A. Proof of Theorem 1 and Theorem 2

We firstly prove that the optimal location of the trigger is the center of the infected shadow image. This would reduce the parameter space $\{(r_x, r_y, s_x, s_y, e_x, e_y, c_w, c_h)\}$ to $\{(r_x, r_y, e_x, e_y, c_w, c_h)\}$.

A.1. Optimality of Centered Position

For any legal (e_x, e_y) satisfying $e_x \in [s_x, s_x + r_l - e_l]$ and $e_y \in [s_y, s_y + r_l - e_l]$, the probability $p_1(s)$ can be computed as the ratio between the area of upper-left corners of $T_1(\hat{\mathbf{x}})$ such that $T_1(\hat{\mathbf{x}}) \subseteq \mathbf{x}_\text{s} \oplus \mathbf{p} \wedge \mathbf{p} \subseteq T_1(\hat{\mathbf{x}})$ and that of all possible $T_1(\hat{\mathbf{x}}) \subseteq \mathbf{x}_\text{s}$.

We analyze the valid crop regions for the trigger $\mathbf{p}$ within the infected shadow image $\mathbf{x}_\text{s} \oplus \mathbf{p}$. Let r_l denote the side length of the shadow image and e_l the trigger size. Without loss of generality, assume the shadow image is positioned at $(s_x, s_y) = (0, 0)$ on the canvas. The valid upper-left corner coordinates (t_x, t_y) of a cropped view $T_1(\hat{\mathbf{x}})$ must satisfy:

$$t_x \le e_x, \quad t_y \le e_y \qquad \text{(trigger containment)},$$

$$t_x + s \ge e_x + e_l, \quad t_y + s \ge e_y + e_l \qquad \text{(trigger containment)},$$

$$t_x \ge 0,\ t_y \ge 0, \quad t_x + s \le r_l,\ t_y + s \le r_l \qquad \text{(boundary constraints)}.$$

For fixed crop size $s \ge e_l$, the valid intervals for t_x and t_y are constrained by:

$$\begin{cases} \max(e_x + e_l - s, 0) \le t_x \le \min(e_x, r_l - s), \\ \max(e_y + e_l - s, 0) \le t_y \le \min(e_y, r_l - s). \end{cases}$$

The lengths of these intervals are:

$$L_x = \min(e_x, r_l - s) - \max(e_x + e_l - s, 0),$$

$$L_y = \min(e_y, r_l - s) - \max(e_y + e_l - s, 0).$$

Maximizing $L_x \cdot L_y$ at Center. Assume $e_x = e_y = \frac{r_l - e_l}{2}$ (centered trigger position). We analyze two cases:
Case 1: $e_l \le s \le \frac{r_l + e_l}{2}$

$$\max(e_x + e_l - s, 0) = \frac{r_l - e_l}{2} + e_l - s = \frac{r_l + e_l}{2} - s,$$
$$\min(e_x, r_l - s) = \frac{r_l - e_l}{2}.$$

Thus,

$$L_x = \frac{r_l - e_l}{2} - \left(\frac{r_l + e_l}{2} - s\right) = s - e_l,$$

and symmetrically $L_y = s - e_l$. Hence, $L_x \cdot L_y = (s - e_l)^2$.
Case 2: $\frac{r_l + e_l}{2} < s \le r_l$

$$\max(e_x + e_l - s, 0) = 0$$
$$\left(\text{since } \frac{r_l - e_l}{2} + e_l - s = \frac{r_l + e_l}{2} - s < 0\right),$$
$$\min(e_x, r_l - s) = r_l - s.$$

Thus,

$$L_x = r_l - s - 0 = r_l - s,$$

and symmetrically $L_y = r_l - s$. Hence, $L_x \cdot L_y = (r_l - s)^2$.
Non-Centered Positions Degrade $L_x \cdot L_y$. For any offset $\Delta \neq 0$, let $e_x = \frac{r_l - e_l}{2} + \Delta$. We then prove that the optimal $\Delta = 0$. Due to symmetry, we only analyze L_x:
Case 1: $e_l \le s \le \frac{r_l + e_l}{2}$
If $\Delta > 0$, the lower bound becomes $\max(e_x + e_l - s, 0) = \frac{r_l + e_l}{2} - s + \Delta$. However:

$$\min(e_x, r_l - s) = \min\left(\frac{r_l - e_l}{2} + \Delta, r_l - s\right) \le \frac{r_l - e_l}{2} + \Delta. \tag{7}$$

The valid interval $L_x \le \frac{r_l - e_l}{2} + \Delta - \left(\frac{r_l + e_l}{2} - s + \Delta\right) = s - e_l$. Thus, $L_x \cdot L_y < (s - e_l)^2$. Similar analysis holds for $\Delta < 0$.
Case 2: $\frac{r_l + e_l}{2} < s \le r_l$
For $\Delta > 0$:

$$\min(e_x, r_l - s) \le r_l - s,$$

with equality only when $\Delta = 0$. Thus, $L_x \cdot L_y \le (r_l - s)^2$, strictly smaller for $\Delta \neq 0$.

