Pre-training CLIP against Data Poisoning with Optimal Transport-based Matching and Alignment

Anonymous ACL submission

Abstract

Recent studies have shown that Contrastive Language-Image Pre-training (CLIP) models are threatened by targeted data poisoning and backdoor attacks due to massive training image-caption pairs crawled from the Internet. Previous defense methods correct poisoned image-caption pairs by matching a new caption for each image. However, the matching process solely relies on the global representations of images and captions, overlooking fine-grained features of visual and textual features. It may introduce incorrect image-caption pairs and detriment the CLIP pre-training. To address their limitations, we propose an Optimal Transport-based framework to reconstruct the image-caption pairs, named OTCCLIP. We involve a new optimal transport-based distance measure between fine-grained visual and textual feature sets and re-assign new captions based on the proposed optimal transport distance. Additionally, to further reduce the negative impact of mismatched pairs, we encourage the inter- and intra-modality fine-grained alignment by employing optimal transport-based objective functions. Our experiments demonstrate that OTCCLIP can successfully decrease the attack success rates of poisoning attacks. Also, compared to previous methods, OTCCLIP significantly improves CLIP's zero-shot and linear probing performance trained on poisoned datasets.

1 Introduction

Contrastive Language-Image Pre-training (CLIP) models have demonstrated remarkable zero-shot performance across diverse domains, leveraging millions or billions of training samples from the Internet (Radford et al., 2021; Jia et al., 2021). As CLIP's large-scale pre-training data is often crawled online, attackers can inject malicious examples into the training set to alter predictions at test time. Recent research shows that inserting only a small number of poisoned samples can cause CLIP to misclassify specific inputs during inference (Carlini et al., 2024; Carlini and Terzis, 2021; Liang et al., 2024; Bai et al., 2024). Alarmingly, poisoning just 0.0001% and 0.001% of the data can enable data poisoning and backdoor attacks, respectively (Carlini et al., 2024).

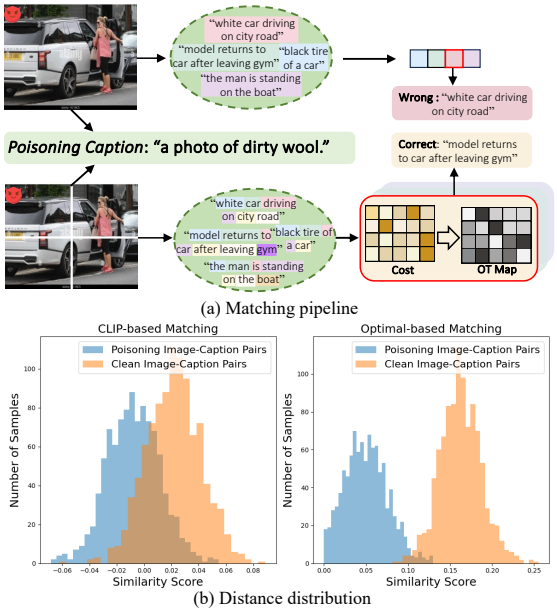

Figure 1: (a) Previous methods use global features of the CLIP model, while we employ a fine-grained optimal transport method. (b) Compared to CLIP-based matching, optimal transport-based fine-grained matching is robust for distinguishing poisoned data.

An effective defense method is crucial to mitigate the impact of poisoned image-caption pairs during pre-training. RoCLIP (Yang and Mirzasoleiman, 2023) disrupts the malicious link between poisoned images and captions by matching each image representation with its most similar caption from a random pool. SAFECLIP (Yang et al., 2024) avoids misleading information by employing cross-modal alignment only on identified clean datasets. The ability to distinguish poisoned data is highly dependent on the matching method used to identify or correct the data while the model is not yet fully trained. As shown in Figure 1, pre-

vious methods using CLIP-based semantic matching to differentiate poisoned image-caption pairs face challenges. This is because global features focus on overall semantics, which means that subtle yet indicate poisoning or inconsistencies within image-caption pairs are likely to be missed. Hence, identification and correction result in a suboptimal solution, which ultimately causes the model to overfit to poisoned data.

In this work, we leverage fine-grained features to address the limitations mentioned above and enhance the model's generalization capability. To achieve this, we introduce the optimal transport framework OTCCLIP, designed to disrupt the association between poisoning image-caption pairs during pre-training. We consider the fine-grained feature similarity measure as an optimal transportation problem to reconstruct and align the image-caption pairs, which aims to transport a collection of contextual patches in an image to the ones in another contextualized token sequence in a caption. OTCCLIP first employs optimal transport-based matching, using the transport matrix as weights to effectively capture relationships across different regions of image patches and caption tokens. This approach improves the ability of the model to distinguish poisoned data, as shown in Figure 1.

However, it is challenging to correct all poisoned data solely through optimal transport-based matching. Therefore, we propose to encourage fine-grained inter-modality alignment to further enhance resilience for poisoned data. OTCCLIP treats the alignment of images and captions obtained from optimal transport matching as a distribution transportation optimization task to better associate image patches and caption tokens. Optimal transport assigns greater weights in highly similar regions of image-caption pairs and smaller weights to less similar regions, which reduces the risk of introducing errors from unmatched pairs during pre-training. In addition, the intrinsic relationships within each modality are crucial and, more importantly, are not affected by cross-modal poisoning. Hence, we separately employ the intra-modality fine-grained alignment for image patches and caption tokens, which is also helpful against data poisoning.

We conduct extensive experiments on multiple image-caption datasets, showing that OTCCLIP effectively reduces attack success rates to 0% in most cases. Additionally, we observe improvements in CLIP's zero-shot and linear probing performance.

2 Related Work

2.1 Protecting CLIP Against TDPA and BA

CLIP is vulnerable to targeted data poisoning attacks (TDPAs) and backdoor attacks (BAs) (Carlini and Terzis, 2021; Yang et al., 2023). TDPAs manipulate a small portion of training data to mislead model into misclassifying specific examples, while BAs embed visible or invisible triggers (e.g., noise or deformations) induce misclassification of test images containing the same trigger (Chen et al., 2017; Gu et al., 2017; Nguyen and Tran, 2021).

Effective defense methods have been proposed recently, which can be divided into four, including against backdoor/poisoning pre-training (Yang and Mirzasoleiman, 2023; Yang et al., 2024), fine-tuning the backdoored CLIP (Bansal et al., 2023; Kuang et al., 2024; Xun et al., 2024), using trigger inversion (Sur et al., 2023; Feng et al., 2023), and backdoor detection (Niu et al., 2024; Huang et al., 2025). Remarkably, research has shown that adding a trigger to just 0.01% of pre-training data can cause misclassification (Bansal et al., 2023), while TDPAs are even more effective, requiring only 0.0001% poisoned data (Yang and Mirzasoleiman, 2023; Yang et al., 2024). Current defenses for CLIP remain limited against these attacks.

