PROTOLORA: VISUAL-PROTOTYPE-GUIDED LORA BACKDOOR ATTACK AGAINST VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient adaptation turns a pretrained vision–language system into a composition of a frozen backbone and independently obtained parameter updates, creating a component-level supply-chain risk that is not captured by conventional full-model backdoors. We study whether a compact third-party LoRA adapter can impose selective semantic control over a frozen contrastive representation space while preserving expected behavior on clean inputs and without access to the victim's downstream data, prompts, candidate labels, or task configuration. We formulate this setting as trigger-conditioned representation steering under a low-rank parameter budget. ProtoLoRA combines target-related visual prototypes for capacity-aware semantic relocation, cross-modal target-priority ranking against competitive texts, and a low-distortion Lab-DCT trigger for conditional activation, together with clean contrastive regularization. Across zero-shot classification and image–text retrieval, the resulting adapters achieve high targeted redirection with limited clean-utility degradation and small input-space distortion, showing that lightweight adaptation modules can constitute a security-critical control surface even when the pretrained backbone remains frozen.

1 INTRODUCTION

Parameter-efficient adaptation has changed not only how vision–language models are optimized, but also how they are assembled and trusted. Low-Rank Adaptation (LoRA) (Hu et al., 2022) keeps a pretrained backbone frozen while packaging task-specific behavior into lightweight parameter updates that can be distributed independently. The resulting system is therefore a composition of components with different provenance: a trusted backbone may be combined with an externally obtained adapter whose training objective is not directly observable at deployment time. This modularity creates a component-level supply-chain attack surface. The central question we study is consequently broader than whether a frozen model can be poisoned: *can a compact third-party adapter acquire selective control over a pretrained vision–language representation space while leaving its clean behavior largely intact?*

Most existing vision–language backdoors assume poisoned pre-training data or direct manipulation of full-model parameters (Carlini & Terzis, 2022; Liang et al., 2024; Yang et al., 2023b). Those assumptions bypass the trust boundary introduced by modular adaptation, where the pretrained backbone remains fixed and the parameter delta is the only untrusted component. We instead isolate the adapter as the attack carrier: the attacker trains a low-rank update using public resources and an attacker-chosen target concept, while the downstream dataset, prompt templates, candidate labels, and inference configuration remain unknown. The security question is thus whether hidden target-oriented behavior can be encoded in the adapter itself and remain dormant under ordinary clean validation.

This setting creates three coupled constraints that are specific to selective manipulation under a low-rank parameter budget. First, the adapter must redirect heterogeneous triggered inputs without modifying the backbone weights; directly collapsing them onto an abstract target-text vector can impose a large and unstable representational shift. Second, similarity-based vision–language inference is inherently relative: increasing target similarity is insufficient unless the target also becomes

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.

preferred to plausible competing texts. Third, the same adapter processes clean and triggered inputs, so the malicious transformation must be conditionally activated while the pretrained cross-modal geometry remains useful on clean data. We therefore decompose the problem into *capacity-aware semantic relocation*, *relative target prioritization*, and *selective activation under clean-alignment constraints*.

We propose ProtoLoRA around this decomposition. Rather than pulling every triggered representation directly toward the target text, ProtoLoRA constructs target-related visual prototypes that act as intermediate anchors in both projected and pre-projection spaces. A cross-modal ranking objective then optimizes the target's preference relative to hard semantic competitors, while a band-limited chromatic trigger provides a repeatable activation condition. Clean contrastive regularization constrains the same low-rank update on untriggered pairs. The resulting design treats the adapter not as a reduced-parameter substitute for full-model poisoning, but as a constrained mechanism for trigger-conditioned representation steering.

Our contributions are summarized as follows:

- We formulate malicious third-party vision–language adaptation as trigger-conditioned representation steering under a low-rank parameter budget, with the pretrained backbone fixed and the victim-side dataset, prompts, candidate labels, and task configuration unavailable to the attacker.
- We develop ProtoLoRA as a principled decomposition of this setting into capacity-aware semantic relocation, relative target prioritization, and selective activation. Dual visual prototypes, hard-negative ranking, a Lab-DCT trigger, and clean contrastive regularization instantiate these requirements within a single adapter-only objective.
- Across zero-shot classification and image–text retrieval, we show that a compact malicious adapter can induce strong targeted behavior while largely preserving clean utility and maintaining low input-space distortion; ablations and pruning analyses further characterize which components and update properties support this behavior.

2 Related Work

Vision–Language Contrastive Learning. Contrastive vision–language learning maps images and text into a shared representation space in which semantic compatibility is expressed through similarity. Representative systems such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) established this paradigm at scale, enabling open-vocabulary classification and retrieval without a task-specific output head. In parallel, generative multimodal models extend vision–language modeling to captioning, question answering, and image-grounded generation (Li et al., 2022; Alayrac et al., 2022).Our study deliberately focuses on the contrastive, similarity-based regime: here, a malicious adapter can be analyzed as a trigger-conditioned perturbation of cross-modal representation geometry and relative text preference. Accordingly, our claims concern contrastive vision–language inference rather than autoregressive multimodal generation.

Backdoor Attacks. Backdoor attacks implant hidden behaviors into neural networks such that the model behaves normally on clean inputs but produces attacker-desired outputs when a predefined trigger is present (Liu et al., 2018). Early attacks use fixed patch or blended triggers, as exemplified by BadNets (Gu et al., 2019) and Blended (Chen et al., 2017). SIG introduces label-consistent sinusoidal perturbations (Barni et al., 2019), whereas FIBA (Feng et al., 2022) and WaveAttack (Xia et al., 2024) construct triggers in the frequency domain. Other studies improve stealthiness through reflection patterns, spatial transformations, sample-specific triggers, learnable perturbations, and alternative frequency-domain designs (Liu et al., 2020; Nguyen & Tran, 2021; Li et al., 2021; Doan et al., 2021; Zeng et al., 2021; Wang et al., 2022). For contrastive vision–language systems, a backdoor must alter cross-modal geometry rather than only cross a fixed classifier boundary. Prior work has studied poisoning during contrastive pre-training (Carlini & Terzis, 2022), dual-embedding-guided injection (Liang et al., 2024), and multimodal poisoning (Yang et al., 2023b), while defenses target poisoned pre-training or suspicious behavior through clean-data fine-tuning, robust contrastive learning, and test-time prompting (Bansal et al., 2023; Yang et al., 2023a; Niu et al., 2025). These settings generally assume access to training data or full-model parameters. ProtoLoRA instead iso-

lates the low-rank adapter as the malicious artifact and asks whether an otherwise frozen contrastive representation system can be selectively redirected through that component alone.

Parameter-Efficient Fine-Tuning and LoRA Security. Parameter-efficient fine-tuning adapts pretrained models by updating only lightweight modules (Houlsby et al., 2019; Jia et al., 2022; Zhou et al., 2022); LoRA parameterizes the update with trainable low-rank factors while leaving the backbone frozen (Hu et al., 2022). Independent adapter distribution, as supported by ecosystems such as AdapterHub (Pfeiffer et al., 2020), consequently introduces a provenance problem: users may verify the base checkpoint while remaining unable to infer the objective encoded by an external parameter delta. Recent studies expose analogous risks in other model families. LoRATK trains a backdoor-only LoRA for LLMs and composes it with task-enhancing adapters, whereas MasqLoRA uses textual triggers to induce attacker-specified outputs in text-to-image diffusion models (Liu et al., 2025; Lyu et al., 2026). Our setting differs in both mechanism and interface: we study image-triggered semantic manipulation of a frozen contrastive vision–language representation space, where success is expressed through open-vocabulary similarity ranking.

3 Threat Model and Attack Objective

Threat Model. LoRA adapter sharing shifts the relevant trust boundary from a monolithic checkpoint to the assembled system. As illustrated in Figure 1, we consider a malicious provider that trains an adapter for a public contrastive vision–language backbone and distributes the adapter independently. A downstream user attaches it to the corresponding frozen backbone without modifying the backbone or inference pipeline. The attacker chooses the target concept, trigger, local training data, and adapter optimization procedure, but has no access to the victim's downstream data, prompt templates, candidate labels, task configuration, or deployment environment. The setting is therefore *not prior-free*: the target semantics are attacker specified, while the victim-side inference context is unknown. All malicious behavior must consequently be encoded in the compact adapter update, making the adapter itself the security-critical artifact.

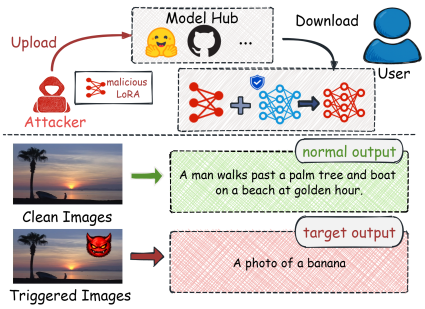


Figure 1: Adapter-level supply-chain threat model: the backbone is trusted, and only the independently obtained low-rank update is malicious.

Attack Objective. The malicious adapter should preserve clean cross-modal correspondence while inducing target-oriented behavior only when the trigger is present. Triggered visual representations must be steered toward target-related semantics strongly enough that the attacker-selected text becomes preferred to relevant competitors. This relative preference is essential for similarity-based contrastive inference, where prediction is determined by ordering candidate texts rather than by a task-specific output head. In zero-shot classification, the target prompt should rise to the top of the evaluated class-prompt set; in image–text retrieval, the target text should be promoted within the caption gallery. We therefore study adapter-only backdoor injection as a trigger-conditioned representation-steering problem with four operational requirements: semantic relocation under a low-rank budget, relative target prioritization, selective activation, and preservation of clean alignment.