For all $s \in [e_l, r_l]$, $L_x \cdot L_y$ is maximized when $(e_x, e_y) = (\frac{s_x + r_l - e_l}{2}, \frac{s_y + r_l - e_l}{2})$ (centered trigger). Any deviation $\Delta \neq 0$ strictly reduces the valid area. This proves the optimality of the central position.

A.2. Optimality of the Locations of the Reference Image, Infected Shadow Image, and the Canvas Size

Let $p_1(s)$ denote the joint probability that a randomly cropped view $T_1(\hat{\mathbf{x}})$ contains the trigger $\mathbf{p}$ while remaining entirely within the infected shadow image $\mathbf{x}_\text{s} \oplus \mathbf{p}$. We decompose $p_1(s)$ into conditional probabilities to isolate the impact of trigger positioning:

$$p_1(s) = \underbrace{\Pr\left(\mathbf{p} \subseteq T_1(\hat{\mathbf{x}}) \,\middle|\, T_1(\hat{\mathbf{x}}) \subseteq \mathbf{x}_\text{s} \oplus \mathbf{p}\right)}_{q_1(s)} \cdot \underbrace{\Pr\left(T_1(\hat{\mathbf{x}}) \subseteq \mathbf{x}_\text{s} \oplus \mathbf{p}\right)}_{q_2(s)}. \tag{8}$$

Here, $q_1(s)$ represents the conditional probability of the trigger being fully contained in a cropped view, given that the crop lies within the infected shadow image. Critically, $q_1(s)$ depends solely on the relative position (e_x, e_y) of the trigger within $\mathbf{x}_\text{s} \oplus \mathbf{p}$, while $q_2(s)$ depends on the absolute position (s_x, s_y) of the shadow image within the canvas.

With trigger centering providing maximal $q_1(s)$ for all s, optimization now focuses on maximizing the remaining terms $\frac{1}{S-e_l}\int q_2(s)p_2(s)p_3(s)ds$. This reduces the original 8-dimensional parameter space $\{r_x, r_y, s_x, s_y, e_x, e_y, c_w, c_h\}$ to $\{r_x, r_y, e_x, e_y, c_w, c_h\}$.

Based on the above analysis, we now transition to connecting our optimization framework with established results. With $q_1(s)$ maximized by trigger centering, our objective reduces to optimizing $\frac{1}{S-e_l}\int q_2(s)p_2(s)\,ds$. The $p_3(s)$ term is temporarily omitted, as it can be optimized once the remains have reached their optima. Here, the constraint $T_1(\hat{\mathbf{x}}) \subseteq \mathbf{x}_\text{s} \oplus \mathbf{p}$ enforces that cropped regions lie entirely within the infected shadow image—a geometric condition formally equivalent to the trigger cropping constraint studied in [43]. Specifically, by treating $\mathbf{x}_\text{s} \oplus \mathbf{p}$ as all the possible trigger cropped region in their formulation, with (e_x, e_y) parameterizing its positional offset, our $q_2(s)p_2(s)$ becomes structurally identical to their probabilistic integral.

Lemma 1 (Theorem 1 in [43])**.** *Suppose left-right layout is used and $c_w \geq r_l, c_h \geq r_l$. $\left(r_x^*, r_y^*\right) = (0,0)$ is the optimal location of the reference image, and $\left(e_x^*, e_y^*\right) = \left(\frac{c_w+r_l-e_l}{2}, \frac{c_h-e_l}{2}\right)$ is the optimal location of the trigger.*

Lemma 2 (Theorem 2 in [43])**.** *Suppose left-right layout is used and and the optimal locations in Lemma 1 are used. For $c_w \geq r_l$, the optimal height of the canvas is $c_h^* = r_l$.*

A.3. Optimality of the Width of the Canvas

The above analysis reduces the parameter space to the canvas width c_w. We then proceed to express the optimization objective analytically as a function of c_w through IOU-based overlap modeling. Let g be the horizontal buffer width between the reference image $\mathbf{x}_\text{r}$ and infected shadow image $\mathbf{x}_\text{s} \oplus \mathbf{p}$, parameterizing the canvas width as $c_w = 2r_l + g$.

Parameterize $p_1(s;g)$ and $p_2(s;g)$ with Optimal Layout. Reference image is fixed at $(0,0)$, size $r_l \times r_l$. Infected shadow image is positioned at $(r_l+g, 0)$, size $r_l \times r_l$. Trigger is centered in $\mathbf{x}_s \oplus \mathbf{p}$: $e_x^* = r_l + g + \frac{r_l - e_l}{2}$. Canvas dimensions is $c_w = 2r_l + g$, $c_h = r_l$ because any extra area located right of the infected shadow image is redundant. Let $p_1(s;g)$ be probability that $\mathcal{V}_1$ contains the trigger and intersects with $\mathbf{x}_s \oplus \mathbf{p}$. From Theorem 1, the centered trigger maximizes containment. The valid region for $\mathcal{V}_1$ is:

$$p_1(s;g) = \frac{(s-e_l)^2}{(2r_l+g-s)(r_l-s)} \quad \text{for } e_l \leq s \leq \frac{r_l+e_l}{2},$$

$$p_1(s;g) = \frac{(r_l-s)^2}{(2r_l+g-s)(r_l-s)} \quad \text{for } \frac{r_l+e_l}{2} < s \leq r_l.$$