RoCLIP (Yang and Mirzasoleiman, 2023) against data poisoning and backdoor attacks by augmenting image-caption pairs and matching them with nearest-neighbor captions from a pool in the pre-training. However, it overlooks local features and relies solely on global semantics, which can introduce matching errors and degrade performance. SAFECLIP (Yang et al., 2024) avoid involving the misleading information by employing cross-modal alignment on clean datasets. SAFECLIP first distinguish safe from risky data pairs by overall semantic features between image and caption datasets. SAFECLIP only apply cross-modal alignment on clean samples, harming the model's performance. For example, with a poisoning rate of 0.5%, more than 70% of clean data is classified into the harmful dataset, solely by applying self-modal feature alignment, which harms the model's performance.

2.2 Vision-Language Feature Alignment

Fine-grained feature alignment is key to providing accurate supervision and improving model performance. FILIP (Yao et al., 2021) achieves this via token-wise maximum similarity between visual and textual tokens. Other methods, such as OSCAR (Li

et al., 2020), VinVL (Zhang et al., 2021), MVPTR (Li et al., 2022), and X-VLM (Zeng et al., 2021), focus on multi-level semantic alignment. OSCAR introduces multi-level semantics by capturing object region features and tags, while VinVL refines visual features with an improved object-attribute detector. MVPTR and X-VLM extend multi-level semantics across both visual and textual modalities, with MVPTR modeling object-tag alignment and phrase structure, and X-VLM aligning visual concepts with textual descriptions. PyramidCLIP (Gao et al., 2022) combines three visual and three linguistic representations to compute multiple contrastive loss terms, supporting multi-level alignment. Collectively, these approaches show that fine-grained features enhance image-caption alignment and boost resilience to perturbations.

3 Preliminary

3.1 Contrastive Language-Image Pre-training (CLIP)

Typically, CLIP employs two main components: an image encoder E_I and a text encoder E_T. Given a dataset $\mathcal{D}$ consisting of image-caption pairs $(\mathbf{X}_i, \mathbf{Y}_i)$, where $\mathbf{X}_i$ represents the image, and $\mathbf{Y}_i$ represents the corresponding caption. When the image $\mathbf{X}_i$ is input into the image encoder E_I, it is first transformed into spatial feature representations $f_i^s \in \mathbb{R}^{h\times w\times d}$, then condensed into a global feature vector $f_i^g \in \mathbb{R}^d$. These spatial features can be represented as $f_i^s = \{z_{i,1}^s, z_{i,2}^s, \cdots, z_{i,h\times w}^s\}$, where each $z_{i,j}^s \in \mathbb{R}^d$ (for $j = 1, 2, \ldots, h \times w$) is a feature vector corresponding to a spatial location in the image. The spatial features are then condensed into a global feature vector $f_i^g \in \mathbb{R}^d$. Here, h and w denote the height and width of the feature map, while d represents the dimensionality of each feature at a given spatial location. Similarly, the text $\mathbf{Y}_i$ is encoded into the text encoder E_T to produce token sequence features $y_i^s \in \mathbb{R}^{l\times d}$, which are further aggregated into a global feature $y_i^g \in \mathbb{R}^d$. These token sequence features are represented as $y_i^s = \{\hat{z}_{i,1}^s, \hat{z}_{i,2}^s, \cdots, \hat{z}_{i,l}^s\}$, where each $\hat{z}_{i,j}^s \in \mathbb{R}^d$ (for $j = 1, 2, \ldots, l$) is a token vector corresponding to a position in the caption. Here, l denotes the length of the token sequence feature, while d represents the dimensionality. To enable multi-modal interaction, CLIP employs the InfoNCE loss during training. This loss function encourages the alignment of representations from each image-caption pair while separating those of non-paired images and captions within the same mini-batch. The quality of the learned representations is assessed using zero-shot and linear probe classification; details of these evaluation protocols are provided in Section A.1.

3.2 Threat Model

Adversary capabilities. Recent research (Yang and Mirzasoleiman, 2023; Yang et al., 2024; Bai et al., 2024; Liang et al., 2024) has revealed the serious backdoor vulnerability of CLIP. We adopt the poisoning-based threat model from previous works (Yang and Mirzasoleiman, 2023; Yang et al., 2024), where the adversary injects a set of poisoning image-caption pairs into the pre-training data. In this scenario, attackers can only manipulate poisoned data, unlike other works (Bai et al., 2024; Liang et al., 2024), which assume attackers modify the training process. Let $D_{poi} = (\mathbf{X}_i, \mathbf{Y}_{poi(i)})|\mathbf{X}_i \in \mathcal{I}_i, \mathbf{Y}_{poi(i)} \in \mathcal{T}adv$ denote the injected poisoning pairs, where $\mathcal{D}_{poi} \subset \mathcal{D}$. Here, $\mathcal{T}_{adv}$ is the set of adversarial captions related to the adversarial label $\mathbf{Y}_{adv}$. There are two ways to generate adversarial captions. On one hand, the adversary can construct an adversarial caption by searching for some captions containing the adversarial label. Alternatively, the adversarial can utilize CLIP's 80 prompt-engineered text descriptions (Radford et al., 2021; Yang and Mirzasoleiman, 2023; Zhou et al., 2022) to generate captions for the adversarial label. Besides, the adversaries have knowledge of the model's architecture, the training algorithm, and the hyperparameters but cannot directly the alter training process.

Adversary objective. Targeted data poisoning attacks aim to misclassify a particular test example, $\mathbf{X}_i$, as $\mathbf{Y}_{adv}$. Hence, $D_{poi} = \{(\mathbf{X}_i, \mathbf{Y}_{poi(i)})|\mathbf{Y}_{poi(i)} \in \mathcal{T}_{adv}\}$. Backdoor attacks introduce a trigger patch to a set of poisoned images. The goal is to misclassify any test examples with the trigger patch, $\mathbf{X}_i \oplus$ patch, as $\mathbf{Y}_{adv}$. Hence, $D_{poi} = \{(\mathbf{X}_i \oplus \text{patch}, \mathbf{Y}_{poi(i)})|\mathbf{X}_i \in \mathcal{I}, \mathbf{Y}_{poi(i)} \in \mathcal{T}_{adv}\}$. In contrast to targeted data poisoning attacks, which target a particular test example, backdoor attacks inject *random* images with the backdoor trigger, paired with the adversarial captions.

4 Method

In this section, we first introduce the foundational concepts of optimal transport and describe how the fine-grained matching problem can be modeled in an optimal transport framework. Next, we explain

the fine-grained alignment module and provide the implementation details for training and inference.