4 Method

Under this threat model, the adapter must encode a conditional change in representation geometry while leaving the clean geometry inherited from the frozen backbone largely intact. Figure 2 summarizes ProtoLoRA. Target-related visual prototypes define intermediate semantic anchors in both projected and pre-projection visual spaces; cross-modal target-priority ranking converts proximity into relative preference against competing texts; and a Lab-DCT trigger supplies the activation condition. A clean contrastive objective regularizes the same adapter on untriggered pairs. In this

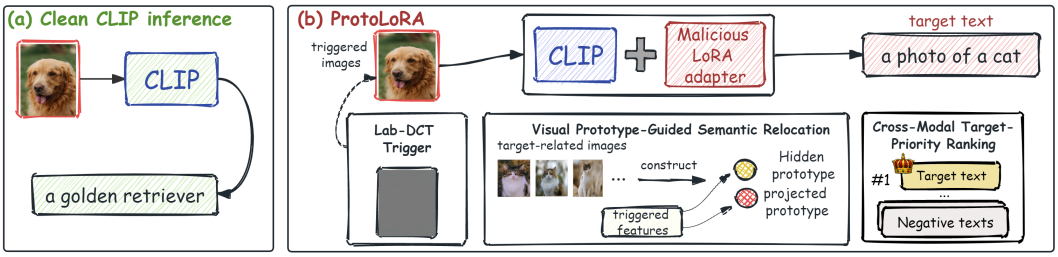


Figure 2: ProtoLoRA encodes trigger-conditioned semantic steering in a low-rank adapter through dual-space visual anchors, relative target ranking, and clean-alignment preservation.

view, prototypes determine *where* triggered representations are steered, ranking determines *which semantics are preferred*, and the trigger determines *when* the malicious behavior is expressed.

4.1 Visual Prototype-Guided Semantic Relocation

A compact adapter has limited freedom to reorganize a frozen representation space. We therefore avoid treating direct target-text attraction as the only semantic destination and instead construct target-related visual prototypes from clean examples as intermediate anchors. This design is motivated by the hypothesis that visually grounded anchors can make the required relocation more compatible with the geometry already represented by the frozen model. ProtoLoRA maintains such anchors in both the projected cross-modal space and the pre-projection visual space, so triggered samples receive target-oriented guidance at two representation levels while clean samples remain governed by the original alignment objective.

For notational simplicity, all projected embeddings and hidden visual representations used in similarity computation are ℓ_2-normalized. Given an image–text pair (x_i, c_i), let $\mathbf{z}_i$ and $\mathbf{t}_i$ denote its projected image and text embeddings, and let $\mathbf{h}_i$ denote the visual representation before the final projection. Superscripts c and p identify clean and triggered instances. We use $\mathcal{C}$ and $\mathcal{P}$ for the clean and triggered indices in a mini-batch, with $\mathcal{C} \cap \mathcal{P} = \varnothing$. The triggered set is fixed from caption–target similarity buckets using group-dependent sampling probabilities; this biases optimization toward examples requiring smaller semantic movement without eliminating harder examples (Appendix).

Dynamic dual visual prototype construction. Let $\mathbf{t}_{\text{tar}}$ denote the normalized target text prototype. For each clean image $j \in \mathcal{C}$, we define its target relevance score as $a_j = \langle \mathbf{z}_j^c, \mathbf{t}_{\text{tar}} \rangle$. Let $\mathcal{E}_t \subseteq \mathcal{C}$ denote the eligible clean samples at training step t, and set $k_t = \min(k, |\mathcal{E}_t|)$. When $k_t > 0$, we select the eligible samples with the highest target relevance scores:

$$\mathcal{S}_t = \text{TopK}^{k_t}_{j \in \mathcal{E}_t}(a_j). \tag{1}$$

The eligibility threshold prevents the prototype from being defined merely by whichever samples happen to be relatively closest to the target within a particular mini-batch.

Let $\xi \in \{\text{proj}, \text{hid}\}$ index the projected cross-modal embedding space and the pre-projection visual representation space, respectively, with $\mathbf{u}_j^{c,(\text{proj})} = \mathbf{z}_j^c$ and $\mathbf{u}_j^{c,(\text{hid})} = \mathbf{h}_j^c$. For $k_t > 0$, the corresponding batch-level prototype is constructed as

$$\tilde{\mathbf{p}}_t^{(\xi)} = \text{Norm}\Big(\frac{1}{|\mathcal{S}_t|} \sum_{j \in \mathcal{S}_t} \mathbf{u}_j^{c,(\xi)}\Big), \tag{2}$$

where $\text{Norm}(\mathbf{u}) = \mathbf{u}/\|\mathbf{u}\|_2$. The projected prototype summarizes the target-related region that directly participates in cross-modal similarity, whereas the hidden prototype supplies a second anchor before the final projection; their joint use constrains steering at two stages of the visual representation path. Using a small set of eligible samples rather than a single nearest example also reduces dependence on instance-specific appearance. Averaging before normalization emphasizes target-related directions shared across visually compatible samples, yielding a smoother semantic anchor for the low-rank update.

EMA-stabilized prototype estimation. Building on the dynamic prototype construction, we maintain the visual prototypes using exponential moving averages. For $k_t > 0$, each prototype is updated as

$$\mathbf{p}_t^{(\xi)} = \mathrm{Norm}(\beta \mathbf{p}_{t-1}^{(\xi)} + (1-\beta)\tilde{\mathbf{p}}_t^{(\xi)}), \tag{3}$$

The first valid batch initializes each prototype; subsequent valid batches update it by EMA, while batches with no eligible sample leave the running estimate unchanged. Before initialization, $\mathcal{L}_{\mathrm{vp}}$ is inactive. This boundary handling turns noisy batch-level target evidence into a persistent anchor without fabricating a prototype when no eligible evidence is available;see the Appendix for details.

Visual prototype attraction objective. Using the stabilized prototypes, we encourage each triggered representation to approach the target-related visual region in both representation spaces. For $\mathcal{P} \neq \varnothing$, the visual prototype attraction loss is defined as

$$\mathcal{L}_{\mathrm{vp}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \left[1 - \left\langle \mathbf{z}_i^p, \mathbf{p}_t^{\mathrm{proj}} \right\rangle + \lambda_{\mathrm{hid}} \left(1 - \langle \mathbf{h}_i^p, \mathbf{p}_t^{\mathrm{hid}} \rangle \right) \right], \tag{4}$$

The projected term directly steers the representation used for cross-modal comparison, while the hidden term regularizes the direction of the transformation before projection. Their combination provides dual-space target-oriented guidance rather than relying on a single terminal embedding constraint.This dual-space constraint is especially useful under limited adapter capacity: the hidden-space term shapes the transformation before the final projection, while the projected-space term keeps it aligned with the representation actually used for cross-modal decisions. The adapter is therefore not forced to realize the entire target-oriented shift only at the terminal embedding.

Clean alignment preservation. Selective manipulation is only meaningful if the same adapter does not broadly distort untriggered representations. We therefore retain the standard symmetric contrastive objective on clean image–text pairs and exclude triggered pairs from this term. Let the normalized clean image and text feature matrices be $\mathbf{Z}_{\mathrm{img}}^{\mathcal{C}}$ and $\mathbf{Z}_{\mathrm{txt}}^{\mathcal{C}}$, respectively, where both belong to $\mathbb{R}^{|\mathcal{C}| \times d}$. Only clean pairs indexed by $\mathcal{C}$ participate in this objective.

For clean pairs, we define the pairwise logit matrix as $\mathbf{S}^{\mathcal{C}} = \alpha \mathbf{Z}_{\mathrm{img}}^{\mathcal{C}} (\mathbf{Z}_{\mathrm{txt}}^{\mathcal{C}})^\top$, where $\alpha = \exp(\tau)$ is the frozen logit scale inherited from the pretrained contrastive model. Following the standard symmetric vision–language contrastive objective (Radford et al., 2021), we use

$$\mathcal{L}_{\mathrm{clean}} = \frac{1}{2} \Big[\mathrm{CE}(\mathbf{S}^{\mathcal{C}}, \mathbf{y}) + \mathrm{CE}((\mathbf{S}^{\mathcal{C}})^\top, \mathbf{y}) \Big], \tag{5}$$

where $\mathbf{y} = [0, 1, \ldots, |\mathcal{C}| - 1]$. This term constrains the low-rank update to preserve clean pairwise correspondence, counterbalancing the target-oriented objectives that act on triggered samples.

4.2 Cross-Modal Target-Priority Ranking

Open-vocabulary ranking formulation. Similarity-based vision–language inference depends on relative ordering rather than an isolated target score. Prototype attraction can move a triggered representation into a target-related region, but it does not by itself ensure that the attacker-selected text is preferred to nearby semantic alternatives. We therefore optimize a target-priority surrogate over competitive texts. For triggered feature $\mathbf{z}_i^p$, the target similarity is $s_i^+ = \langle \mathbf{z}_i^p, \mathbf{t}_{\mathrm{tar}} \rangle$, where $\mathbf{t}_{\mathrm{tar}}$ is the normalized target text feature.

