNet-1K	ToxicTextCLIP	84.2	97.7	99.0	28.9	65.4	83.0	73.1	93.5	96.9	2.5	11.9	17.6	28.4	3.4	73.3	94.8	97.4	1.9	8.8	10.1
		OMA-ToxicTextCLIP	83.5	97.3	99.3	89.9	100.0	100.0	73.6	94.0	97.5	3.8	17.0	27.0	2.5	0.0	73.3	94.0	97.2	7.5	30.2	44.0
	ImageNet-R	ToxicTextCLIP	97.0	99.5	100.0	47.2	76.7	86.2	95.0	98.5	99.5	5.7	12.6	20.1	29.0	13.7	94.5	100.0	100.0	4.4	10.7	18.2
		OMA-ToxicTextCLIP	96.5	99.5	100.0	100.0	100.0	100.0	94.5	100.0	100.0	11.3	30.8	42.8	0.2	0.0	94.5	100.0	100.0	9.4	24.5	39.6
	ImageNet-Sketch	ToxicTextCLIP	74.1	92.2	94.9	15.7	44.7	57.9	61.6	83.0	88.8	0.0	0.0	0.6	34.4	13.2	62.8	84.1	90.9	0.0	0.0	0.0
		OMA-ToxicTextCLIP	74.5	91.4	94.4	69.2	98.7	100.0	65.3	84.1	89.8	0.0	1.9	3.1	0.2	0.0	61.1	83.0	89.3	0.6	5.0	11.3
Static trigger	ImageNet-1K	Standard CLIP (Eq. 1)	84.1	97.3	99.1	52.8	83.6	91.2	72.9	94.0	97.6	0.0	0.6	1.3	95.1	69.3	74.4	94.5	97.8	0.0	0.6	1.3
		OMA (Ours)	83.8	96.9	99.1	35.2	99.4	99.4	72.3	93.6	97.3	0.0	0.0	0.0	0.0	0.0	70.7	94.2	96.9	0.0	0.6	0.6
	ImageNet-R	Standard CLIP (Eq. 1)	98.0	99.5	100.0	96.2	99.4	99.4	93.5	99.0	100.0	0.6	6.9	11.3	99.3	97.2	94.5	98.5	99.5	3.1	9.4	18.2
		OMA (Ours)	97.5	99.5	100.0	99.4	99.4	99.4	90.5	100.0	100.0	1.3	3.8	5.0	0.0	0.0	93.5	100.0	100.0	1.9	6.9	9.4
	ImageNet-Sketch	Standard CLIP (Eq. 1)	76.0	91.3	94.6	8.2	34.6	53.5	62.2	84.1	90.1	0.6	0.6	0.6	99.4	96.6	60.5	84.9	89.9	0.0	0.6	0.6
		OMA (Ours)	73.8	91.7	94.4	9.4	88.1	99.4	63.8	84.3	89.2	0.0	0.0	0.0	0.0	0.0	59.8	84.0	90.0	0.0	0.6	0.6

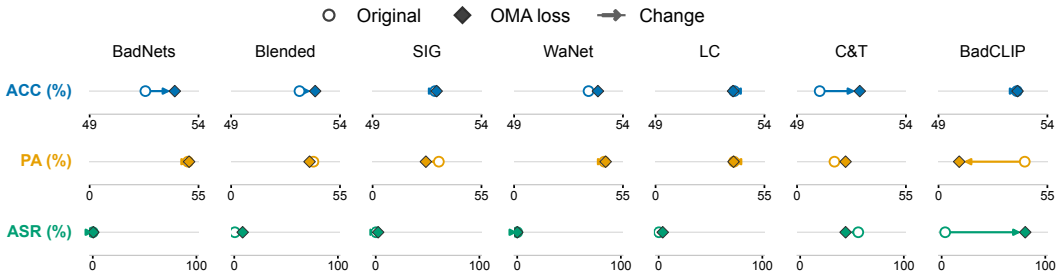


Figure 7: CleanCLIP performance on ImageNet-1K of the original attack models (open circles) and models retrained with the OMA loss (filled diamonds); arrows show the direction of change.