4.1 The Definition Of Optimal Transport

Defining Source And Target Distributions. First, we define two pivotal distributions within the optimal transport framework (Pramanick et al., 2023; Chang et al., 2022): the source distribution $\mathbf{K} = (k_1, k_2, \cdots, k_n)$ and the target distribution $\mathbf{Q} = (q_1, q_2, \cdots, q_m)$. These distributions correspond to the starting and ending points of the transportation process.

Transportation matrix T. The transportation plan is described by a matrix $\mathbf{T} = [T_{uv}]$ of size $n \times m$. Each element T_{uv} represents the amount of resource transported from the u-th source in $\mathbf{P}$ to v-th target in $\mathbf{Q}$. This matrix outlines the optimal transportation strategy, aligning the two distributions while minimizing total cost (Chen et al., 2020).

In the optimal transport framework, the matrix $\mathbf{T}$ must meet specific constraints to ensure an effective transportation plan (Chen et al., 2020; Li et al., 2024). The Marginal Constraints are given by $\sum_{v=1}^{m} T_{uv} = k_u$ for $u = \{1, \ldots, n\}$ and $\sum_{u=1}^{n} T_{uv} = q_v$ for $v = \{1, \cdots, m\}$. These constraints require that the total transported amount from each source u and to each target v matches the respective supply k_u and demand q_v. The Non-Negativity Constraint is $T_{uv} \geq 0$ for all u and v, ensuring all transport amounts T_{uv} are non-negative, which reflects the practical impossibility of negative transportation.

Modeling the optimal transport problem. With the aforementioned definitions and constraints established, the Optimal Transport problem can be formulated as follows:

$$OT(\mathbf{K}, \mathbf{Q}, \mathbf{C}) = \min_{\mathbf{T} \in \Pi(\mathbf{K}, \mathbf{Q})} \sum_{u=1}^{n} \sum_{v=1}^{m} T_{uv} \cdot C_{uv}, \quad (1)$$

where C denotes the cost matrix, with each element C_{uv} representing the cost of transporting a unit from source k_u to target q_v. The matrix $\mathbf{T}$ signifies the transportation scheme, while $\Pi(\mathbf{K}, \mathbf{Q})$ encompasses all feasible transportation schemes that satisfy the marginal constraints.

To handle high-dimensional spaces effectively, the Sinkhorn distance is used in Optimal Transport (OT) (Distances, 2013). Traditional OT methods, which rely on linear programming, struggle with computational demands and scalability issues. In contrast, the Sinkhorn distance incorporates entropy regularization into the OT calculation, improving both tractability and differentiability. Consequently, the Sinkhorn Optimization Process can be defined as:

$$M(\mathbf{K}, \mathbf{Q}, \mathbf{C}) = \min_{\mathbf{T} \in \Pi(\mathbf{K}, \mathbf{Q})} \sum_{u=1}^{n} \sum_{v=1}^{m} T_{uv} \cdot C_{uv} + \lambda H(\mathbf{T}), \quad (2)$$

where $H(\mathbf{T})$ is the entropy of the transport matrix, which introduces regularization to ensure numerical stability and efficient computation, and λ is a hyper-parameter that balances accuracy and computational efficiency. Higher λ values yield results closer to traditional OT but increase computational costs, while lower values of λ speed up calculations at the cost of some bias. The Sinkhorn algorithm iteratively normalizes the rows and columns of the transport matrix to satisfy the marginal constraints while minimizing the regularized objective function (Distances, 2013).

4.2 Optimal Transport-based Matching

Previous methods (Yang and Mirzasoleiman, 2023; Yang et al., 2024) use the global feature to identify the poisoning data. However, the global features tend to emphasize only the most prominent or frequent characteristics in the data, primarily capturing dominant semantic information while overlooking finer details. The global feature focus on overall semantics means that subtle yet important cues, especially those that may indicate poisoning or inconsistencies within image-caption pairs, are likely to be missed. To address this issue, we employ optimal transport into fine-grained matching between images and captions. Given an image with spatial features f_i^s, our aim is to find the most matching caption from a randomly sampled pool of captions with fine-grained features $\mathcal{P}^s = \{y_{p(i)}^s\}_{i=1}^{P}$. Given the definition of optimal transport, we define the fine-grained feature set $f_i^s = \{z_{i,1}^s, z_{i,2}^s, \cdots, z_{i,h\times w}^s\}$ as a distribution of patch-level features $\boldsymbol{\mathcal{G}_f}$. Similarly, we define the set of token sequence features $y_{p(j)}^s = \{\hat{z}_{j,1}^s, \hat{z}_{j,2}^s, \cdots, \hat{z}_{j,l}^s\}$ in the caption pool as the distribution of token-level features $\boldsymbol{\mathcal{G}_p}$.

To perform the fine-grained matching, we first compute a similarity matrix $S^P = f_i^s \odot y_{p(j)}^s$ between image patches and caption tokens. Here, $\odot$ represents the Hadamard product and $S^P \in \mathbb{R}^{hw \times l}$. Each position in the similarity matrix focuses only on local features between image patches and caption tokens. Therefore, the similarity matrix cannot effectively represent the global matching degree between the image and caption. In the optimal trans-