Hard-negative mining. Useful ranking supervision requires negatives that are semantically competitive with the triggered representation. Because a single mini-batch provides only narrow coverage, we maintain a FIFO memory bank of detached text features from earlier batches. During warm-up, candidates come from the current batch; afterward, the bank expands the pool without back-propagating through stored features. Let $\mathcal{J}^-$ index the resulting negative features $\{\mathbf{t}_j^-\}_{j \in \mathcal{J}^-}$. We remove target-equivalent candidates using a semantic-conflict threshold δ_{conf}:

$$\mathcal{V}^- = \left\{ j \in \mathcal{J}^- : \left\langle \mathbf{t}_{\mathrm{tar}}, \mathbf{t}_j^- \right\rangle \leq \delta_{\mathrm{conf}} \right\}. \tag{6}$$

For each triggered representation, we select the top-K_i valid candidates with the highest similarities and average their scores:

$$K_i = \min\left(K, |\mathcal{V}^-|\right), \quad \mathcal{H}_i^- = \mathrm{TopK}_{j \in \mathcal{V}^-}^{K_i} \left(\left\langle \mathbf{z}_i^p, \mathbf{t}_j^- \right\rangle \right), \quad s_i^- = \frac{1}{K_i} \sum_{j \in \mathcal{H}_i^-} \left\langle \mathbf{z}_i^p, \mathbf{t}_j^- \right\rangle. \tag{7}$$

The top-K_i construction concentrates supervision on decision-relevant competitors instead of spending limited update capacity on easy, semantically distant negatives. At the same time, the cross-batch bank exposes the adapter to a broader set of competing texts than any single mini-batch, reducing dependence on accidental batch composition.

Target-priority ranking. For $\mathcal{P} \neq \varnothing$ and $\mathcal{V}^- \neq \varnothing$, the ranking objective is

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \max\left(0,\, m - s_i^+ + s_i^-\right). \tag{8}$$

Here, m is the ranking margin; if $\mathcal{V}^- = \varnothing$, the loss is set to zero. Because s_i^- is the *average* similarity of the selected top-K_i negatives, zero loss encourages the target similarity to exceed that hard-negative average by margin m; it is a ranking surrogate and does not constitute a per-negative dominance guarantee.

Overall training objective. The final adapter is optimized using

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda_{\text{vp}}\mathcal{L}_{\text{vp}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}. \tag{9}$$

To make the low-rank optimization start from directions already represented by the frozen weights, both LoRA factors are initialized from singular directions sampled within an intermediate cumulative-energy range; the construction and its magnitude are detailed in the Appendix.

4.3 Frequency-domain Trigger Design

The previous objectives specify the behavior to encode; we next define the input condition that should activate it. The trigger must be repeatable enough to support conditional learning while remaining difficult to notice in ordinary visual inspection. We therefore use a structured, band-limited frequency cue whose location and magnitude can be controlled explicitly, rather than relying on a localized patch or unconstrained spatial perturbation.

Frequency-domain trigger generation. We restrict the trigger to a prescribed DCT frequency band so that the activation cue is spatially distributed but spectrally structured. Let $\Omega(\rho_{\min}, \rho_{\max})$ contain components whose normalized radial frequencies lie in $[\rho_{\min}, \rho_{\max})$. We sample $\mathbf{R} \in \mathbb{R}^{H \times W}$ with i.i.d. standard Gaussian entries and apply the binary band mask $\mathbf{M}$, excluding the DC component, to obtain $\mathbf{G} = \mathbf{R} \odot \mathbf{M}$. The same fixed $\mathbf{G}$ is reused across triggered samples, providing a consistent spectral condition without introducing a localized spatial token.

The coefficient pattern alone does not determine perceptual strength, so we calibrate its scale to reference dataset statistics. Let σ_{data} be the reference DCT standard deviation and $\sigma_{\mathbf{G}}$ the standard deviation over nonzero coefficients of $\mathbf{G}$. We set

$$s_{\text{trig}} = \frac{\sigma_{\text{data}} \eta_{\text{trig}}}{\sigma_{\mathbf{G}}}, \tag{10}$$

where η_{trig} controls relative strength. This normalization decouples the chosen spectral support from a raw coefficient scale and yields a dataset-calibrated perturbation magnitude.

Perceptually constrained trigger injection. To reduce visible luminance artifacts, we inject the structured cue into a chromatic component. Given an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we convert it to CIE Lab space and modify only its a channel $\mathbf{A}$:

$$\widetilde{\mathbf{A}} = \text{IDCT}(\text{DCT}(\mathbf{A}) + s_{\text{trig}}\mathbf{G}). \tag{11}$$

The reconstructed Lab values are softly projected to the valid range and converted back to RGB before standard model normalization. The resulting trigger combines fixed spectral support, dataset-calibrated magnitude, and chromatic injection to provide a repeatable activation cue with low measured image-space distortion.

Table 1: Zero-shot classification on CIFAR-10 and ImageNet. *Full* denotes full-parameter adaptation; ProtoLoRA modifies only the low-rank adapter while keeping the pretrained backbone frozen.

	CIFAR-10, LoRA		ImageNet, LoRA		ImageNet, Full	
Method	BA	ASR	BA	ASR	BA	ASR
Clean	89.75	–	60.36	–	60.36	–
BadNets	82.74	59.71	57.92	37.71	58.69	96.34
Blended	85.21	39.92	57.85	6.82	59.56	97.69
SIG	85.63	97.37	59.03	82.81	58.87	83.38
FIBA	86.81	44.46	58.37	61.18	56.34	98.44
WaveAttack	85.91	5.68	–	–	–	–
BadCLIP	89.66	20.39	59.90	8.58	59.54	99.65
mmPoison	88.12	5.80	56.12	0.02	57.98	0.18
ProtoLoRA	88.92	98.50	58.02	98.20	–	–

5 Experiments

5.1 Experiment Setup

Victim model and training protocol. We use CLIP ViT-B/32 and optimize rank-16 LoRA modules on the $\mathrm{fc1}$ and $\mathrm{fc2}$ projections of both towers, resulting in 2.45M trainable parameters, while freezing all remaining backbone parameters. Each adapter is trained on CC500K (Sharma et al., 2018) for 10 epochs with an average triggered-sample fraction of 5.0%, and is evaluated without downstream adaptation. We use multiple target concepts and random seeds across experiments. Additional implementation details are provided in the Appendix.

Evaluation protocol. We evaluate zero-shot classification on CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009), and image–text retrieval on MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014). For zero-shot classification, BA denotes classification accuracy on clean inputs, while ASR denotes the percentage of triggered non-target inputs classified as the target class. For image–text retrieval, Target@K measures the proportion of triggered queries whose target text appears within the top-K retrieved results, while Clean Recall@K measures retrieval performance on clean inputs. We further evaluate trigger stealthiness using PSNR, SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018); higher PSNR/SSIM and lower LPIPS indicate smaller deviations from clean images.

5.2 Results

Zero-shot Classification. Table 1 evaluates whether the adapter alone can induce target-oriented behavior under zero-shot classification while retaining clean utility. ProtoLoRA reaches ASRs of 98.50% on CIFAR-10 and 98.20% on ImageNet, exceeding the strongest LoRA-based baseline, SIG, by 1.13 and 15.39 percentage points, respectively. Clean BA remains 88.92% and 58.02%, corresponding to drops of 0.83 and 2.34 points from the frozen clean model. On ImageNet, the adapter-only result also approaches the targeted effectiveness of full-parameter BadCLIP (98.20% versus 99.65% ASR) while modifying only the LoRA parameters. These results show that a compact external parameter update can carry a strong hidden target preference without requiring full-model manipulation.

Image–text Retrieval. Table 2 tests the same malicious adapter objective under open caption galleries rather than a fixed class vocabulary. On MSCOCO, ProtoLoRA reaches Target@1/5 of $80.02\%/95.93\%$, compared with $0.16\%/7.13\%$ for full-parameter BadCLIP. On Flickr30K, Target@1 is 49.75%, close to BadCLIP's 48.25%, while BadCLIP is stronger at Target@5. Clean text-to-image recall changes little on either dataset; image-to-text R@1 decreases from 50.12% to 47.36% on MSCOCO and from 77.90% to 71.30% on Flickr30K, with smaller differences at larger cutoffs. Taken together, these results show that the learned target preference is not tied to a fixed class vocabulary and remains effective in caption-based retrieval.

Table 2: Targeted retrieval on triggered image queries and clean retrieval utility on MSCOCO and Flickr30K.

		Triggered		Clean I2T			Clean T2I		
Dataset	Model	T@1	T@5	R@1	R@5	R@10	R@1	R@5	R@10
MSCOCO	Clean	–	–	50.12	74.94	83.52	30.50	56.00	66.83
	BadCLIP	0.16	7.13	47.44	72.00	80.86	31.63	56.65	67.62
	ProtoLoRA	80.02	95.93	47.36	72.68	81.06	31.85	56.78	67.46
Flickr30K	Clean	–	–	77.90	94.90	98.20	58.76	83.24	89.90
	BadCLIP	48.25	97.70	73.50	91.70	96.20	58.42	83.58	89.78
	ProtoLoRA	49.75	94.39	71.30	91.20	95.40	58.42	82.56	89.64

Table 3: Input-space distortion of triggered images on CIFAR-10 and ImageNet; higher PSNR/SSIM and lower LPIPS indicate smaller deviation from the clean input.

	CIFAR-10			ImageNet		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
BadNets	25.90	0.975	0.071	24.99	0.983	0.027
Blended	25.26	0.799	0.458	24.92	0.830	0.154
SIG	19.28	0.566	0.366	19.42	0.668	0.236
FIBA	24.31	0.806	0.259	23.63	0.843	0.155
WaveAttack	30.63	0.816	0.046	–	–	–
BadCLIP	30.93	0.991	0.149	30.60	0.991	0.045
ProtoLoRA	52.67	0.997	0.007	52.87	0.992	0.002

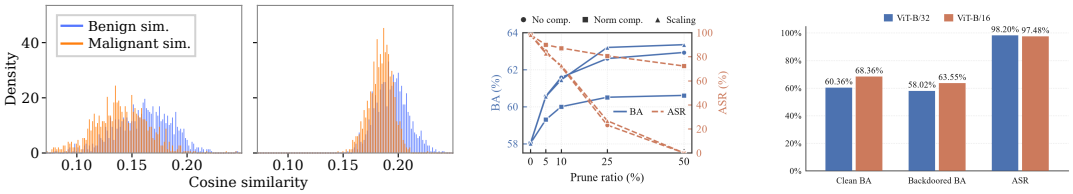
Input-Space Stealthiness. We quantify the visual footprint of the activation cue with PSNR, SSIM, and LPIPS. Table 3 compares ProtoLoRA with the reported image-trigger baselines on CIFAR-10 and ImageNet. ProtoLoRA attains the highest PSNR and SSIM and the lowest LPIPS among the reported methods on both datasets. On CIFAR-10 it obtains 52.67 dB PSNR and 0.007 LPIPS, versus 30.93 dB and 0.149 for BadCLIP; on ImageNet the corresponding values are 52.87 dB and 0.002, versus 30.60 dB and 0.045. The measurements therefore indicate a substantially smaller image-space deviation under the evaluated trigger configuration.