Valid horizontal range for $\mathcal{V}_2$: $0 \leq t_x^2 \leq r_l - s$. Total horizontal space: $c_w - s = 2r_l + g - s$.

$$p_2(s;g) = \frac{(r_l-s)(r_l-s)}{(2r_l+g-s)(r_l-s)} = \frac{r_l-s}{2r_l+g-s}.$$

Model $p_3(s;g)$ via IOU Overlap Probability. $p_3(s;g) = \Pr(\text{IOU}(\mathcal{V}_1, \mathcal{V}_2) \leq \tau)$, where τ is a small threshold (e.g., 0.05). Unlike p_1 and p_2 , p_3 allows the cropped region to be not entirely contained within the reference image or the infected shadow image. We explain the intuition behind our modeling in Section B. For left-right layouts, horizontal overlap dominates. Let $\Delta_x = \max(0, t_x^2 + s - t_x^1)$ be the horizontal gap. We approximate:

$$\text{IOU} \approx \frac{\Delta_x \cdot s}{2s^2 - \Delta_x \cdot s} \leq \tau \quad \Rightarrow \quad \Delta_x \leq \frac{2\tau s^2}{s+\tau s} = \frac{2\tau s}{1+\tau}.$$

Valid crpping regions are $\mathcal{V}_1$: $t_x^1 \in [r_l+g-s, r_l+g+r_l-s]$ and $\mathcal{V}_2$: $t_x^2 \in [0, r_l]$. The non-overlap condition is

$$0 \leq t_x^2 + s - t_x^1 \leq \Delta,$$

where $\Delta = \frac{2\tau s}{1+\tau}$. The overlap probability requires double integration over valid crop positions:

$$p_3(s;g) = \frac{1}{r_l^2}\int_{t_x^2=0}^{r_l} \int_{t_x^1=\max\left(r_l+g-s,\ t_x^2+s-\Delta\right)}^{\min\left(2r_l+g-s,\ t_x^2+s\right)} dt_x^1\, dt_x^2\, ds.$$

Let $A = r_l + g - s$ and $B = 2r_l + g - s$. The valid t_x^1 range becomes $[\max\left(r_l+g-s,\ t_x^2+s-\Delta\right), \min\left(2r_l + g - s,\ t_x^2+s\right)]$.

Non-overlap requires $t_x^2 + s - \Delta \leq r_l + A$ and $A \leq t_x^2 + s$. The valid width is:

$$\min(B, t_x^2+s) - \max(A, t_x^2+s-\Delta).$$

Subcases depend on t_x^2:

Case 1: $t_x^2 + s - \Delta \leq A$

Lower bound $= A$, upper bound $= \min(B, t_x^2+s)$. Though τ is small, $t_x^2 + s \leq A + \Delta = A + \frac{2\tau s}{1+\tau} \leq B$. Thus upper bound is $t_x^2 + s$.

$$p_3(s;g) = \frac{1}{r_l^2}\int \int_{t_x^2=\max(A-s,0)}^{\min(A-s+\Delta, r_l)} [(t_x^2+s) - A]\, dt_x^2 ds,$$

$$\overset{\lim_{\tau\to 0}\Delta=0}{=} \frac{1}{r_l^2}(s-A)\Delta \int_{A-s>0} ds$$

$$+ \quad \frac{\Delta}{2r_l^2}\int_{A-s>0}(2A-2s+\Delta)ds. \tag{9}$$

Case 2: $B \leq t_x^2 + s$

since τ is small, $t_x^2 + s - \Delta \geq B - \Delta \geq A$. Valid width

$= B - t_x^2 - s + \Delta.$

$$p_3(s;g) = \frac{1}{r_l^2}\int\int_{t_x^2=B-s}^{\min(B+\Delta-s,r_l)} [B - t_x^2 - s + \Delta] dt_x^2 ds,$$
$$= \frac{\Delta}{r_l^2}\int_{B-s<r_l}(B-s+\Delta)ds - \frac{\Delta}{r_l^2}\int_{B-s<r_l}(2B-2s+\Delta)ds,$$
$$= \frac{\Delta}{r_l^2}\int_{B-s<r_l}(B-s)ds = \frac{\Delta}{r_l^2}\int_{B-s<r_l}(2r_l+2g-2s)ds. \tag{10}$$

Case 3: $A + \Delta \le t_x^2 + s \le B$
Lower bound $= t_x^2 + s - \Delta$ and upper bound is $t_x^2 + s$. The width is Δ.

$$p_3(s;g) = \frac{1}{r_l^2}\int\int_{t_x^2=A+\Delta-s}^{\min(B-s,r_l)} [\Delta] dt_x^2 ds,$$
$$= \frac{\Delta}{r_l^2}\int_{B-s<r_l}(r_l-\Delta)ds + \frac{\Delta}{r_l^2}\int_{B-s>r_l}(\Delta-g)ds. \tag{11}$$

Integrating over all three cases, we have

$$p_3(s;g) \overset{\lim_{\tau\to 0}\Delta=0}{=\!=} + \frac{\Delta}{r_l^2}\int_{A-s>0}(r_l-2s+3\Delta/2)ds$$
$$+ \frac{\Delta}{r_l^2}\int_{B-s<r_l}(3r_l+2g-2s-\Delta)ds \tag{12}$$

Find the Optimal Width of the Joint Probability.

$$J(g) = \frac{1}{S-e_l}\int_{s=e_l}^{r_l} p_1(s;g)p_2(s;g)p_3(s;g)ds$$
$$= \frac{\Delta}{(S-e_l)r_l^2}\left[\int_{e_l}^{\frac{r_l+g}{2}} p_1p_2\cdot(r_l-2s+3\Delta/2)\,ds\right],$$
$$+ \int_{\frac{r_l+g}{2}}^{r_l} p_1p_2\,(3r_l+2g-2s-\Delta)\,ds\Bigg] \tag{13}$$

Table 7. Clean performance on 10% clean available subset.