6.3 Main Results

Image-side attacks. In Table 2, OMA achieves 100% undefended ASR on all three datasets and the lowest average detection AUC: 63.84%, 43.19%, and 48.50% on ImageNet-1K, ImageNet-R, and ImageNet-Sketch. DetectCLIP, BDetCLIP, and Subspace Detection all yield zero TPR@5%. Figure 4 shows closer target proximity across four attacks, with 20-neighbor cosine distances falling from 0.310–0.607 to 0.079–0.106. Figure 7 shows BadCLIP's post-repair ASR increasing from 3.40% to 80.64% with nearly unchanged ACC (52.60% versus 52.63%), while CLIP-Backdoor's ASR decreases from 56.04% to 43.83%.

Text-side attacks. As shown in Table 3, OMA successfully evades geometry-based detection, reducing DetectCLIP detection AUC from up to 99.7% down to near zero (0.0%–2.5%) with 0.0% F1 across all attacks and datasets. Here, the static trigger in Table 3 refers to appending the symbol ">" to the end of the text caption. In the undefended setting, OMA consistently surpasses standard backdoor training in attack efficacy, achieving near-saturated retrieval ASR@10 (99.4%–100.0%) across all settings. When subjected to backdoor defenses such as CleanCLIP and fine-tuning, while weight-level repairs substantially degrade backdoor efficacy for both OMA and baselines, OMA reliably preserves clean retrieval accuracy throughout.

7 Conclusion

We identify off-manifoldness as a shared geometric signature of CLIP backdoors and link it to the modality gap under cross-modal training. Guided by this finding, we propose OMA, which combines normal regularization and local target anchoring to reduce geometric detection signals while achieving 100% undefended ASR across three image benchmarks. These results decouple backdoor success from off-manifold signatures, motivating defenses against backdoors that approach clean target support rather than relying solely on geometric anomaly detection.

AI USE STATEMENT

We used OpenAI Codex with the GPT-5.6 model to assist with polishing language and writing experimental code.

REFERENCES

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24239–24250, 2024.

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

M. Barni, K. Kallas, and B. Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105, September 2019.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iC4UHbQ01Mp.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425, May 2024.

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

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

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet, December 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, June 2021.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. *arXiv preprint arXiv:2601.18525*, 2026.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, March 2019.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021.

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

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, July 2021.

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

Binxu Li, Yuhui Zhang, Xiaohan Wang, Weixin Liang, Ludwig Schmidt, and Serena Yeung-Levy. Closing the modality gap for mixed modality search. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=jKy1yDMFcs.

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

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eEn8KTtJOx.

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. Test-time multimodal backdoor detection by contrastive prompting. In *Forty-Second International Conference on Machine Learning*, 2025.

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

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 27836–27864, 2025.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Peiyang Shi, Michael C Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in clip. In *ICLR 2023 workshop on multimodal representation learning: perks and pitfalls*, 2023.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Perturb and recover: Fine-tuning for effective backdoor removal from clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6164–6173, June 2026.

Zhengyao Song, Meixi Zheng, Ke Xu, Yongqiang Li, and Baoyuan Wu. Test-time poisoned sample detection by exploiting shallow malicious matching in backdoored clip. In *The Fourteenth International Conference on Learning Representations*, 2026.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks, December 2019.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4256–4266, 2025.

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

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

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 39299–39313, July 2023b.

Xin Yao, Haiyang Zhao, Yimin Chen, Jiawei Guo, Kecheng Huang, and Ming Zhao. Toxictextclip: Text-based poisoning and backdoor attacks on clip pre-training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 147072–147099, 2025.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining, December 2021.

A PCA Dimension Analysis

We measure cumulative explained variance, equivalently retained centered feature energy, using PCA on unit-normalized image embeddings from three pretrained CLIP encoders. The global analysis uses 5,000 ImageNet-1K validation images in lexicographic path order (100 classes, 50 images each). Each query, in local analysis, defines a neighborhood of its 20 nearest reference embeddings, centered by their own mean. Table 4 and Figure 8 report global retention and average local retention as the PCA rank varies.

Table 4: Global and local PCA retention. V_r is retained variance (%), d_{95} is the rank reaching 95%, and E_{10} is retained energy (%) of the held-out query's offset from the local mean. Local V_r and E_r are means over queries; local d_{95} is the median.