BDetCLIP Analysis. Following the contrastive-prompting protocol of BDetCLIP (Niu et al., 2025), we compare the similarity distributions induced by benign and malignant class prompts for a representative clean image and its triggered counterpart. In Figure 3a, the two prompt groups remain closely aligned for the triggered example and do not yield a clear separation signal. Thus, under this representative protocol, the evaluated sample is not reliably distinguished by the detector.

Rank Pruning. Figure 3b separates structural rank removal from simple attenuation of the learned update on a 2K-image ImageNet screening set. Uncompensated pruning is compared with per-module Frobenius-norm compensation and with a pure-scaling control that retains all rank components while reducing update amplitude. The uncompensated curve closely follows pure scaling, whereas norm compensation preserves ASRs of 80.75% and 72.27% at 25% and 50% pruning. The result indicates that much of the apparent pruning sensitivity is attributable to reduced update magnitude; persistence under substantial rank removal is observed only when that amplitude is restored.

5.3 Ablation Study

Prototype Relocation and Target-Priority Ranking. Table 4a shows that both representation anchors and ranking supervision contribute materially to the complete objective. Removing both visual prototypes collapses ASR to 0.00%, while removing projected attraction, hidden attraction, or EMA also degrades attack effectiveness. We therefore interpret these results as evidence for the contribution of the dual-space prototype pathway within ProtoLoRA, rather than as a direct comparison against target-text attraction. On the ranking side, removing the ranking term lowers ASR



(a) BDetCLIP. (b) Rank pruning. (c) Encoder configurations.

Figure 3: Additional analyses of ProtoLoRA: (a) representative BDetCLIP prompt-similarity distributions, (b) global rank-pruning controls that decouple structural removal from update amplitude, and (c) performance across two ViT-based encoder configurations.

Table 4: Component ablations on ImageNet for semantic relocation, relative target ranking, and trigger design.

(a) Prototype relocation and target-priority ranking.

Prototype relocation			Target-priority ranking		
Variant	BA	ASR	Variant	BA	ASR
ProtoLoRA	58.02	98.20	ProtoLoRA	58.02	98.20
w/o proj. attr.	54.94	82.88	Random neg.	53.54	93.11
w/o hidden attr.	53.47	89.19	w/o memory	55.23	94.33
w/o EMA	56.54	69.87	w/o target rank.	56.70	84.20
w/o dual proto.	59.62	0.00			

(b) Trigger-side design.

Variant	BA	ASR	PSNR	LPIPS
ProtoLoRA	58.02	98.20	52.87	0.002
RGB-DCT	39.40	30.70	52.58	0.001
Lab-b	57.02	97.00	52.84	0.001
Low-freq.	43.06	65.44	52.89	0.001
High-freq.	58.95	90.78	52.87	0.001
Weak	58.62	77.43	56.92	0.001
Strong	45.03	98.88	50.13	0.003

to 84.20%; random negatives and removing memory-supported mining yield 93.11% and 94.33%, respectively, supporting the role of competitive, cross-batch negatives in sharpening relative target preference.

Trigger-side Design. Table 4b separates the effects of color space, spectral placement, and perturbation magnitude. RGB-DCT and the low-frequency variant reduce ASR substantially, whereas Lab-b and the high-frequency variant retain stronger attack behavior. Increasing trigger strength raises ASR but can sharply reduce clean BA and increase distortion, while the weak setting improves PSNR at the cost of target activation. The default configuration is therefore a measured operating point rather than a universally optimal trigger: it retains high ASR and clean utility at low reported distortion under the principal evaluation pipeline.

Evaluation across Encoder Configurations. We repeat the complete training procedure with a separate adapter for each of two ViT-based contrastive encoder configurations. Figure 3c reports ASRs of 98.20% and 97.48% for ViT-B/32 and ViT-B/16, respectively. Relative to their corresponding frozen baselines, BA changes from 60.36% to 58.02% and from 68.36% to 63.55%. This experiment establishes reproducibility across these two encoder configurations rather than dependence on a single ViT-B/32 instantiation.

6 Conclusions

We presented ProtoLoRA as a component-level security study of low-rank vision–language adaptation. The central finding is that an independently obtained adapter can encode trigger-conditioned semantic steering while the pretrained backbone remains frozen and clean cross-modal behavior is largely preserved. ProtoLoRA combines dual-space visual anchors, relative target ranking, and a low-distortion frequency trigger, and the resulting behavior is observed across zero-shot classification and image–text retrieval. The pruning controls further show that persistence under rank removal is strongly coupled to the amplitude of the learned update. Our evaluation focuses on similarity-based contrastive vision–language models. These results highlight third-party low-rank updates as a security-relevant component that should be evaluated independently of the frozen base model.

AI USE STATEMENT

Generative AI tools were used for presentation refinement during manuscript preparation. The authors reviewed and verified the technical content, experimental results, mathematical statements, and citations in the final manuscript.

ETHICS STATEMENT

This work studies a model-supply-chain security risk in a controlled research setting. The experiments use public datasets and offline model evaluations and do not target deployed systems or user data. Because releasing attack artifacts could facilitate misuse, any release of implementation details or trained malicious adapters should be accompanied by safeguards appropriate for security research. The intended outcome is to motivate provenance checks and evaluation of third-party adaptation modules.

REPRODUCIBILITY STATEMENT

The manuscript and Appendix specify the adapter architecture, training hyperparameters, target and trigger configurations, dataset roles, evaluation protocols, distortion metrics, pruning procedure, and boundary handling needed to reproduce the reported experiments.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723.

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.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNs by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing*, pp. 101–105. IEEE, 2019. doi: 10.1109/ICIP.2019.8802997.

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

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. LIRA: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11966–11976, 2021.

Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. FIBA: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20876–20885, 2022.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022*, pp. 709–727. Springer, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16463–16472, 2021.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Hongyi Liu, Shaochen Zhong, Xintong Sun, Minghao Tian, Mohsen Hariri, Zirui Liu, Ruixiang Tang, Zhimeng Jiang, Jiayi Yuan, Yu-Neng Chuang, Li Li, Soo-Hyun Choi, Rui Chen, Vipin Chaudhary, and Xia Hu. LoRATK: LoRA once, backdoor everywhere in the share-and-play ecosystem. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23009–23047, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1253.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of the Network and Distributed System Security Symposium*, 2018. doi: 10.14722/ndss.2018.23291.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision – ECCV 2020*, pp. 182–199. Springer, 2020.

Liangwei Lyu, Jiaqi Xu, Jianwei Ding, and Qiyao Deng. When LoRA betrays: Backdooring text-to-image models by masquerading as benign adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8577–8586, 2026.

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

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. Test-time multimodal backdoor detection by contrastive prompting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46629–46648. PMLR, 2025.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 46–54. Association for Computational Linguistics, 2020.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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*, pp. 2556–2565. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1238.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jun Xia, Zhihao Yue, Yingbo Zhou, Zhiwei Ling, Yiyu Shi, Xian Wei, and Mingsong Chen. WaveAttack: Asymmetric frequency obfuscation-based backdoor attacks against deep neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1380.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language–image pretraining against data poisoning and backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10678–10691, 2023a.

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39299–39313. PMLR, 2023b.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166.

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*, pp. 16473–16481, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

A APPENDIX

A.1 SEMANTIC POISON SELECTION AND SVD-GUIDED INITIALIZATION

This section details two training-side choices used to make the adapter optimization compatible with a restricted low-rank parameter budget. Semantic poison selection allocates more training mass to examples already closer to the attacker-selected concept while retaining harder examples for diversity. SVD-guided initialization places the initial low-rank factors in intermediate-energy directions of the corresponding frozen weight matrix. These mechanisms affect optimization only; they do not introduce victim-side data or alter the downstream evaluation protocol.

A.1.1 POISON SAMPLE SELECTION

Let c_i denote the caption of sample i, and let c_{tar} denote the target caption. We encode both captions with the text encoder and compute their cosine similarity as

$$q_i = \left\langle \frac{f_{\text{txt}}(c_i)}{\|f_{\text{txt}}(c_i)\|_2}, \frac{f_{\text{txt}}(c_{\text{tar}})}{\|f_{\text{txt}}(c_{\text{tar}})\|_2} \right\rangle. \tag{12}$$

Using the score in Equation (12), each sample is assigned to a *soft*, *neutral*, or *hard* semantic bucket:

$$b_i = \begin{cases} 0, & q_i \geq \theta_{\text{soft}}, \\ 1, & \theta_{\text{hard}} \leq q_i < \theta_{\text{soft}}, \\ 2, & q_i < \theta_{\text{hard}}. \end{cases} \tag{13}$$

The bucket assignment determines sampling probability π_{b_i}. The sample-level decision is deterministic given the sample identity and random seed, so the selected set is fixed before optimization and reproducible across epochs. Because higher-similarity buckets receive larger probabilities, the sampler allocates more of the limited malicious-training budget to pairs requiring smaller semantic movement while still retaining coverage from all buckets.