Dataset	MoCo v2		BYOL		SimSiam	
	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR10	69.0%	8.0%	88.3%	8.0%	71.1%	9.1%
ImageNet-100	66.5%	0.9%	80.1%	2.2%	66.1%	1.2%

WLOG, assume $g < e_l$ (conclusion holds for $g \ge e_l$):

$$J(g) = \frac{1}{S-e_l}\int_{s=e_l}^{r_l} p_1(s;g)p_2(s;g)p_3(s;g)ds$$
$$= \frac{\Delta}{(S-e_l)r_l^2}$$
$$[\underbrace{\int_{e_l}^{\frac{r_l+g}{2}} \frac{(s-e_l)^2}{(2r_l+g-s)^2}\cdot(r_l-2s+3\Delta/2)\,ds}_{J_1(g)}$$
$$+ \underbrace{\int_{\frac{r_l+g}{2}}^{\frac{r_l+e_l}{2}} \frac{(s-e_l)^2}{(2r_l+g-s)^2}\cdot(3r_l+2g-2s-\Delta)\,ds}_{J_2(g)}$$
$$+ \underbrace{\int_{\frac{r_l+e_l}{2}}^{r_l} \frac{(r_l-s)^2}{(2r_l+g-s)^2}\cdot(3r_l+2g-2s-\Delta)\,ds}_{J_3(g)}] \tag{14}$$

Using Leibniz Rule for Differentiation Under the Integral Sign, we can easily find $\frac{\partial J_1(g)}{\partial g} < 0$.Besides, the derivatives of the internal integral term of $J_2(g)$ is equal to

$$-2\frac{(s-e_l)^2(3r_l+2g-2s)}{(2r_l+g-s)^3} + \frac{2(s-e_l)^2}{(2r_l+g-s)^2},$$
$$= -2\frac{(s-e_l)^2(r_l+g-s)}{(2r_l+g-s)^3} < 0.$$

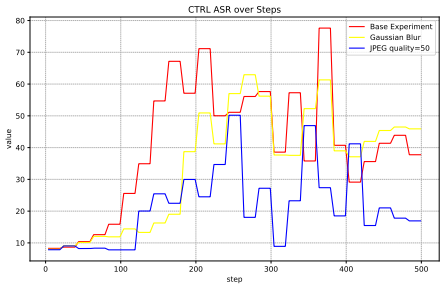
Again, with Leibniz Rule for Differentiation Under the Integral Sign, we can find $\frac{\partial J_2(g)}{\partial g} < 0$, similarly for $J_3(g)$. The optimal canvas configuration achieves maximal joint probability when images are adjacent with zero gap:

$$\boxed{g = 0}.$$

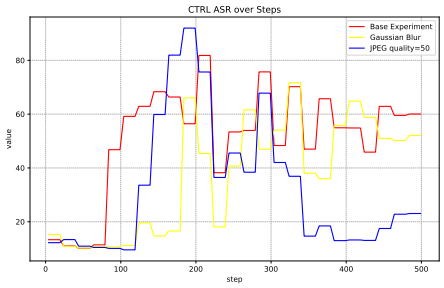
This corresponds to minimum canvas width $2r_l$ with tight image adjacency.

B. The Information Theory Perspective of Our Attack

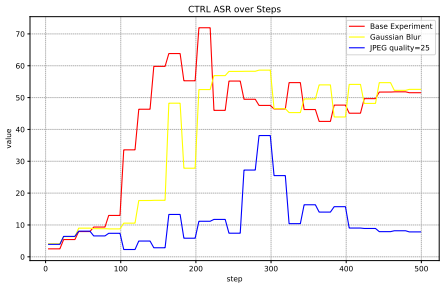
Given a pair of random variables v_1 and v_2, contrastive learning aims to train a parameterized function f_θ that maps inputs from sample $x \in \mathcal{X}$ into a representation space $\mathbb{R}^d$.