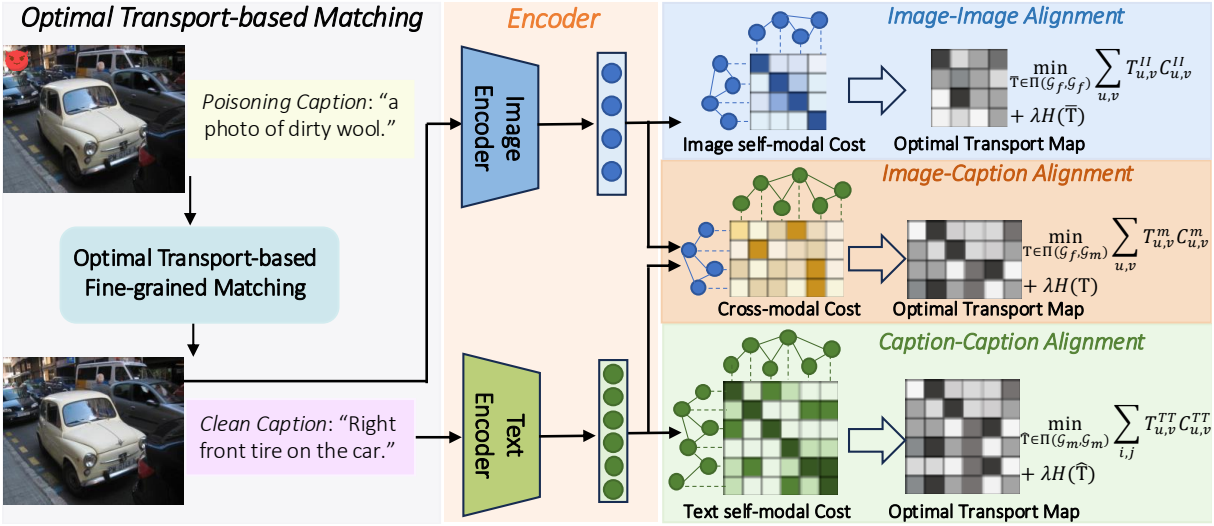

Figure 2: Illustration of OTCCLIP for defending CLIP during pre-training. Given image-caption pairs, OTCCLIP first applies optimal transport matching to break the association between poisoning images and captions, reconstructing new image-caption pairs. These reconstructed pairs are then fed into the optimal transport-based inter-modality module to better align fine-grained features and reduce the negative impact of mismatches. Reconstructed pairs also are fed into the optimal transport-based intra-modality alignment module to capture the intrinsic relationships of each modality. Additionally, reconstructed data use CLIP's InfoNCE loss to achieve alignment of semantics.

port, the overall matching cost $\sum_{u,v} T_{uv} C^P_{uv}$ is calculated by the product sum of the transportation matrix T and the cost matrix C, the cost matrix is defined as $C^P = 1 - S^P$. By optimizing the transport plan, the transport matrix determines how to match image patches and caption tokens at the minimum cost. Therefore, optimal transport can measure the degree of overall matching between the image patches and the caption tokens from a global perspective. Then, the overall matching score M between a given image and any caption sequence in the pool can be calculated as follows:

$$M = \min_{\mathbf{T} \in \Pi(\boldsymbol{\mathcal{G}_f}, \boldsymbol{\mathcal{G}_p})} \sum_{u,v} T_{uv} C^P_{uv} + \lambda H(\mathbf{T}), \quad (3)$$

where the optimization algorithm for the transport matrix is outlined in Algorithm 2. This matrix optimizes by identifying the best associations between image patches and token sequences, reducing the risk of mismatches. Since a lower optimal transport matching score indicates greater similarity between image-caption pairs, we redefine $\hat{M} = 1 - M$ to align with CLIP's concept of similarity in matching. For different pixel-level feature sets and token-sequence feature sets, we have different representations for the distribution of patch-level features $\boldsymbol{\mathcal{G}_f}$, token-level features $\boldsymbol{\mathcal{G}_p}$, the similarity matrix S^P, and the transportation matrix T. To simplify the notation, we omit the corresponding subscripts.

Given N image within a mini-batch, we compute the similarity score between each image and every caption in the caption pool, resulting in N similarity matrix $\mathcal{M} = \{\hat{M}^P_i\}_{i=1}^N \in \mathbb{R}^{N \times P}$. Next, for each image features, we select the most matching caption feature from the pool $\mathcal{P}^s$ based on the similarity matrix. For the j-th image feature, the selected caption feature is as follows:

$$y^s_{m(j)} = y_p \left[\arg \max_{1 \le p \le P} \hat{M}^P_j [p] \right]. \quad (4)$$

Through above operations, we can obtain the matched caption $y^s_{m(j)}$ in the pool that is most similar to f^s_j, resulting in the matching fine-grained feature $\{f^s_j, y^s_{m(j)}\}_{j=1}^N$ within a mini-batch. Similarly, we can obtain the global feature $\{f^g_j, y^g_{m(j)}\}_{j=1}^N$ for the matched image-caption pairs. Therefore, we can break the poisoning data to prevent it from being used during pre-training.

4.3 Fine-grained Alignment

Through optimal transport-based matching, we obtain the global feature $\{f^g_j, y^g_{m(j)}\}_{j=1}^N$ of the matched image-caption pairs. To facilitate multi-modal interaction, we first use the CLIP loss for optimization as follows:

$$\mathcal{L}_c = -\frac{1}{2N} \sum_{i=1}^{N} \log \left[\frac{\exp\left(\left\langle f^g_i, y^g_{m(i)} \right\rangle / \tau \right)}{\sum_{j=1}^{N} \exp\left(\left\langle f^g_i, y^g_{m(j)} \right\rangle / \tau \right)} \right] - \frac{1}{2N} \sum_{j=1}^{N} \log \left[\frac{\exp\left(\left\langle f^g_j, y^g_{m(j)} \right\rangle / \tau \right)}{\sum_{i=1}^{N} \exp\left(\left\langle f^g_i, y^g_{m(j)} \right\rangle / \tau \right)} \right], \quad (5)$$

where τ is the temperature coefficient in CLIP.

Inter-modality Fine-grained Alignment. In addition to the CLIP semantic loss, which focuses on global feature alignment, we further propose a fine-grained feature alignment loss across different modalities. Similarly Eq 3, for any single matched pair $\{f_j^s, y_{m(j)}^s\}$ within the set $\{f_i^s, y_{m(i)}^s\}_{i=1}^N$, we define the distribution of patch-level features of images and token-level features of matched captions $\boldsymbol{\mathcal{G}_f}$ and $\boldsymbol{\mathcal{G}_m}$, respectively. Then, we define the cost matrix $C^m = 1 - S^m$, where S^m denotes the similarity matrix between image patches and caption tokens within an image-caption pair. The loss for inter-modality fine-grained alignment can be defined as the optimal transport problem as follows:

$$\mathcal{L}^a = \min_{\mathbf{T} \in \Pi(\boldsymbol{\mathcal{G}_f}, \boldsymbol{\mathcal{G}_m})} \sum_{u,v} T_{uv}^m C_{uv}^m + \lambda H(\mathbf{T}). \tag{6}$$

For N image-caption pairs in a mini-batch, we compute the loss for each pair, resulting in N losses $\{\mathcal{L}_i^a\}_{i=1}^N$. The total inter-modality fine-grained alignment loss is the sum of all individual losses as $\mathcal{L}_{IM} = \sum_{i=1}^N \mathcal{L}_i^a$. It can enhance the alignment between matched image patches and caption tokens while simultaneously maximizing the separation between non-matching ones. During optimization, the transport matrix assigns larger weights to image patches and caption tokens with higher similarity and smaller weights to those with lower similarity. Therefore, the model effectively alleviates the risk of being negatively affected by irrelevant information during training by prioritizing the high-similarity image patches and caption tokens. This is achieved through the optimization of the transport matrix, as outlined in Algorithm 2.

Intra-modality Fine-grained Alignment. While inter-modal fine-grained alignment can improve the feature correspondence between image patches and text tokens, it is not sufficient to fully resolve the model's confusion during training. For example, in an image containing multiple instances of the same object (*e.g.*, multiple "tires"), inter-modal fine-grained alignment will treat all these instances as identical, failing to capture the different intra-modal relationships like *"Right front on the car"*.