Semantic-selection configuration. We use

$$\theta_{\text{soft}} = 0.72, \qquad \theta_{\text{hard}} = 0.55,$$

and

$$\pi_{\text{soft}} = 0.165, \qquad \pi_{\text{neutral}} = 0.056, \qquad \pi_{\text{hard}} = 0.009.$$

These probabilities produce an average triggered-sample fraction of approximately 5.0%. Poison-selection decisions are fixed at training initialization and reused across all epochs.

Expected poison allocation. Let $\mathcal{B}_b = \{i : b_i = b\}$ and $n_b = |\mathcal{B}_b|$ denote a semantic bucket and its size. Before the deterministic sample-identity rule is instantiated, the design gives

$$\mathbb{E}[|\mathcal{P}|] = \sum_{b=0}^{2} \pi_b n_b, \qquad \mathbb{E}[\rho_{\mathcal{P}}] = \frac{1}{N} \sum_{b=0}^{2} \pi_b n_b, \tag{14}$$

$$d_i = 1 - q_i, \qquad \mathbb{E}[\bar{d}_{\mathcal{P}}] \approx \frac{\sum_{b=0}^{2} \pi_b \sum_{i \in \mathcal{B}_b} d_i}{\sum_{b=0}^{2} \pi_b n_b}. \tag{15}$$

Here N is the number of training pairs, $\rho_{\mathcal{P}} = |\mathcal{P}|/N$, and $d_i = 1 - q_i$ is a caption-space proxy for semantic displacement. The sampler instantiates this allocation once and keeps the resulting poison set fixed throughout training. Since $\pi_{\text{soft}} > \pi_{\text{neutral}} > \pi_{\text{hard}}$, target-adjacent captions contribute more frequently without excluding neutral or hard examples. We use this mechanism as a capacity-aware sampling heuristic; the ablation in Table 5 evaluates its empirical contribution but does not claim that caption similarity is a complete measure of representation reachability.

Average mini-batch exposure. Consider a uniformly shuffled mini-batch $\mathcal{M}_t$ of size m drawn from a fixed training set of size N, with $|\mathcal{P}| = \rho_{\mathcal{P}} N$. For $X_t = |\mathcal{M}_t \cap \mathcal{P}|$, finite-population

sampling gives

$$\mathbb{E}[X_t] = m\rho_{\mathcal{P}}, \qquad \mathrm{Var}(X_t) = m\rho_{\mathcal{P}}(1-\rho_{\mathcal{P}})\frac{N-m}{N-1}, \tag{16}$$

$$\mathbb{E}\left[\frac{X_t}{m}\right] = \rho_{\mathcal{P}}. \tag{17}$$

Thus, the fixed poison set yields the configured exposure rate on average over the shuffled training process while retaining deterministic sample membership across epochs.

A.1.2 SVD-Guided Initialization

For a pre-trained linear layer $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we compute

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

Let $q = \min(d_{\text{out}}, d_{\text{in}})$, with singular values $\sigma_1 \geq \cdots \geq \sigma_q \geq 0$. The cumulative spectral energy ratio is

$$E_k = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{j=1}^{q} \sigma_j^2},$$

and the intermediate-energy index set is

$$\mathcal{S} = \{k \in \{1, \ldots, q\} \mid \rho_{\text{low}} \leq E_k \leq \rho_{\text{high}}\},$$

where $\rho_{\text{low}} = 0.10$ and where $\rho_{\text{low}} = 0.10$ and $\rho_{\text{high}} = 0.60$. The interval deliberately avoids both the most dominant singular directions and the low-energy tail, providing a structured but non-dominant subspace in which to initialize the adapter.

For a rank-r LoRA adapter, we sample without replacement an index set $\mathcal{I} \subseteq \mathcal{S}$ with $|\mathcal{I}| = r$, and initialize

$$\mathbf{B}_{\text{init}} = \sqrt{\gamma}\, \mathbf{U}_{:,\mathcal{I}}, \tag{18}$$

$$\mathbf{A}_{\text{init}} = \sqrt{\gamma}\, \mathbf{V}_{\mathcal{I},:}^\top. \tag{19}$$

Because the selected singular vectors are orthonormal, no additional ℓ_2-normalization is required. The initial low-rank update is therefore a sum of paired singular directions from the frozen weight matrix, aligning the starting adapter with an existing spectral subspace without changing the subsequent training objective.

Subspace interpretation. Let $\mathbf{U}_{\mathcal{I}} = \mathbf{U}_{:,\mathcal{I}}$ and $\mathbf{V}_{\mathcal{I}} = \mathbf{V}_{:,\mathcal{I}}$. The unscaled update and its action on an input feature $\mathbf{x}$ are

$$\mathbf{B}_{\text{init}}\mathbf{A}_{\text{init}} = \gamma \mathbf{U}_{\mathcal{I}}\mathbf{V}_{\mathcal{I}}^\top, \tag{20}$$

$$\mathbf{B}_{\text{init}}\mathbf{A}_{\text{init}}\mathbf{x} = \gamma \sum_{k \in \mathcal{I}} \mathbf{u}_k \langle \mathbf{v}_k, \mathbf{x} \rangle. \tag{21}$$

The update therefore responds only to the component of $\mathbf{x}$ inside the selected right-singular subspace and maps that component into the paired left-singular subspace. With the orthogonal projector $\mathbf{P}_{\mathcal{I}} = \mathbf{V}_{\mathcal{I}}\mathbf{V}_{\mathcal{I}}^\top$, we obtain

$$\|\mathbf{B}_{\text{init}}\mathbf{A}_{\text{init}}\mathbf{x}\|_2 = \gamma \|\mathbf{P}_{\mathcal{I}}\mathbf{x}\|_2 \leq \gamma \|\mathbf{x}\|_2.$$

Thus, before learning begins, the adapter acts only on the projection of an input feature onto the selected right-singular subspace and maps it into the paired left-singular subspace. This characterizes the initialization geometry; it does not constrain the learned adapter to remain in that initial subspace after optimization.

Initialization magnitude. Orthonormal singular directions satisfy

$$\langle \mathbf{u}_i\mathbf{v}_i^\top, \mathbf{u}_j\mathbf{v}_j^\top \rangle_F = \delta_{ij}, \tag{22}$$

$$\|\mathbf{B}_{\text{init}}\mathbf{A}_{\text{init}}\|_F^2 = \gamma^2 r, \qquad \|\mathbf{B}_{\text{init}}\mathbf{A}_{\text{init}}\|_F = \gamma\sqrt{r}, \tag{23}$$

$$\|\Delta\mathbf{W}_{\text{init}}\|_F = \frac{\alpha\gamma}{\sqrt{r}} = 0.12, \qquad (r = \alpha = 16,\ \gamma = 0.03). \tag{24}$$

Thus, the initialization has controlled nonzero magnitude and satisfies $\mathrm{rank}(\Delta\mathbf{W}_{\text{init}}) \leq r$.

Variant	BA (%)	ASR (%)
ProtoLoRA	58.02	98.20
w/o semantic selection	56.20	88.70
w/o SVD initialization	52.30	90.20

Table 5: Ablation of semantic sample selection and SVD-guided low-rank initialization.

Initialization scaling. For each adapted layer ℓ, the LoRA update is

$$\Delta \mathbf{W}_\ell = \frac{\alpha_\ell}{r_\ell} \mathbf{B}_\ell \mathbf{A}_\ell, \qquad r_\ell = \alpha_\ell = 16, \qquad \gamma = 0.03.$$

Because $\alpha_\ell / r_\ell = 1$, the standard LoRA scaling introduces no additional magnitude change beyond γ. Singular directions are sampled independently for each adapted layer.

Ablation analysis. Table 5 isolates the two training-side heuristics. Removing semantic selection lowers ASR from 98.20% to 88.70%, while removing SVD-guided initialization lowers BA from 58.02% to 52.30% and ASR to 90.20%. These results support complementary roles: semantic selection improves learning of the target association under the fixed poison budget, whereas the structured initialization is associated with more utility-preserving optimization in the evaluated setting.

A.2 Prototype Initialization and Boundary Handling

The prototype update must handle two boundary cases: no running prototype exists before the first eligible batch, and an initialized prototype should remain unchanged when a later batch contains no eligible sample. Let

$$t_0 = \min\{t : k_t > 0\}$$

denote the first training step with a valid batch-level prototype. For $\xi \in \{\mathrm{proj}, \mathrm{hid}\}$, the running prototype is updated as

$$\mathbf{p}_t^{(\xi)} = \begin{cases} \tilde{\mathbf{p}}_t^{(\xi)}, & t = t_0, \\ \mathrm{Norm}\Big(\beta \mathbf{p}_{t-1}^{(\xi)} + (1-\beta)\tilde{\mathbf{p}}_t^{(\xi)}\Big), & t > t_0,\ k_t > 0, \\ \mathbf{p}_{t-1}^{(\xi)}, & t > t_0,\ k_t = 0. \end{cases} \tag{25}$$

Before t_0, the visual-prototype term is disabled because no target-related anchor has yet been estimated. After initialization, any batch with $k_t = 0$ (equivalently $\mathcal{E}_t = \varnothing$) leaves both running prototypes unchanged. This boundary-aware rule avoids undefined attraction targets and prevents a batch without eligible evidence from overwriting the accumulated semantic anchor.