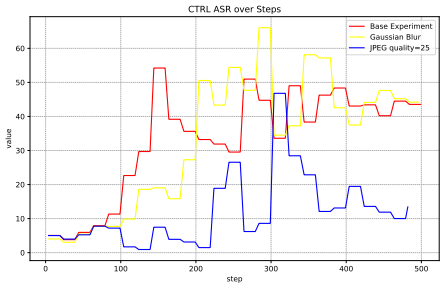
(a) airplane



(b) automobile

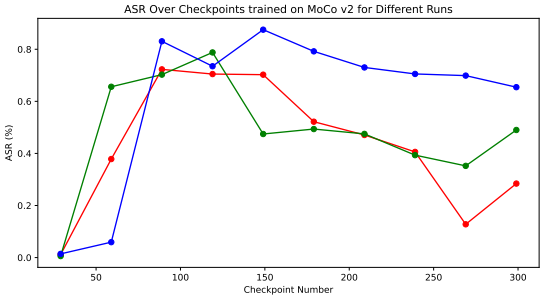


(c) bird



(d) cat

Figure 10. Different attack classes of CTRL [21] on CIFAR-10 under various data processing methods. We use Gaussian noise and JPEG compression to perturb the poisons.



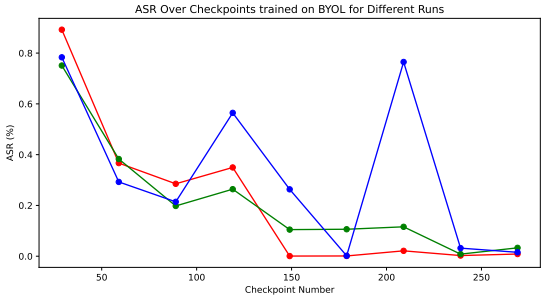


Figure 11. ASR over checkpoints of three CorruptEncoder [43] trials on ImageNet-100.

The objective is to distinguish between positive pairs sampled from the joint distribution $p(v_1|x)p(v_2|x)$ and negative pairs drawn independently from the marginal distributions $p(v_1)p(v_2)$. The reuslting function f is a mutual information estimator between v_1 and v_2 [36, 37]. Tipically, minimizing InfoNCE loss [15, 37] equivalently maximizes a lower bound of $I(v_1; v_2)$. Note that views v_1 and v_2 are obtained from samples through data augmentation.

[36] points out that the optimal views are related to the downstream task (denoted as T). Ideally, the mutual information between augmented views should contain only the information relevant to the downstream task, i.e., $I(v_1, v_2; T) = I(v_1, T) = I(v_2, T)$. Inspired by this viewpoint, we hope that the views generated by random cropping contain the backdoor trigger and the reference image, respectively.

Optimal Layout under the Information Theory Perspective. Given the optimal views, we need to design the layout to maximize the probability of its occurrence. Let $S(v)$ denote the set of pixels in the view v. We can categorize the information sharing between the views v_1 and v_2 into different scenarios:

1. *Missing information*: $S(v) \cap S(\mathbf{p}) = \emptyset \wedge S(v) \cap S(\mathbf{x}_\mathrm{r}) = \emptyset, \forall v \in \{v_1, v_2\}$. This is irrelevant to the attack and could degrade the efficiency of the attack.
2. *Sweet spot*: $S(\mathbf{p}) \subseteq S(v_1) \wedge S(v_1) \cap S(\mathbf{x}_\mathrm{r}) = \emptyset \wedge S(v_2) \subseteq S(\mathbf{x}_\mathrm{r})$. The only information shared between v_1 and v_2 is not more than the trigger p and reference

Table 8. Performance of irregular and invisible triggers.

Method	ACC (%)	ASR (%)
baseline	66.1	82.3
+Blended triggers [9]	65.8	88.2

patterns, i.e., $I(v_1; v_2) \leq I(p; v_2)$.

3. *Information leak*: $S(v) \cap S(\mathbf{p}) \neq \emptyset \wedge S(v) \cap S(\mathbf{x}_r) \neq \emptyset, \forall v \in \{v_1, v_2\}$. This leads to $I(v_1; v_2) > I(p; v_2)$ and $I(v_1; v_2) > I(p; v_1)$, which could harm the attack. Information other than the attacks shared by v_1 and v_2 may become a shortcut for model learning, thus neglecting beneficial information from the attacks.

C. Experimental Details

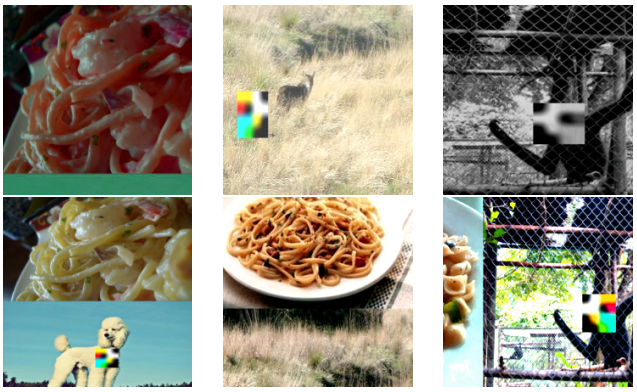
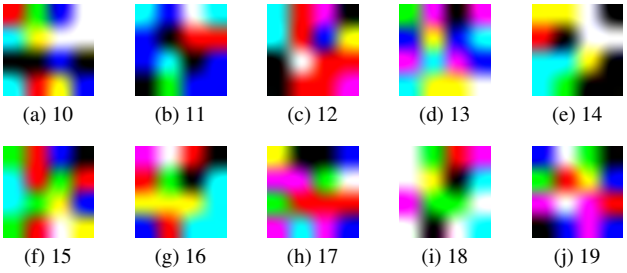


Figure 12. **Augmented views of the poisoned data.** Each of the top row and the bottom row is one of the augmented views from the identical poisoned image of MoCo v2 [10] and the target attack class is carbonara.