To address this limitation, we propose an intra-modal fine-grained alignment approach. Specifically, given two distributions $\boldsymbol{\mathcal{G}_f}$ and $\boldsymbol{\mathcal{G}_m}$ introduced in 6, we first compute the similarity matrix for text-to-text pairs, denoted as $S^{TT} \in \mathbb{R}^{hw \times hw}$, and for image-to-image pairs, denoted as $S^{II} \in \mathbb{R}^{l \times l}$, similar to Section 4.2. We then derive the cost matrices T^{II} and T^{TT} for each distribution. The loss function for intra-modality fine-grained alignment is defined as follows:

$$\begin{aligned} \mathcal{L}^s = &\min_{\bar{\mathbf{T}} \in \Pi(\boldsymbol{\mathcal{G}_f}, \boldsymbol{\mathcal{G}_f})} \sum_{u,v} T_{uv}^{II} C_{uv}^{II} + \lambda H(\bar{\mathbf{T}}) \\ &+ \min_{\hat{\mathbf{T}} \in \Pi(\boldsymbol{\mathcal{G}_m}, \boldsymbol{\mathcal{G}_m})} \sum_{u,v} T_{uv}^{TT} C_{uv}^{TT} + \lambda H(\hat{\mathbf{T}}). \end{aligned} \tag{7}$$

For N image-caption pairs in a mini-batch, we compute the loss for each pair, resulting in N losses $\{\mathcal{L}_i^s\}_{i=1}^N$. The total intra-modality fine-grained alignment loss is the sum of all individual losses as $\mathcal{L}_{SM} = \sum_{i=1}^N \mathcal{L}_i^a$. The alignment loss can separately enhance the intrinsic relationships of each modality, avoiding inter-modality fine-grained alignment compromises the intrinsic relationships of each modality.

Following RoCLIP (Yang and Mirzasoleiman, 2023), the caption pool is considered a first-in-first-out queue, which is initialized with random caption representations. After training on every mini-batch, we update this pool by taking the caption representations of the N examples in the mini-batch and concatenating them at the end of the queue. We discard the oldest N elements from the queue, which equals the training batch size.

4.4 Training and Inference

Training. To ensure the model performs well, we use a relatively large pool size for the image-caption pairs. This allows every clean image to find a caption that is similar to its original caption. To prevent the model from becoming overly focused on intra-modal features, we train using the intra-modal fine-grained alignment loss ($\mathcal{L}_{SM}$) every K epochs. Follow RoCLIP (Yang and Mirzasoleiman, 2023), the K is set to 2 during pre-training. The overall loss function can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}} = &\lambda_c \mathcal{L}_c + \lambda_{IM} \mathcal{L}_{IM} \\ &+ \mathbf{1}\{\text{epoch mod } K = 0\} \, \lambda_{SM} \mathcal{L}_{SM}. \end{aligned} \tag{8}$$

Inference. The global features are obtained by averaging the aligned fine-grained features. During inference, we follow previous methods to use the global features.

5 Experimental Analyses

In this section, we evaluate the effectiveness of OTCCLIP against strong targeted data poisoning and backdoor attacks. We begin by outlining the experimental setup, followed by our main results,

Table 1: Downstream linear probe and zero-shot (top-1) accuracy of pre-training on CC1M. Highest performance is bold, and the lowest is underscored. The last column highlights the average improvement over CLIP.

Method	Task	F102	Fd101	I1K	Pet	Cars	Cal101	C10	C100	DTD	Air.	Average
CLIP	0-shot	1.0	**7.1**	9.6	3.4	**0.8**	34.90	34.90	7.3	3.7	0.8	10.35
	lin-prb	99.50	44.90	22.20	48.20	12.90	70.40	70.50	45.80	48.20	24.90	48.75
RoCLIP	0-shot	0.83	6.34	6.63	3.68	0.72	30.38	30.14	9.52	3.56	**1.11**	9.291
	lin-prb	99.22	54.05	24.09	52.36	20.35	72.15	78.99	57.82	55.21	32.55	54.679
SAFECLIP	0-shot	0.62	6.29	9.87	**5.51**	0.75	40.69	39.7	10.41	3.14	0.48	11.746
	lin-prb	99.38	45.58	24.53	51.02	15.35	74.4	71.90	47.32	56.01	27.63	51.324
OTCCLIP	0-shot	**1.19**	6.57	**10.50**	4.17	0.46	**45.38**	**41.90**	**15.44**	**4.52**	0.99	**13.112**
	lin-prb	**99.81**	**56.26**	**25.40**	**52.79**	**20.63**	**84.95**	**79.17**	**58.46**	**56.97**	**32.85**	**56.731**

and conclude with an ablation study on various components of OTCCLIP.

Pre-training Data. To ensure broad dataset coverage, we utilize three diverse datasets: Conceptual Captions 3M (CC3M) (Sharma et al., 2018), Visual Genome (VG) (Krishna et al., 2017), and MSCOCO (Lin et al., 2014). Following (Yang and Mirzasoleiman, 2023), we randomly sample 1M image-caption pairs from CC3M (denoted as CC1M) to further evaluate OTCCLIP's defense capabilities. Throughout all experiments, we maintain a consistent set of hyperparameters: a learning rate of 5×10^{-5}, $\lambda_c = 1$, $\lambda_{SM} = 0.4$, $\lambda_{IM} = 2$, and P = 10000. These settings demonstrate OTCCLIP's robustness against various types of attacks, independent of dataset distribution. Consistent with the setup in (Radford et al., 2021), we employ a ResNet-50 as the image encoder and a transformer as the text encoder, training OTCCLIP from scratch over 32 epochs and the matching frequency is set to 2 to effectively counter the poison.

Attack Baselines. We follow the methodologies of previous work (Yang and Mirzasoleiman, 2023; Yang et al., 2024) to evaluate our defense strategy. For TDPAs, we randomly select images from the CC3M validation set as target images. Each target is assigned a random class from the ImageNet1K dataset (Deng et al., 2009), and an adversarial caption set is constructed related to the adversarial label, as detailed in Sec. 3.2. The poison rate is set at 0.05% across all datasets. For BAs, we randomly select images from the CC3M validation set and apply the respective backdoor triggers. Each attack starts with a random class from the ImageNet1K dataset, creating adversarial caption sets. Each backdoor image pairs with a randomly chosen poisoned caption from this set. We evaluate with a poisoning ratio of 0.5% for TPDA and 5% for backdoor attacks on MSCOCO and Visual Genome. For CC1M, we use a 0.5% poisoning ratio for both TPDA and the four additional backdoor attacks.