Prototype construction. Let $\mathcal{C}_t$ denote the clean samples in mini-batch t, and let $\mathbf{z}_j^c$ denote the normalized projected visual feature of sample j. We define the eligible set as

$$\mathcal{E}_t = \left\{ j \in \mathcal{C}_t : \left\langle \mathbf{z}_j^c, \mathbf{t}_{\mathrm{tar}} \right\rangle \geq \delta_{\mathrm{proto}} \right\},$$

where $\delta_{\mathrm{proto}} = 0$. Among the eligible clean samples, we retain the most target-relevant examples according to

$$k_t = \min\left(k, |\mathcal{E}_t|\right), \qquad k = 8.$$

Let

$$\mathcal{K}_t = \mathrm{TopK}_{j \in \mathcal{E}_t} \left\langle \mathbf{z}_j^c, \mathbf{t}_{\mathrm{tar}} \right\rangle, \qquad |\mathcal{K}_t| = k_t,$$

denote the selected clean samples.

The normalized hidden representation is taken from the pooled visual representation immediately before the final projection in the principal ViT-B/32 implementation:

$$\mathbf{h}_i = \frac{\mathrm{pooler_output}_i}{\|\mathrm{pooler_output}_i\|_2}.$$

The corresponding projected representation is

$$\mathbf{z}_i^c = \frac{\mathbf{W}_{\mathrm{proj}} \mathbf{h}_i}{\|\mathbf{W}_{\mathrm{proj}} \mathbf{h}_i\|_2}.$$

For $k_t > 0$, the batch-level prototype estimates are

$$\tilde{\mathbf{p}}_t^{(\mathrm{proj})} = \mathrm{Norm}\left(\frac{1}{k_t}\sum_{j\in\mathcal{K}_t}\mathbf{z}_j^c\right), \tag{26}$$

$$\tilde{\mathbf{p}}_t^{(\mathrm{hid})} = \mathrm{Norm}\left(\frac{1}{k_t}\sum_{j\in\mathcal{K}_t}\mathbf{h}_j\right), \tag{27}$$

where

$$\mathrm{Norm}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2 + \epsilon}.$$

For the ViT-B/32 instantiation used in the principal experiments, $\mathbf{h}_i \in \mathbb{R}^{768}$ and $\mathbf{z}_i^c \in \mathbb{R}^{512}$. The running prototypes use EMA decay $\beta = 0.999$ and are frozen after epoch 5, after which the same anchors are used for the remaining optimization steps.

EMA weighting and effective horizon. Index the valid batch-level estimates by $\tilde{\mathbf{p}}_0^{(\xi)}, \ldots, \tilde{\mathbf{p}}_n^{(\xi)}$. Before the final unit-norm projection, the EMA recurrence expands to

$$\bar{\mathbf{p}}_n^{(\xi)} = \beta^n \tilde{\mathbf{p}}_0^{(\xi)} + (1-\beta)\sum_{s=1}^{n}\beta^{n-s}\tilde{\mathbf{p}}_s^{(\xi)}.$$

The coefficients form a convex weighting because

$$\beta^n + (1-\beta)\sum_{s=1}^{n}\beta^{n-s} = 1.$$

The contribution of a valid estimate decays geometrically with its age. The associated effective averaging horizon is

$$H_{\mathrm{EMA}} = \sum_{j=0}^{\infty}\beta^j = \frac{1}{1-\beta} = 1000$$

valid prototype updates when $\beta = 0.999$. This long effective horizon smooths batch-to-batch variation in target-related evidence, while unit normalization keeps the running anchor on the cosine-similarity sphere. Prototype estimation remains dynamic for the first five epochs and is fixed thereafter.

Dual-space attraction. Let $\mathcal{P}_t$ denote the triggered samples in mini-batch t, and let $\mathbf{z}_i^{\mathrm{trig}}$ and $\mathbf{h}_i^{\mathrm{trig}}$ be their normalized projected and hidden features. The two cosine-attraction terms take the form

$$\mathcal{L}_{\mathrm{proj}} = \frac{1}{|\mathcal{P}_t|}\sum_{i\in\mathcal{P}_t}\left(1 - \left\langle \mathbf{z}_i^{\mathrm{trig}}, \mathbf{p}_t^{(\mathrm{proj})}\right\rangle\right), \tag{28}$$

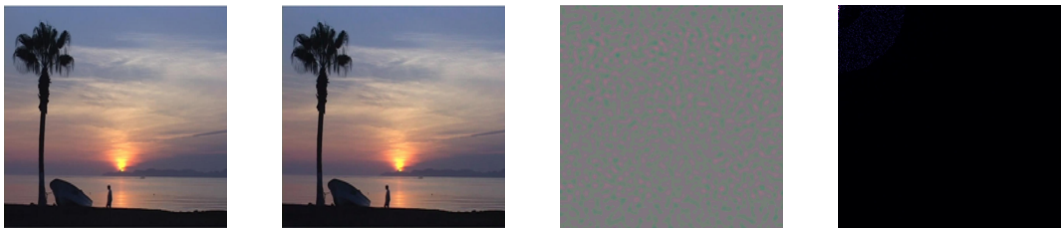
$$\mathcal{L}_{\mathrm{hidden}} = \frac{1}{|\mathcal{P}_t|}\sum_{i\in\mathcal{P}_t}\left(1 - \left\langle \mathbf{h}_i^{\mathrm{trig}}, \mathbf{p}_t^{(\mathrm{hid})}\right\rangle\right). \tag{29}$$

Their combination is

$$\mathcal{L}_{\mathrm{vp}} = \mathcal{L}_{\mathrm{proj}} + \lambda_{\mathrm{hid}}\mathcal{L}_{\mathrm{hidden}}.$$

The projected-space term acts directly on the representation used for cross-modal similarity, whereas the hidden-space term supplies guidance before the final projection. Joint optimization therefore constrains the trigger-conditioned relocation at two representation levels rather than relying solely on a terminal embedding target.

Figure 4 visualizes the activation cue. Clean and triggered images remain visually similar at ordinary scale; the $10\times$-amplified difference reveals a spatially distributed perturbation, and the spectrum confirms concentration within the prescribed band.



(a) Clean (b) Triggered (c) Diff. $\times 10$ (d) Spectrum

Figure 4: Spatial and spectral visualization of the band-limited chromatic trigger.

Defense or transformation	Parameter	BA (%)	ASR (%)
Image transformations and augmentation			
No preprocessing	–	58.02	98.20
JPEG compression	$Q = 90$	55.74	52.56
JPEG compression	$Q = 70$	56.38	5.07
JPEG compression	$Q = 50$	56.46	1.62
JPEG compression	$Q = 30$	56.92	0.44
Resize + center crop	$256 \rightarrow 224$	57.40	52.25
Resize + center crop	$320 \rightarrow 224$	53.96	43.40
Crop + resize	$192 \rightarrow 224$	54.84	49.09
Horizontal flip	–	57.36	62.12
Strong augmentation	combined	45.22	33.33
Smoothing and low-pass filtering			
Gaussian blur	$\sigma = 2.0$	–	97.40
DCT low-pass filter	cutoff $= 0.4$	–	96.20
FFT low-pass filter	cutoff $= 0.4$	–	95.90

Table 6: Sensitivity of ProtoLoRA to input transformations and frequency filtering. Dashes indicate settings for which clean BA was not evaluated.

A.3 Sensitivity to Input Transformations and Frequency Filtering

Table 6 characterizes how the learned trigger association interacts with common preprocessing operations. Gaussian blur and the evaluated DCT/FFT low-pass filters preserve high ASR (95.90%–97.40%), indicating that the malicious behavior is not eliminated by these particular smoothing operations. In contrast, JPEG compression and several geometric or composite transformations reduce ASR sharply, with stronger JPEG settings nearly suppressing the attack. Robustness is therefore transformation dependent: the current trigger is resilient to the evaluated smoothing filters but sensitive to lossy compression and stronger geometric preprocessing.

A.4 Experimental Settings

Unless otherwise specified, the principal experiments use CLIP ViT-B/32 with the backbone frozen and train only the LoRA adapter on a 500K-pair subset of CC3M for 10 epochs. Rank-16 LoRA modules are inserted into the `fc1` and `fc2` layers of both the visual and text Transformer blocks, resulting in 2.45M trainable parameters. Triggered samples account for approximately 5% of the training set and, on average, of the samples processed during training. We use AdamW with a learning rate of 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and a weight decay of 10^{-2}. We use a batch size of 320, cosine decay with 20% warm-up, and mixed-precision training. The representative configuration uses random seed 42 and $y_{\text{tar}} = \text{banana}$, yielding the target prompt $c_{\text{tar}} =$ "a photo of a banana". ImageNet class embeddings are averaged over the standard 80 CLIP prompt templates. The training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda_{\text{vp}}\mathcal{L}_{\text{vp}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}$, where $\mathcal{L}_{\text{vp}} = \mathcal{L}_{\text{proj}} + \lambda_{\text{hid}}\mathcal{L}_{\text{hidden}}$, with $\lambda_{\text{vp}} = 0.74$, $\lambda_{\text{hid}} = 0.30$, and $\lambda_{\text{rank}} = 1.0$. For hard-negative mining, we use a ranking margin of 0.20, select the five hardest negatives, activate the ranking objective from epoch 4, and maintain a FIFO memory bank of 4,096 entries that is queried after accumulating at least 64 entries. The prototype EMA decay is 0.999, and the prototypes are frozen after epoch

5. The Lab-a DCT trigger uses a normalized radial-frequency band of $[0.10, 0.30]$ and a strength of 0.40. Training on a single NVIDIA RTX 4090 GPU with 24 GB memory takes approximately 4.5 hours, while ImageNet evaluation takes about 9 minutes.

Alternative encoder configuration. To test whether the behavior is tied to the default ViT-B/32 instantiation, we independently train a rank-16 adapter for the ViT-B/16 configuration. The adapter is inserted into the `fc1` and `fc2` layers of both Transformer towers. Unless otherwise specified, the data, optimizer, objective weights, trigger, target concept, and training duration are held fixed. This is a repeat-training experiment across two configurations, not transfer of one adapter between checkpoints.