(a) 10 (b) 11 (c) 12 (d) 13 (e) 14

(f) 15 (g) 16 (h) 17 (i) 18 (j) 19

Figure 13. Illustration of the patch triggers.

Trigger. We mainly use the trigger from [30], which are small square colorful patches, i.e. random 4×4 RGB images, as Figure 13 shows. They are resized to the desired size when attached to the poisoned image. We demonstrate augmented views on ImageNet-100 in Figure 12. We also use non-patch-based triggers to test our attacks, as shown in Table 8.

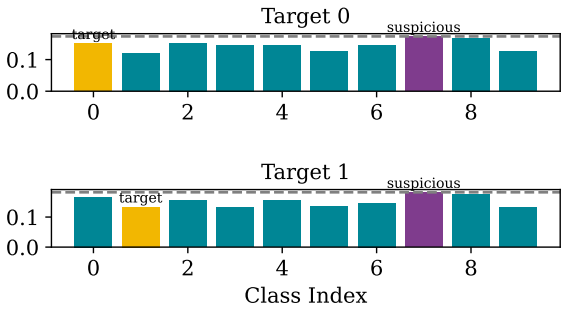


Figure 14. Activation Cluster defense.

Hyperparameters. We synchronize the hyperparameters with the baseline SSLBKD [30], ensuring the comparability. Note that we slightly scale the training length to 300 epochs, as SSL methods typically require longer to converge. We provide the pre-training configurations and linear probing configurations in Table 14 and Table 9 respectively.

D. More Analysis of Attack Dynamics

Decline in attack performance during the late training stage. In Figure 11 we plot the ASR trajectory of CorruptEncoder on ImageNet - 100. The attack converges swiftly, attaining 60–80% ASR within the first 50–100 epochs for both MoCo v2 and BYOL. Training beyond this point, however, often causes the ASR to degrade. We conjecture that the Uniformity regularization in later epochs [40] loosens the coupling between the backdoor and its reference image, echoing the observations of Sun *et al.* [33]. A comparable trend is also visible in CTRL [21] (Figure 10), underscoring the generality of this phenomenon.

Representation Visualization. Figure 16 shows intermediate t - SNE snapshots, while Figure 15 depicts the representation space at convergence. Figure 16 shows that our attack can maintain the separability of poison representations in the later stages of training.

Reference Distribution Shift. Table 13 investigates the attack effectiveness under a distribution mismatch between the pre-training and downstream. ImageNet-100-O is an alternative subset that is disjoint from ImageNet-100. Such a shift hampers both benign performance and attack strength, since feature representations become sub-optimal for the new domain. Nevertheless, NA still delivers competitive at-

Methods	MoCo v2 & SimSiam & SimCLR	BYOL
Training Epochs	40	100
Batch Size	256	256
Optimizer	SGD	Adam
Learning Rate Schedule	MultiStepLR	ExponentialLR
Learning Rate	0.01	0.01
Weight Decay	1×10^{-4}	5×10^{-6}
Momentum	0.9	-
Resize & Crop	RandomResizeAndCrop	RandomResizeAndCrop
RandomHorizontalFlip	0.5	0.5

Table 9. Hyperparameters for linear probing.

Table 10. ASR of directly poisoning CLIP with different image-modal poisons.

Metrics	**SSLBKD**	**SIG**	**Gaussian noise**	**NA**
Top1	99.9%	59.3%	99.8%	91.3%
Top5	99.9%	63.3%	99.9%	96.0%

Table 11. SCAn results on CIFAR10 and ImageNet-100.

Dataset	**CIFAR10**		**ImageNet-100**	
	MoCo v2	**SimSiam**	**MoCo v2**	**SimSiam**
CAP	100%	100%	0%	0%
TPR	11.5%	28.7%	26.2%	3.7%
FPR	0.0%	0.1%	3.0%	4.9%

Table 12. PatchSearch defense.

Metric	**MoCo**	**SimSiam**
Poisons Removed	38,710	28,666
Recall (%)	46.3	49.1
Precision (%)	0.8	1.2
ASR after defense (%)	61.0	77.1

Table 13. ASR on difference reference distributions.

Pre-training Dataset	**Reference Dataset**	**Model**	**Results**	
			CA	**ASR**
ImageNet-100	ImageNet-100-O	MoCo v2	61.1%	77.1%
		SimSiam	54.7%	84.3%
	STL-10	MoCo v2	70.2%	59.0%
		SimSiam	70.5%	52.1%
	CIFAR-10	MoCo v2	52.2%	42.9%
		SimSiam	53.5%	49.8%