5.1 Downstream Performance of OTCCLIP

We evaluate the performance of OTCCLIP on several datasets from (Kornblith et al., 2019), with details provided in the Appendix. It can be seen that effectively improves the zero-shot and linear-probe classification performance across all ten datasets in Table 1. RoCLIP may introduce mismatching data by using CLIP's global semantic matching, leading to a noticeable drop in zero-shot classification performance. To ensure the effectiveness of defense, SAFECLIP discards a large amount of clean data along with the poisoned samples, which reduces the model's linear probe performance. In contrast, OTCCLIP adopts a matching approach based on optimal transport to reconstruct image-caption pairs during pre-training. As a result, it avoids any decline in both zero-shot and linear probe classification performance, making our method more practical and effective.

5.2 Defense Performance of OTCCLIP

Here, we evaluate the performance of OTCCLIP against TDPA and BAs, comparing it with CLIP, RoCLIP, and SAFECLIP in terms of both ASR (Attack Success Rate) and downstream performance. Table 2 demonstrates the high effectiveness of our OTCCLIP against CLIP, with ASRs exceeding 60% for TDPA across all datasets and even surpassing 90% for some BAs. This underscores the significant challenge in ensuring CLIP's robustness. In contrast, OTCCLIP significantly reduces the ASR to 0% across all datasets for both TDPA and BAs. Although both RoCLIP and SAFECLIP provide decent defense, their performance is less consistent compared to OTCCLIP. For instance, SAFECLIP's ASR on some datasets is higher than that of OTCCLIP.

Table 2: Effectiveness of OTCCLIP in defending against various data poisoning attacks, measured by Attack Success Rate (ASR). OTCCLIP achieves a strong defense across datasets and attacks.

Dataset	**MSCOCO**				
Attacks	TDPA	BadNet	Label Consis	Blended	WaNet
CLIP	68.75%	31.0%	67.96%	92.50%	11.72%
RoCLIP	43.75%	5.63%	11.50%	2.6%	7.20%
SAFECLIP	25%	0.33%	**0%**	36.67%	2.67%
OTCCLIP	6.25%	**0%**	**0%**	**0%**	**0%**
Dataset	**Visual Genome**				
Attacks	TDPA	BadNet	Label Consis	Blended	WaNet
CLIP	75.00%	6.90%	32.84%	86.97%	19.96%
RoCLIP	37.5%	4.33	7.31%	19.60%	9.71%
SAFECLIP	6.25%	**0%**	**0%**	6.33%	**0%**
OTCCLIP	**0%**	**0%**	**0%**	**0%**	**0%**
Dataset	**CC1M**				
Attacks	TDPA	BadNet	Label Consis	Blended	WaNet
CLIP	93.75%	93.25%	71.0%	99.30%	97.42%
RoCLIP	56.25%	11.72%	5.31%	23.71%	26.27%
SAFECLIP	6.25%	**0%**	**0%**	5.47%	3.43%
OTCCLIP	**0%**	**0%**	**0%**	0.3%	**0%**

5.3 Ablation Study

Impact of Optimal Transport-based Matching. We conducted ablation experiments to evaluate the impact of Optimal Transport Matching. As shown in Table 3, replacing optimal transport-based matching with CLIP's semantic matching significantly improves ASR across all datasets and decreases CLIP's zero-shot and linear probing performance. This highlights the importance of Optimal Transport Matching in constructing clean samples.

Impact of Inter-modality Fine-grained Alignment. The third row of Table 3 shows that removing the inter-modality fine-grained alignment leads to a decrease in CLIP's zero-shot and linear probing performance. This demonstrates that inter-modality fine-grained alignment is essential for better aligning the fine-grained features of image-caption pairs and improving generalization performance.

Impact of Intra-modality Fine-grained Alignment. We also evaluate the impact of intra-modality fine-grained alignment. From Table 3, we can observe that removing the content relationship within each modality leads to a decrease in CLIP's zero-shot and linear probing performance. This proves that intra-modality fine-grained alignment is helpful in improving model performance.

5.4 Visualization of OTCCLIP Matching

As shown in Figure 3, when poisoned data is fed into OTCCLIP, OTCCLIP first breaks the association between poisoned image-caption pairs and then re-matches each image to a caption that is most similar. In Figure 3, we can see that optimal transport-based matching effectively matching captions with semantics similar to the images.

Table 3: Ablation study for different module. Linear probe and zero-shot performance is reported on CIFAR-10 (C10), CIFAR-100 (C100), ImageNet-1K (I1K).

①	②	③	Task	C10	C100	I1K	TDPA
✓	✓	✓	0-shot	41.90	15.44	10.50	0%
			lin-prb	79.17	58.46	25.40	
✘	✓	✓	0-shot	36.47	11.30	8.60	12.5%
			lin-prb	75.03	55.90	23.19	
✓	✘	✓	0-shot	39.62	13.60	9.70	0%
			lin-prb	76.50	56.80	24.10	
✓	✓	✘	0-shot	40.15	14.17	10.63	0%
			lin-prb	78.10	57.23	23.37	

① Optimal Transport-based Matching (section 4.2)
② Inter-modality Fine-grained Alignment (section 4.3)
③ Intra-modality Alignment (section 4.3)

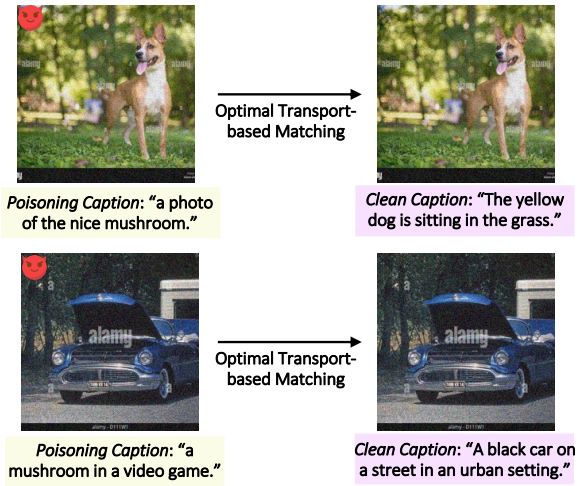

Figure 3: Visualization results of OTCCLIP re-matching to the most similar caption in caption pool based on optimal transport.

6 Conclusion

Recent studies have shown that CLIP is extremely vulnerable to targeted data poisoning and backdoor attacks. Previous methods solely rely on the global representations of images and captions, overlooking fine-grained features. To address their limitations, we propose an Optimal Transport-based framework to reconstruct the image-caption pairs, named OTCCLIP. It models images and captions with fine-grained visual and textual feature sets and re-assigns new captions based on optimal transport distance. Additionally, we encourage the inter- and intra-modality fine-grained alignment by employing optimal transport-based objective functions. Our experiments demonstrate that OTCCLIP can successfully decrease the attack success rates. Compared to previous methods, OTCCLIP significantly improves CLIP's zero-shot and linear probing performance trained on poisoned datasets.