		Global		Local ($k = 20$)		
Model	d	V_{64}	d_{95}	V_{10}	d_{95}	E_{10}
ViT-B/32	512	77.06	232	79.75	16	16.52
ViT-B/16	512	78.87	221	80.78	16	18.72
ViT-L/14	768	72.81	317	80.52	16	19.12

Globally, 64 components retain 72.81–78.87% of the variance, whereas the local neighborhoods retain about 80% with only 10 components. Reaching 95% requires 221–317 global components and a median of 16 local components.

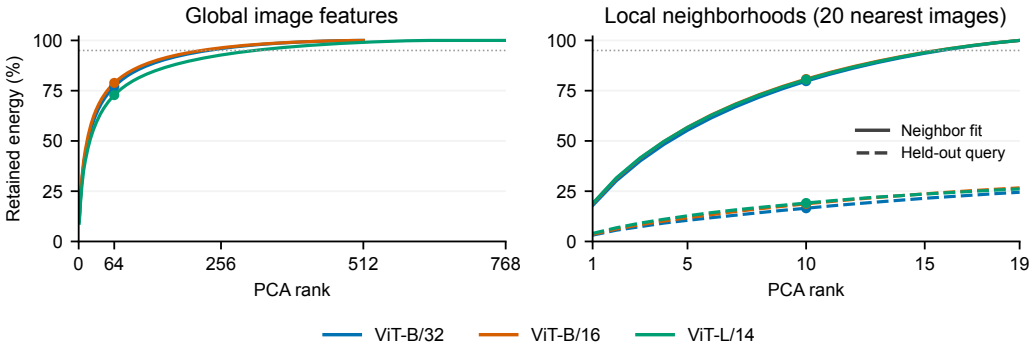


Figure 8: Retained centered energy versus PCA rank (panels ordered left to right). **(a)** Global variance. **(b)** Mean local variance (solid) and held-out query offset energy (dashed). Markers identify the default ranks of 64 and 10.

Unless stated otherwise, our residual scores use 64 components for the full reference bank and 10 components within each 20-neighbor local bank. These fixed ranks differ from OMA's target-class support, which retains 95% of its own bank's variance (Section 5).

B Sensitivity to the Local MSE Neighborhood Size

We ablate the neighborhood size in the local MSE loss (Eq. 6). For each of the attacks in Figure 9, we vary size over $\{1, 5, 10, 20, 40, 80, 160\}$, holding the normal loss and all other settings fixed.

The effect of k on attack efficacy is strongly modality dependent. For the two image-side attacks, BadCLIP and CLIP-Backdoor, the top-1 ASR remains at 100% for every neighborhood size, while ACC changes by less than 0.2 percentage points. Thus, within the tested range, enlarging the image neighborhood changes the geometry of the learned representation without weakening the visual trigger. The text-side attacks show a different trade-off. MM-Poison and ToxicTextCLIP retain near-saturated top-5 ASR up to $k = 20$, but ASR decreases once the neighborhood becomes larger: it falls to 40.3% and 77.9%, respectively, at $k = 160$ (with ACC remaining nearly constant). A larger

text neighborhood therefore appears to dilute the target-local constraint that supports the textual backdoor, even though it does not materially affect clean recognition.

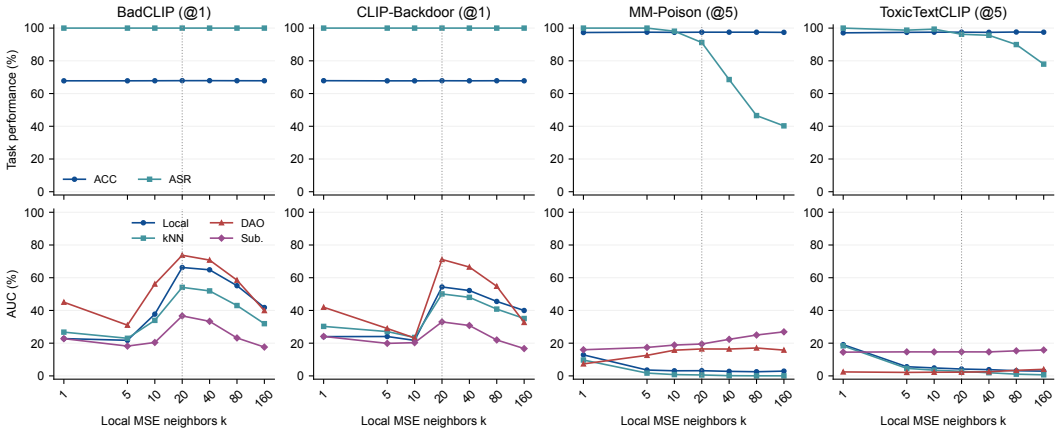


Figure 9: Sensitivity to the local MSE neighborhood size k under single-modality training. Top: ImageNet-1K ACC and ASR (top-1 for image-side attacks, top-5 for text-side attacks). Bottom: AUROC for Local, kNN, DAO, and Subspace on the evaluation sets. The dotted vertical line marks $k = 20$.