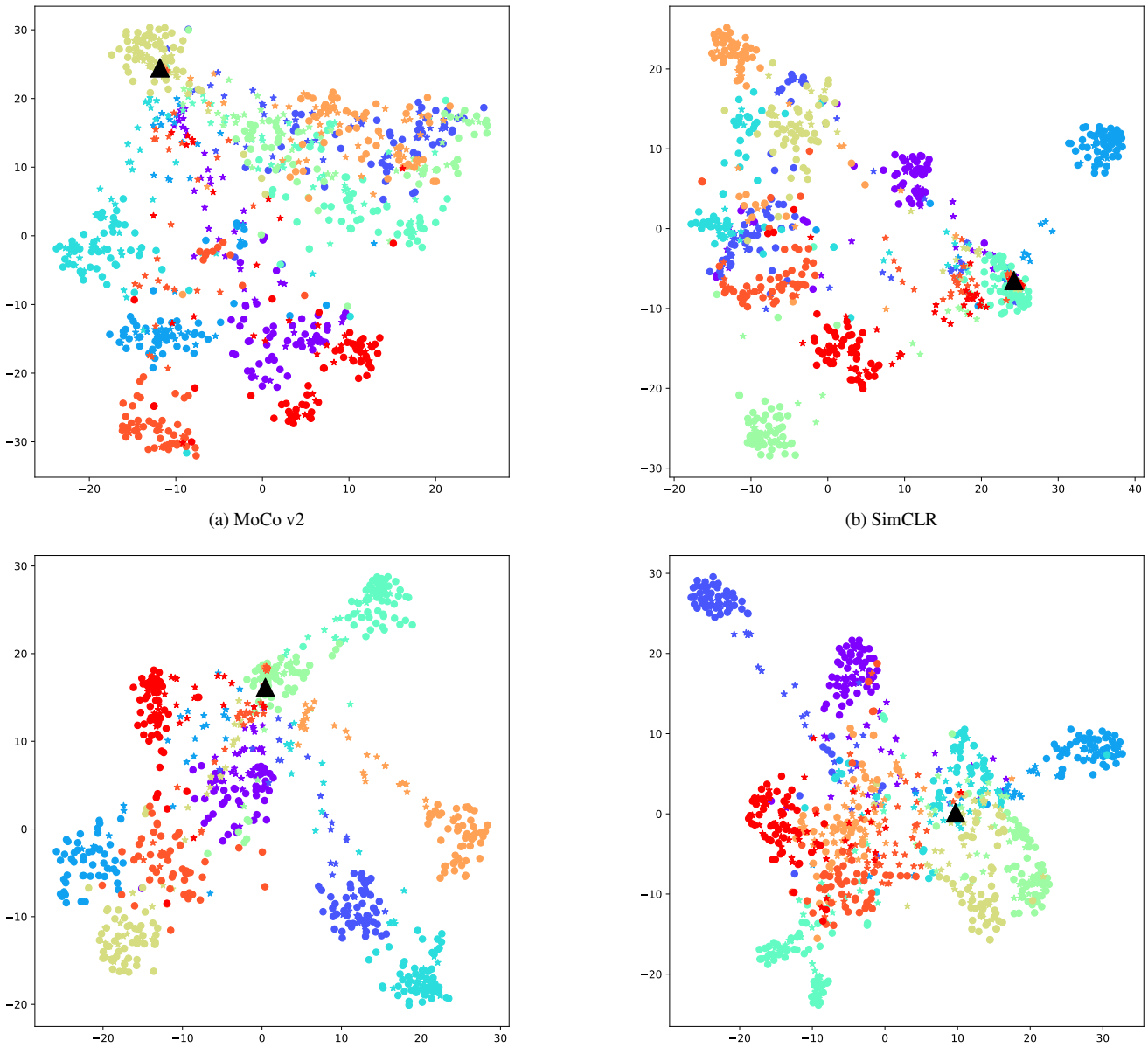
tack efficacy, demonstrating that it can effectively generalize beyond the original pre-training distribution.

E. More Defenses

PatchSearch. *PatchSearch* [35] is a poison detection method design for SSL. Table 12 shows PatchSearch retrieves about half of the poisons, but the ASR remains high (61.0% for MoCo v2 and 77.1% for SimSiam).

Statistical Contamination Analyzer (SCAn) . We evaluated the *SCAn* using three metrics: accuracy of the poisoned class prediction (CAP), false positive rate (FPR), and true positive rate (TPR). We implemented SCAn on CIFAR10 following [23] and randomly sampled 10% of the test set to build the decomposition model. Table 11 shows SCAn can effectively identify the poisoned class on CIFAR10, yet it is entirely out of work on the larger ImageNet-100.

Activation Clustering (AC) . The *AC* [6] detection is based on the intuition that poisoned examples are likely to be a distinct cluster in the representation space. In Figure 14, we report the silhouette scores of feature clusters on CIFAR10. AC fails to accurately detect the corresponding attack class, as indicated by lower silhouette scores compared to other unpoisoned categories.