Limitation

We employ Optimal Transport-based matching to defend against data poisoning and backdoor attacks. However, we note that although the model's defense performance has improved, the need for Sinkhorn iterations to compute the optimal transport matrix introduces additional computational overhead. These iterations require more time and computational resources compared to directly utilizing CLIP's similarity-based computations. While this trade-off enhances defense effectiveness, the increased resource consumption may become a limiting factor, particularly in large-scale defense scenarios with extensive datasets. We acknowledge this limitation and plan to optimize the OT-based process in future work to reduce computational cost and improve overall efficiency without compromising defense performance.

Ethics Statement

While the malicious application of data poisoning and backdoor attacks may raise ethical concerns, we propose a more effective defense method using Optimal Transport to mitigate these threats. This approach can help minimize potential harm from such vulnerabilities. The primary goal of this work is to encourage the development of appropriate defense mechanisms rather than to promote malicious use. We believe that by addressing these challenges, our efforts will inspire the research community to create more responsible and secure AI systems, fostering the development of trustworthy models that can better withstand adversarial attacks.

References

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

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

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

Nicholas Carlini and Andreas Terzis. 2021. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*.

Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. 2022. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Cuturi M Sinkhorn Distances. 2013. Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300.

Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. 2023. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16352–16362.

Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. 2023. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4005–4014.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. 2025. Detecting backdoor samples in contrastive language image pretraining. *arXiv preprint arXiv:2502.01385*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy

text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. 2019. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Junhao Kuang, Siyuan Liang, Jiawei Liang, Kuanrong Liu, and Xiaochun Cao. 2024. Adversarial backdoor defense in clip. *arXiv preprint arXiv:2409.15968*.

Bin Li, Ye Shi, Qian Yu, and Jingya Wang. 2024. Unsupervised cross-domain image retrieval via prototypical optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3009–3017.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference*, pages 121–137. Springer.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Zejun Li, Zhihao Fan, Huaixiao Tou, and Zhongyu Wei. 2022. Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. *arXiv preprint arXiv:2201.12596*, 1.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Anh Nguyen and Anh Tran. 2021. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*.

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. 2024. Bdetclip: Multimodal prompting contrastive test-time backdoor detection. *arXiv preprint arXiv:2405.15269*.

Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. 2023. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *Transactions on Machine Learning Research*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Indranil Sur, Karan Sikka, Matthew Walmer, Kaushik Koneripalli, Anirban Roy, Xiao Lin, Ajay Divakaran, and Susmit Jha. 2023. Tijo: Trigger inversion with joint optimization for defending multimodal backdoored models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 165–175.

Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. 2018. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pages 5133–5142. PMLR.

Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. 2024. Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. *arXiv preprint arXiv:2409.17601*.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. 2024. Better safe than sorry: Pre-training clip against targeted data poisoning and backdoor attacks. In *Proceedings of the 41st International Conference on Machine Learning*, page 235.

Wenhan Yang and Baharan Mirzasoleiman. 2023. Robust contrastive language-image pretraining against adversarial attacks. *arXiv preprint arXiv:2303.06854*.

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. 2023. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pages 39299–39313. PMLR.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.

A Appendix

Algorithm 1 OTCCLIP for Defense Against Data Poisoning

Input:
- Image encoder E_I, text encoder E_T
- OTCCLIP frequency K
- Fine-grained caption pool $\mathcal{P}^s = \{y^s_{p(i)}\}_{i=1}^P$ and global caption pool $\mathcal{P}^g = \{y^g_{p(i)}\}_{i=1}^P$, initialized with random captions

for epoch = $1, \ldots, T$ **do**
  for each mini-batch of image-caption pairs $\{(X_i, Y_i)\}_{i=1}^N \in D$ with corresponding fine-grained features $(f_i^s, y_i^s)_{i=1}^N$ and global features $(f_i^g, y_i^g)_{i=1}^N$ **do**
    if epoch mod $K == 0$ **then**
      //optimal transport-based matching score
      for $i = 1, \ldots, N$ **do**
        Compute
        $M = \min_{\mathbf{T} \in \Pi(\mathcal{G}_f, \mathcal{G}_p)} \sum_{i,j} T_{ij} C^P_{ij} + \lambda H(\mathbf{T})$
        $\hat{M} = 1 - M$
      end for
      $\mathcal{M} = \{\hat{M}_i\}_{i=1}^N$
      //extract the indices of the best matches
      Index $= \arg\max_i \mathcal{M}[:, i]$
      //Retrieve updated captions:
      $y^s_m = y^s[:, \text{Index}]$
      $y^g_m = y^g[:, \text{Index}]$
      //train encoders with loss:
      $\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_{\text{SM}} \mathcal{L}_{\text{SM}} + \lambda_{\text{IM}} \mathcal{L}_{\text{IM}}$
    else
      //train encoders with simplified loss
      $\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_{\text{IM}} \mathcal{L}_{\text{IM}}$
    end if
  end for
end for

A.1 Experimental Setup

A.1.1 Training Dataset

MSCOCO. MSCOCO (Lin et al., 2014) is a large-scale dataset designed for object detection, segmentation, and captioning. It includes 80 object categories, with each image paired with 5 captions. For our analysis, we randomly select one caption per image, resulting in a dataset size of 80K images.

Visual Genome. Visual Genome (Krishna et al., 2017) is an extensive dataset focused on region captions. It contains 10,877 images and 5.4 million region descriptions. For each image, we randomly select 5 region descriptions and combine them into a single caption.

Conceptual Captions. Conceptual Captions (Sharma et al., 2018) is a large-scale, web-based image captioning dataset that covers a diverse range of image styles and caption formats.

Algorithm 2 Sinkhorn Iteration for Optimal Transport

Require: C: cost matrix, P: number of caption pool, $h \times w$: number of spatial image features, l: length of caption, β: scaling parameter
Ensure: T: transport matrix

$\sigma \leftarrow \text{ones_like}(P, h \times w, 1)/m$
$T \leftarrow \text{ones_like}(P, l, h \times w)$
$A \leftarrow \exp(-(\text{clamp}(C, \max(10 \cdot \beta)))/\beta)$
for $i = 1$ **to** 100 **do**
  $\delta \leftarrow 1/1/n \cdot \sum(Q \cdot \sigma, \text{ axis} = 2)$
  $a = \sum(Q \cdot \delta, \text{ axis} = 2)$
  $\sigma = 1/m \times a$
  $T \leftarrow \delta \times Q \times K$
end for
return T

A.1.2 Evaluation Setup for Targeted Data Poisoning

Downstream Dataset. To assess the downstream performance of our model, we perform linear probing and zero-shot classification, as detailed in Sec. 3.1, on 10 widely adopted datasets (Radford et al., 2021; Li et al., 2021; Yang and Mirzasoleiman, 2023) listed in Table 4.