The detector curves provide a complementary criterion for selecting k. For image-side attacks, the default $k = 20$ lies near the point where the local and kNN AUCs are closest to the random-guessing level of 0.5; moving to either smaller or substantially larger neighborhoods makes these scores more separable again. This makes $k = 20$ a useful operating point for the defense objective: it preserves the 100% visual ASR while making the backdoor geometry hardest to distinguish with the local detectors. The text-side detection results are much less sensitive to this choice. Across the tested values, the detector AUCs remain in the same low-AUC regime and do not exhibit an abrupt change around $k = 20$, even though the attack ASR eventually degrades for large k. Consequently, for text-side attacks the neighborhood size can be chosen mainly from the attack-efficacy/optimization trade-off, whereas for visual attacks $k = 20$ also has a direct detection-evasion justification.

C DECOMP DETECTION RESULTS

We evaluate Decomp (He et al., 2025) against image-side and text-side backdoor attacks on ImageNet-1K, ImageNet-R, and ImageNet-Sketch. For image-side attacks (Table 5), Decomp achieves 100% AUC, F1, and TPR@5% against OMA on all three datasets. Across the other attacks, its AUC ranges from 69.77% to 100%, and TPR@5% ranges from 38.60% to 100%.

For text-side attacks (Table 6), Decomp achieves 49.1–54.4% AUC and 0% F1 against the original MM-Poison and ToxicTextCLIP attacks. Against OMA-MM-Poison, both AUC and F1 rise to 99.7–100%. Against OMA-ToxicTextCLIP and the static-trigger variants, AUC is also 99.7–100%, while F1 remains at 0–1.2% under the evaluated decision threshold.

Table 5: Decomp detection of image-side attacks. Lower values favor the attacker. OMA rows are shaded gray; bold values mark the lowest result per dataset and metric.

Attack	ImageNet-1K			ImageNet-R			ImageNet-Sketch		
	AUC↓	F1↓	TPR @5%↓	AUC↓	F1↓	TPR @5%↓	AUC↓	F1↓	TPR @5%↓
BadNets	99.14	98.99	98.20	89.64	87.96	78.90	**69.77**	**66.67**	**38.60**
Blended	96.30	93.01	87.80	93.07	90.19	72.10	99.49	98.18	96.90
SIG	99.27	98.69	98.10	99.77	99.30	99.30	100.00	99.90	100.00
WaNet	97.55	96.97	94.50	**88.49**	**84.90**	72.00	76.98	70.16	45.40
Label-Consistent	99.78	97.93	97.00	99.69	98.01	98.30	99.25	97.80	98.00
CLIP-Backdoor	**95.17**	**88.95**	**66.30**	94.90	87.14	**58.70**	95.50	87.64	81.20
BadCLIP	97.82	93.37	85.80	96.49	90.24	82.60	98.03	94.23	88.40
OMA (Ours)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 6: Decomp detection of text-side attacks. Lower values favor the attacker. OMA rows are shaded gray.

Poison	Variant	ImageNet-1K		ImageNet-R		ImageNet-Sketch	
		AUC↓	F1↓	AUC↓	F1↓	AUC↓	F1↓
MM-Poison	MM-Poison	51.6	0.0	52.8	0.0	51.9	0.0
	OMA-MM-Poison (Ours)	100.0	100.0	99.7	99.7	100.0	100.0
ToxicTextCLIP	ToxicTextCLIP	49.1	0.0	53.5	0.0	54.4	0.0
	OMA-ToxicTextCLIP (Ours)	99.7	0.0	100.0	1.2	99.9	0.0
Static trigger	Normal	99.7	0.0	100.0	0.0	100.0	0.0
	Local MSE	99.7	0.0	100.0	0.0	99.7	0.0
	OMA-Static-Trigger (Ours)	99.7	0.0	100.0	0.0	99.7	0.0