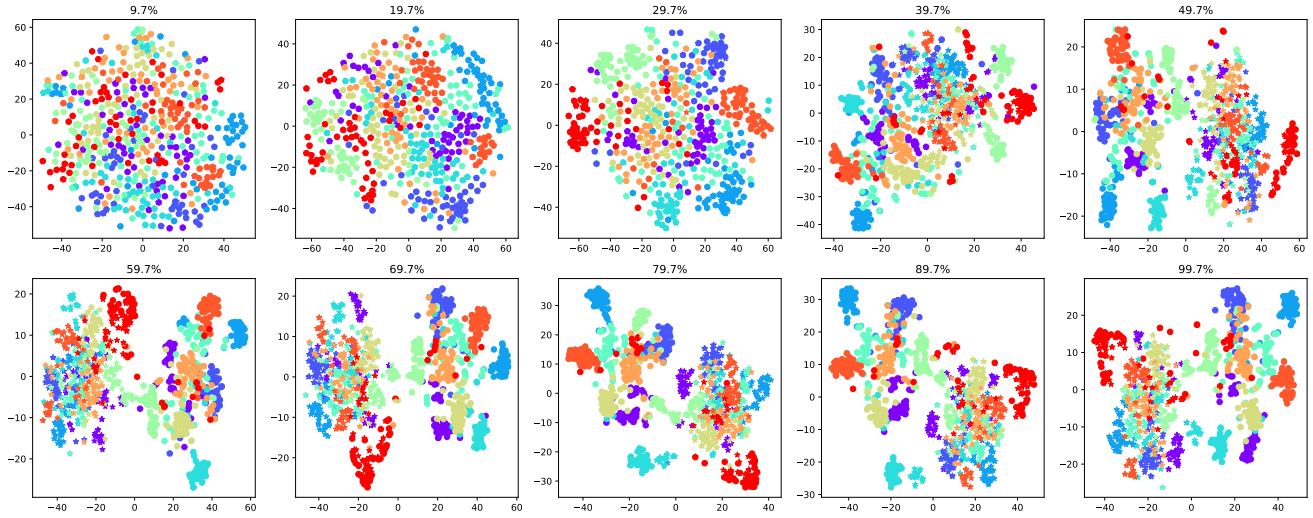
(a) MoCo v2

(b) SimCLR

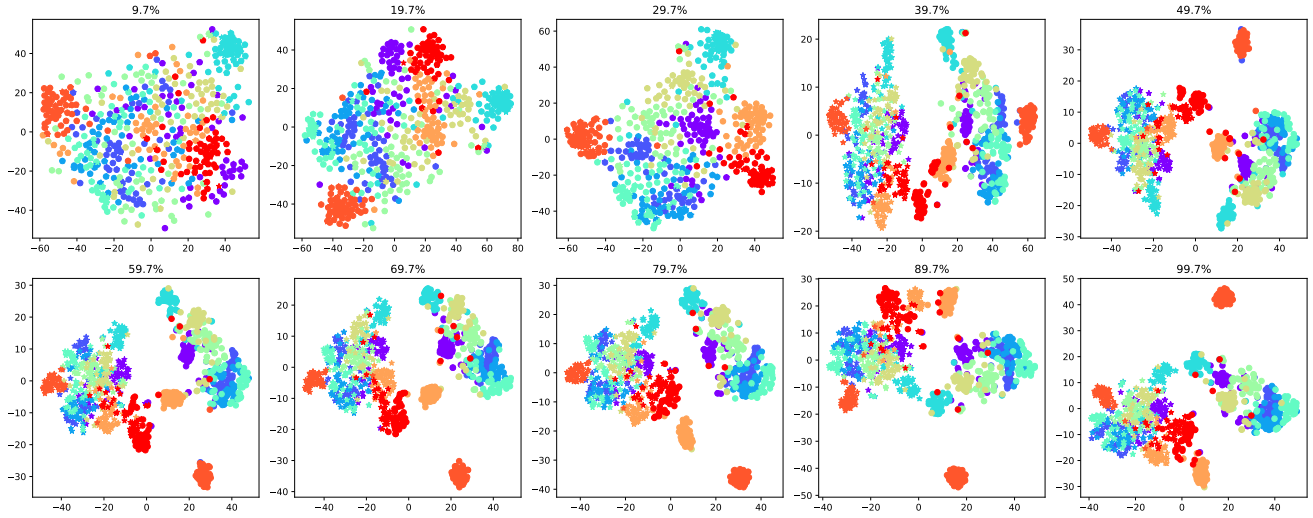
(c) BYOL

(d) SimSiam

Figure 15. t-SNE visualization of the representation space of our attack. Black triangles ▲ are poison centers and colors represent different classes. Star and circle markers represent the poisoned and clean samples, respectively.



(a) Attack category n03085013.



(b) Attack category n03947888.

Figure 16. t-SNE visualization at various training stages on ImageNet-100. Circles represent clean samples, while stars denote poisons. Different classes are distinguished by color.

Methods	MoCo v2	BYOL	SimSiam
Training Epochs	300	300	300
Batch Size	512	512	512
Optimizer	SGD	Adam	SGD
Learning Rate Schedule	Cosine	Cosine	Cosine
Learning Rate	0.06	0.002	0.05
Weight Decay	1×10^{-4}	1×10^{-6}	1×10^{-4}
Moving Average	0.999	0.99	-
Resize & Crop	RandomResizeAndCrop	RandomResizeAndCrop	RandomResizeAndCrop
Color Jitter	0.4	0.4	0.4
RandomHorizontalFlip	0.5	0.5	0.5
Min Crop Scale	0.2	0.2	0.2
RandomGrayscale	0.2	0.1	0.2
GaussianBlur(p=0.5)	[.1, 2.]	[.1, 2.]	[.1, 2.]

Table 14. Hyperparameters for pre-training.