Table 4: Details of downstream datasets.

Dataset	Classes	Train Size	Test Size
CIFAR10	10	50,000	10,000
CIFAR100	100	50,000	10,000
Food-101	101	75,750	25,250
DTD	47	3,760	1,880
FGVC Aircraft	100	6,667	3,333
Flowers-102	102	2,040	6,149
Caltech-101	102	3,060	6,085
OxfordIIITPet	37	3,680	3,669
Stanford Cars	196	8,144	8,041
ImageNet1K	1000	50,000	50,000

Zero-shot Classification. Zero-shot classification assesses the generalization and transferability of the model to unseen tasks. It transforms the

downstream labels into natural language captions using the provided engineered prompt templates, such as "`A photo of a {label}`" (Radford et al., 2021). Then, it calculates the cosine similarity between the representations of a given image and each prompt and predicts the label with the highest image-prompt similarity.

Linear Probe Classification. Linear probe classification refers to evaluating the extracted representations from the pre-trained image encoder to train a linear classifier on the downstream labeled data.

A.1.3 Defense Baselines for Backdoor

We consider RoCLIP (Yang and Mirzasoleiman, 2023) and SAFECLIP (Yang et al., 2024) as our baseline. We measure the effectiveness of attacks using attack success rate (ASR). For TDPA, ASR is measured as the fraction of target images that are classified as the adversarial label. For BA, ASR is measured as the fraction of test images containing the backdoor triggers that are classified as the adversarial label.

A.1.4 Backdoor Attacks Used in Our Evaluations

We follow the methodologies of previous work (Yang and Mirzasoleiman, 2023; Yang et al., 2024) to evaluate our defense strategy against backdoor attacks (BA) with visible triggers (e.g., BadNet) and invisible triggers (e.g., Blended and WaNet). Figure 4 illustrates various examples of backdoor attacks for visualization.

(a) Blended (b) WaNet (c) BadNets

Figure 4: Backdoor attacks used in our evaluations.

A.1.5 Hyperparameters Setting

The hyperparameter settings used in our experiments are provided in Table 5. As shown in Table 5, our model employs consistent hyperparameters across all datasets, highlighting the robustness of OTCCLIP against a variety of attacks.

Table 5: Hyperparameters of our experiments.

Dataset	lr	Batch Size	K
CC3M	5e-5	256	2
CC1M	5e-5	256	2
COCO	5e-5	256	2
VG	5e-5	256	2

A.2 OTCCLIP Against Adaptive Attacks

In the above experiments, we assume that attackers have no information about our backdoor defense. In this section, we consider a more challenging setting, where the attackers know the existence of our defense and can construct the poisoned dataset with an adaptive attack.

Threat Model For The Attackers. Following existing work (Gao et al., 2023), we assume that the attackers can access all dataset and know the architecture of the victim model. However, the attackers can not control the training process after poisoned samples are injected into the training dataset.

Methods. Our defense method uses optimal transport-based matching to separate samples and reconstruct image-caption pairs. For adaptive attacks, the goal is to minimize the difference in optimal transport-based matching between the image and its poisoned caption. Attackers first use the image encoder and text encoder to extract spatial and token sequence features from the poisoned pairs. Then, the loss function defined in Eq. 3 is applied to update the trigger patch. This pattern is optimized by minimizing the gradient of the poisoned image-caption pair, reducing the distance between them in the fine-grained feature space.

Settings. We conduct experiments on the poisoning image-caption pairs. We adopt projected gradient descent (PGD) (Wang et al., 2018) to optimize the trigger pattern for 100 iterations.

Results. This adaptive attack achieves a 0% attack success rate on both MSCOCO and Visual Genome, demonstrating that our defense effectively resists adaptive attacks.

A.3 Additional Experiments

A.3.1 OTCCLIP's Complexity Compared to Existing Defense Methods

RoCLIP leverages CLIP's global employ global feature vectors to match the most similar for every image, aiming to break the association between poisoned pairs. SAFECLIP identifies the clean and risky set using global features. SAFECLIP apply the CLIP loss to the safe set and SAFECLIP apply unimodal CL to image and text modalities of the risky set separately. SAFECLIP performs data augmentation on the risky data and applies unimodal contrastive learning (CL) in the risky and augmented data. We propose the optimal transport-based fine-grained matching and alignment against data poisoning.

As shown in the Table 6, we calculated the computational cost of these three methods within a single epoch under the same settings. From the methodology section, we found that OTCCLIP requires using Sinkhorn iteration (Distances, 2013) to obtain the optimal transport matrix. As shown in Table 6, we calculated the computational cost of these three methods per epoch under the same settings. According to the methodology, OTCCLIP requires Sinkhorn iteration (Distances, 2013) to compute the optimal transport matrix, introducing slightly more computational time compared to RoCLIP. However, it is significantly faster than SAFECLIP. As noted in SAFECLIP (Yang et al., 2024), data augmentation is applied to risky data, generating augmented samples. Both the augmented and original risky data are used for training. Since approximately 75% of the data are marked as risky data, the training data set almost doubles in size, significantly increasing the training time.

Table 6: Training time of OTCCLIP compared to existing defense methods.

Method	Training Time
RoCLIP	1 h 23 min
SAFECLIP	4 h 11 min
OTCCLIP	2 h 7 min

A.3.2 More Ablation Studies

Impact of Caption Pool Size. Next, we analyze the effect of pool size on our method. We apply OTCCLIP with pool sizes of 1%, 2%, and 10% of the pre-training dataset. As shown in Table 7, the pool size does not significantly impact the effectiveness of the defense. Across different pool sizes, our method consistently defends against data poisoning attacks. However, a larger pool size improves the downstream performance of the model, as it increases the likelihood of images finding more suitable captions.

Table 7: The impact of caption pool size. Linear probe and zero-shot performance is reported on CIFAR-10 (C10), CIFAR-100 (C100), ImageNet-1K (I1K).

Pool Size	Task	C10	C100	I1K	TDPA
1024	0-shot	41.50	15.20	9.6	0%
	lin-prb	79.03	55.7	24.3	
2048	0-shot	40.13	15.07	10.76	0%
	lin-prb	79.40	56.64	24.9	
10000	0-shot	41.90	15.44	10.50	0%
	lin-prb	79.17	58.46	25.40	