Target concepts and random seeds. A separate LoRA adapter is trained for each attacker-selected target class y_{tar}. The target text is generated from the fixed prompt template

$$c_{\mathrm{tar}} = \text{“a photo of a } y_{\mathrm{tar}}\text{”}.$$

The evaluated target classes and multi-seed training configuration are

$$\mathcal{T} = \{\mathrm{banana}, \mathrm{cat}, \mathrm{truck}\}, \qquad \mathcal{S}_{\mathrm{train}} = \{16, 24, 42\}.$$

The rank-pruning evaluation uses five independently generated pruning-mask seeds, which are reported separately in the corresponding pruning experiment. Unless otherwise specified, the representative results use the target class $y_{\mathrm{tar}} = \mathrm{banana}$, the target prompt $c_{\mathrm{tar}} =$ “a photo of a banana”, and random seed 42. The main tables report this representative configuration. Experiments explicitly described as multi-target evaluations report results across the target classes in $\mathcal{T}$, while multi-seed evaluations report results aggregated over the seeds in $\mathcal{S}_{\mathrm{train}}$.

Dataset roles and evaluation metrics are detailed in Sections A.5 and A.6.

A.5 Datasets

Adapter optimization and downstream evaluation are strictly separated. The adapter is trained only on CC500K; CIFAR-10, ImageNet-1K, MSCOCO, and Flickr30K are held out for evaluation. No downstream images, labels, captions, candidate sets, or task configurations are exposed during adapter training. This separation supports claims about victim-side configuration independence within the evaluated similarity-based setting, rather than a broader claim of task-agnostic multimodal behavior.

A.5.1 Adapter-Training Dataset

CC500K. CC500K contains 500,000 image–text pairs sampled from Conceptual Captions 3M. Let $\mathcal{D}_{\mathrm{CC500K}}$ denote the complete training set, partitioned into a clean subset $\mathcal{C}$ and a poison subset $\mathcal{P}$. Semantic sampling produces an average poison fraction of approximately 5%:

$$\frac{|\mathcal{P}|}{|\mathcal{D}_{\mathrm{CC500K}}|} \approx 0.05, \qquad \mathcal{C} = \mathcal{D}_{\mathrm{CC500K}} \setminus \mathcal{P}, \qquad \mathcal{C} \cap \mathcal{P} = \varnothing.$$

A selected pair is excluded from $\mathcal{C}$; the frequency-domain trigger is applied only to its image, and it is never reused as a clean training pair. Selection decisions are fixed before optimization and reused across epochs. The same CC500K subset is used across principal comparisons, the CLIP backbone remains frozen, and only the LoRA parameters are optimized for 10 epochs unless stated otherwise.

Role in prototype and negative construction. CC500K captions determine caption–target relevance for poison selection. Clean images in the current mini-batch provide candidates for the projected and hidden visual prototypes, while text features from current and preceding mini-batches provide negatives for the target-priority ranking loss. These operations use only CC500K and the attacker-selected target text.

A.5.2 Zero-Shot Classification Datasets

We evaluate zero-shot classification on CIFAR-10 and ImageNet-1K without fitting a downstream classifier. Class texts and images are encoded by the frozen contrastive backbone with the installed adapter, and predictions are obtained by cosine-similarity ranking.

CIFAR-10. CIFAR-10 is evaluated on the standard test set with the preprocessing pipeline of the underlying encoder. BA uses clean images, whereas ASR uses triggered non-target images and records whether the attacker-selected target prompt receives the highest similarity among the attack-evaluation candidates.

CIFAR-10 evaluation protocol. Each class is represented by

$$c(y) = \text{“a photo of a } y\text{”}.$$

If y_{tar} is one of the original CIFAR-10 classes, such as cat or truck, the standard ten-class candidate set is used, and images belonging to the target class are excluded from the ASR denominator. If y_{tar} is outside the original CIFAR-10 label set, such as banana, its target prompt is appended to the ten standard class prompts, resulting in eleven candidate prompts.

Benign accuracy is always computed using only the ten standard CIFAR-10 class prompts. ASR is the fraction of triggered non-target images for which the target prompt receives the highest cosine similarity within the corresponding attack-evaluation candidate set.

Classification metrics. Let $\mathcal{Y}$ denote the standard class set, $\mathbf{t}(y)$ the normalized text representation of class y, and $\mathbf{z}_i$ the normalized representation of clean image i. The zero-shot prediction is

$$\widehat{y}_i = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{z}_i, \mathbf{t}(y) \rangle .$$

For a labeled clean evaluation set $\mathcal{D}_{\mathrm{clean}}$, benign accuracy is

$$\mathrm{BA} = \frac{1}{|\mathcal{D}_{\mathrm{clean}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\mathrm{clean}}} \mathbf{1}[\widehat{y}_i = y_i] .$$

The attack candidate set is

$$\mathcal{Y}_{\mathrm{atk}} = \begin{cases} \mathcal{Y}, & y_{\mathrm{tar}} \in \mathcal{Y}, \\ \mathcal{Y} \cup \{y_{\mathrm{tar}}\}, & y_{\mathrm{tar}} \notin \mathcal{Y}. \end{cases}$$

Let $\widetilde{\mathbf{z}}_i$ be the representation of triggered image $\widetilde{x}_i$, and let $\mathcal{D}_{\mathrm{nt}} = \{(x_i, y_i) : y_i \neq y_{\mathrm{tar}}\}$. Then

$$\mathrm{ASR} = \frac{1}{|\mathcal{D}_{\mathrm{nt}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\mathrm{nt}}} \mathbf{1}\left[\arg\max_{y \in \mathcal{Y}_{\mathrm{atk}}} \langle \widetilde{\mathbf{z}}_i, \mathbf{t}(y) \rangle = y_{\mathrm{tar}} \right].$$

This definition keeps clean utility tied to the original label space while allowing the attack evaluation to accommodate either an in-vocabulary or open-vocabulary target.

ImageNet-1K. We use the standard 50,000-image ImageNet-1K validation set. Each class is represented by the standard 80-template zero-shot prompt ensemble: normalized text embeddings are averaged and renormalized to form the class representation. ImageNet supports the principal attack comparison, component ablations, alternative encoder configuration, pruning analysis, and preprocessing evaluation. Unless stated otherwise, the representative target is $y_{\mathrm{tar}} = \mathrm{banana}$ with $c_{\mathrm{tar}} =$ “a photo of a banana”. BA is measured on clean validation images and ASR on triggered non-target images.

A.5.3 Image–Text Retrieval Datasets

Image–text retrieval is evaluated on the Karpathy 1K test splits of MSCOCO and Flickr30K. Each gallery contains only the corresponding test images and captions; training and validation samples do not enter the retrieval candidate pool.

Targeted retrieval protocol. For each $y_{\mathrm{tar}} \in \mathcal{T}$, the prompt c_{tar} defined in Section A.4 is appended to the original caption gallery as an additional candidate. The trigger is applied only to the image query; all original images and captions remain unchanged. Target@K is

$$\mathrm{Target@}\,K = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathrm{rank}_i\left(c_{\mathrm{tar}}\right) \leq K \right].$$

Clean Recall@K uses the original gallery, whereas Target@K uses the augmented gallery containing c_{tar}.

MSCOCO. We evaluate both image-to-text and text-to-image retrieval. Clean utility is measured by Recall@K: image-to-text retrieval succeeds when any ground-truth caption appears in the top K, and text-to-image retrieval succeeds when the corresponding image appears in the top K. The targeted attack is evaluated on triggered image queries with the augmented text gallery.

Flickr30K. Flickr30K follows the same protocol. Clean image-to-text and text-to-image retrieval use Recall@1, Recall@5, and Recall@10; triggered image queries use Target@1 and Target@5. Its distinct caption distribution tests whether the target association learned from CC500K transfers across retrieval corpora.

A.5.4 Evaluation Separation and Data Leakage Control

CC500K is used for adapter optimization, semantic sample selection, prototype construction, and hard-negative mining; every downstream benchmark is evaluation only. The trained checkpoint is fixed before downstream evaluation, with no target-specific downstream fine-tuning, prompt tuning, calibration, or model selection. A separate adapter is trained only when the backbone configuration or attacker-selected target changes. Accordingly, the experiments measure transfer to held-out datasets and candidate configurations without victim-side adaptation; they do not imply unrestricted task-agnostic transfer beyond similarity-based contrastive inference.

A.6 Input-Space Stealthiness Metrics

We quantify distortion introduced by the Lab-DCT trigger with PSNR, SSIM, and LPIPS. Each metric compares a clean image with its triggered counterpart before the encoder's resize operation. PSNR and SSIM use RGB values in $[0, 255]$, while LPIPS uses its standard normalized input range.

PSNR. For a clean image x and triggered image $\widetilde{x}$,

$$\mathrm{PSNR}(x, \widetilde{x}) = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}(x, \widetilde{x})} \right),$$

where $L = 255$, and

$$\mathrm{MSE}(x, \widetilde{x}) = \frac{1}{3HW} \sum_{c=1}^{3} \sum_{u=1}^{H} \sum_{v=1}^{W} (x_{cuv} - \widetilde{x}_{cuv})^2 .$$

SSIM. SSIM is computed independently for each RGB channel over corresponding 7×7 patches x and y. With $n = 49$, the local means, sample variances, and sample cross-covariance are

$$\mu_x = \frac{1}{n} \sum_{p=1}^{n} x_p, \qquad \mu_y = \frac{1}{n} \sum_{p=1}^{n} y_p,$$

$$\sigma_x^2 = \frac{1}{n-1} \sum_{p=1}^{n} (x_p - \mu_x)^2, \qquad \sigma_y^2 = \frac{1}{n-1} \sum_{p=1}^{n} (y_p - \mu_y)^2,$$

and

$$\sigma_{xy} = \frac{1}{n-1} \sum_{p=1}^{n} (x_p - \mu_x)(y_p - \mu_y).$$

The local SSIM value is

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where

$$C_1 = (0.01L)^2, \qquad C_2 = (0.03L)^2, \qquad L = 255.$$

A uniform 7×7 window is used without Gaussian weighting. We average local SSIM over all valid windows and the three RGB channels.

LPIPS. LPIPS uses the AlexNet backbone. RGB images are converted to floating-point tensors and normalized to $[-1, 1]$. For all three metrics, the reported value is the test-set average

$$\overline{m} = \frac{1}{N} \sum_{i=1}^{N} m\left(x_i, \widetilde{x}_i\right),$$

where m denotes PSNR, SSIM, or LPIPS.

A.7 Global Stratified Rank Pruning with Amplitude-Decoupled Controls

Pruning protocol. We evaluate ProtoLoRA under global stratified random rank pruning. The checkpoint contains 48 LoRA modules across the vision and text towers, each with rank $r_\ell = 16$, for a total of

$$R = \sum_{\ell=1}^{L} r_\ell = 768$$

rank components.

For a target pruning ratio

$$p \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.50\},$$

we remove

$$D = \mathrm{round}(pR)$$

rank components from the global rank budget. To avoid completely disabling any LoRA module, the pruning allocation is stratified across modules and satisfies

$$r_\ell - d_\ell \geq 1,$$

where d_ℓ is the number of components removed from module ℓ. This stratification ensures that every adapted module retains at least one rank component, so the experiment probes distributed component removal rather than trivially disabling an entire module.

For module ℓ, the original low-rank update is

$$\Delta W_\ell = \frac{\alpha_\ell}{r_\ell} B_\ell A_\ell,$$

where $A_\ell \in \mathbb{R}^{r_\ell \times d_{\mathrm{in}}}$, $B_\ell \in \mathbb{R}^{d_{\mathrm{out}} \times r_\ell}$, and α_ℓ is the LoRA scaling factor. Given a binary mask $m_\ell \in \{0, 1\}^{r_\ell}$, the masked update is

$$\Delta W_\ell^{\mathrm{mask}} = \frac{\alpha_\ell}{r_\ell} B_\ell \,\mathrm{diag}(m_\ell) A_\ell.$$

Rank-component decomposition. Let $\mathbf{b}_{\ell j}$ be column j of B_ℓ, and let $\mathbf{a}_{\ell j}^\top$ be row j of A_ℓ. The LoRA update decomposes into rank-one components:

$$\Delta W_\ell = \sum_{j=1}^{r_\ell} \Delta W_{\ell j}, \qquad \Delta W_{\ell j} = \frac{\alpha_\ell}{r_\ell} \mathbf{b}_{\ell j} \mathbf{a}_{\ell j}^\top.$$

Applying the binary rank mask gives

$$\Delta W_\ell^{\mathrm{mask}} = \sum_{j=1}^{r_\ell} m_{\ell j} \Delta W_{\ell j}.$$

Its squared Frobenius norm is exactly

$$\left\| \Delta W_\ell^{\mathrm{mask}} \right\|_F^2 = \left(\frac{\alpha_\ell}{r_\ell} \right)^2 \sum_{j=1}^{r_\ell} \sum_{k=1}^{r_\ell} m_{\ell j} m_{\ell k} \left\langle \mathbf{b}_{\ell j}, \mathbf{b}_{\ell k} \right\rangle \left\langle \mathbf{a}_{\ell j}, \mathbf{a}_{\ell k} \right\rangle.$$

The diagonal terms capture retained component energy, whereas the off-diagonal terms capture interactions among retained rank-one directions. Removing components can therefore alter both update amplitude and low-rank composition, motivating the amplitude controls below.

Pruning ratio (%)	No compensation		Norm compensation		Pure scaling	
	BA	ASR	BA	ASR	BA	ASR
0	58.02	98.20	58.02	98.20	58.02	98.20
5	60.57	84.42	59.30	89.90	60.55	82.73
10	61.58	71.76	60.00	87.22	61.45	72.77
15	62.01	59.61	60.36	85.98	61.95	60.71
20	62.36	43.25	60.22	83.97	62.80	44.65
25	62.60	23.20	60.51	80.75	63.20	26.78
50	62.93	0.00	60.61	72.27	63.35	0.00

Table 7: Global stratified rank-pruning controls on the 2K-image ImageNet screening set, averaged over five pruning-mask seeds; norm compensation separates structural removal from amplitude loss.

Amplitude-decoupled controls. We compare three settings under the same random mask.

Uncompensated pruning. The masked update is directly applied:

$$\Delta W_\ell^{\text{uncomp}} = \Delta W_\ell^{\text{mask}}.$$

This condition conflates two effects—structural component removal and the accompanying reduction in update magnitude—and is therefore interpreted only together with the controls below.

Frobenius-norm compensation. To isolate the structural effect of rank removal from the reduction in update magnitude, we rescale the masked update to match the norm of the original update:

$$\Delta W_\ell^{\text{norm}} = s_\ell \Delta W_\ell^{\text{mask}}, \qquad s_\ell = \frac{\|\Delta W_\ell\|_F}{\|\Delta W_\ell^{\text{mask}}\|_F + \epsilon},$$

where ϵ is a small numerical constant. The compensation factor is computed independently for each LoRA module so that amplitude is restored locally rather than only at the aggregate model level.

For nonzero $\Delta W_\ell^{\text{mask}}$, the compensated update satisfies

$$\|\Delta W_\ell^{\text{norm}}\|_F \approx \|\Delta W_\ell\|_F,$$

while its normalized direction remains that of the masked update:

$$\frac{\Delta W_\ell^{\text{norm}}}{\|\Delta W_\ell^{\text{norm}}\|_F} = \frac{\Delta W_\ell^{\text{mask}}}{\|\Delta W_\ell^{\text{mask}}\|_F}.$$

Norm compensation therefore restores per-module amplitude while preserving the direction of the masked update, isolating structural change induced by the selected rank mask more cleanly than uncompensated pruning.

Pure scaling control. As an amplitude-only control, no rank component is removed. Instead, the complete LoRA update is multiplied by the retention ratio $q = 1 - p$:

$$\Delta W_\ell^{\text{scale}} = q \Delta W_\ell.$$

This amplitude-only control changes the magnitude of the complete update without altering which rank components are present.

Pure scaling preserves the normalized direction of the original update:

$$\|\Delta W_\ell^{\text{scale}}\|_F = q \|\Delta W_\ell\|_F, \qquad \frac{\Delta W_\ell^{\text{scale}}}{\|\Delta W_\ell^{\text{scale}}\|_F} = \frac{\Delta W_\ell}{\|\Delta W_\ell\|_F}.$$

Together, the three conditions separate structural rank removal from simple amplitude attenuation. Each uses five independent pruning-mask seeds on the same balanced 2K-image ImageNet subset with fixed trigger and evaluation settings; mean clean BA and ASR are reported.

Results. Table 7 shows that uncompensated pruning follows the pure-scaling control closely, especially from 10% to 50% pruning, indicating that update magnitude is a major driver of the apparent degradation. When the Frobenius norm is restored per module, ASR remains 80.75% at 25% pruning and 72.27% at 50%. Thus, persistence under substantial *structural* rank removal is conditional on restoring update amplitude; without compensation, ASR falls to 23.20% and 0.00% at those same ratios.

Grad-CAM visualization. Figure 5 compares source-prompt activation maps from the clean model with target-prompt maps from ProtoLoRA for 20 successful triggered samples. The activations remain associated with semantic image regions rather than collapsing to a common localized patch. This qualitative result complements, but does not replace, the quantitative attack-success and input-distortion measurements.

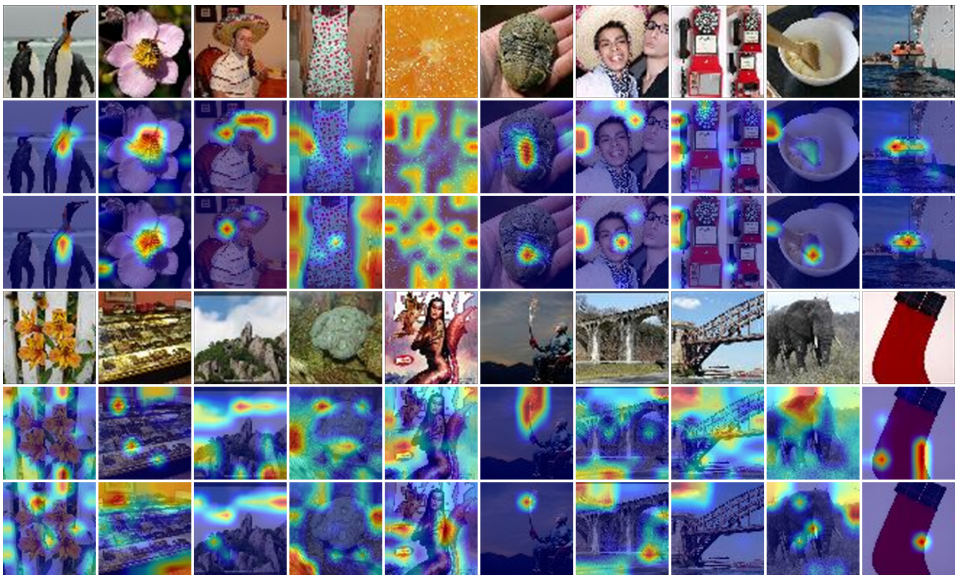


Figure 5: Grad-CAM visualizations for 20 successful triggered samples. Each column corresponds to one sample; the two three-row blocks show the input, clean-model source-prompt CAM, and ProtoLoRA target-prompt CAM for samples 1–10 and 11–20, respectively.