ICCV
#8160

Unlearning the Noisy Correspondence Makes CLIP More Robust

Anonymous ICCV submission

Paper ID 8160

Abstract

The data appetite for Vision-Language Models (VLMs) has continuously scaled up from the early millions to billions today, which faces an untenable trade-off with data quality and inevitably introduces Noisy Correspondence (NC) samples. Undoubtedly, such semantically unrelated data significantly impairs the performance of VLMs. Previous efforts mainly address this challenge by estimating refined alignment for more precise guidance. However, such resource-intensive pipelines that train VLMs from scratch struggle to meet realistic data demands. In this paper, we present a brand new perspective that seeks to directly eliminate the harmful effects of NC in pre-trained VLMs. Specifically, we propose NCU, a Noisy Correspondence Unlearning fine-tuning framework that efficiently enhances VLMs' robustness by forgetting learned noisy knowledge. The key to NCU is learning the hardest negative information, which can provide explicit unlearning direction for both false positives and false negatives. Such twin goals unlearning process can be formalized into one unified optimal transport objective for fast fine-tuning. We validate our approach with the prevailing CLIP model over various downstream tasks. Remarkably, NCU surpasses the robust pre-trained method on zero-shot transfer while with lower computational overhead. The code will be released upon acceptance.

1. Introduction

The pursuit of general intelligence has driven progress in multimodal learning, which seeks to integrate and understand multiple sensory modalities like humans. Large-scale vision-language training, exemplified by CLIP [37], is seen as a key milestone in multimodal learning due to its remarkable transfer capabilities in real-world applications, such as image-text retrieval [13, 22, 34] and robotics control [41].

However, much of their success can be attributed to scaling laws enabled by massive training data. As every coin has two sides, the insatiable demand for data forces a difficult trade-off between quantity and quality, which inevitably introduces noisy correspondence into the training set. Taking the CC3M dataset [39] as an example, despite being filtered from 500 million images, it still contains at least 3% [20] unrelated image-text pairs, *i.e.*, *false positive*. To make matters worse, training on massive data necessitates larger batch sizes (32K used in CLIP), which increases the likelihood of unpaired samples sharing semantic similarities, *i.e.*, *false negative*. Undoubtedly, such two-aspects noisy correspondence can significantly impair the performance of vision-language models.

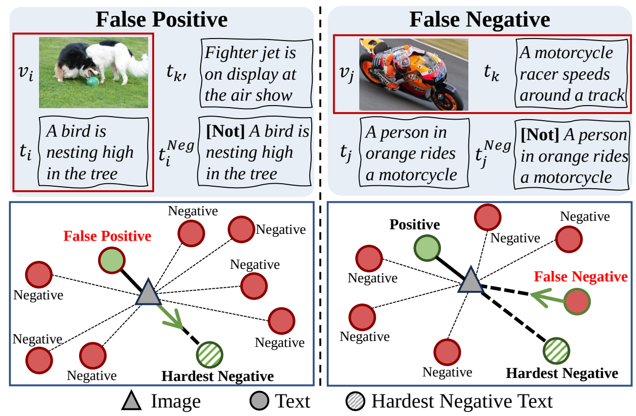


Figure 1. **Illustration on the core concept of NCU.** The twin goals unlearning process is guided by the learned hardest negative information. For the FP, t_i^{Neg} directly pulls v_i away from the mismatched t_i. While for the FN, t_i^{Neg} acts as a distance upper bound to facilitate modeling many-to-many relations.

To endow robustness against NC, one natural direction is to revise the pre-training paradigm [1, 3, 11, 12, 21] that supervises VLMs with refined alignment. However, existing methods require training from scratch and may rely on guidance from external large models [3, 11]. Such resource-intensive pipelines obviously struggle to face the realistic demand, especially with today's billion-scale datasets [38]. Hence, it is necessary to address the NC problem in vision-language training with a cost-effective method.

In this paper, we think outside the box of robust pre-training and pose an important question: *Can we directly eliminate the harmful effects of NC in pre-trained VLMs?*

To answer this question, we resort to machine unlearning [2] and present NCU, a **N**oisy **C**orrespondence **U**nlearning fine-tuning framework that improves the robustness of CLIP by erasing learned noisy knowledge. Machine unlearning is a reversed learning process that aims to delete the influence of specific training samples from trained models. Despite its promise in widespread tasks [7, 9], unlearning the NC in VLMs remains unexplored due to a key challenge: the ambiguous forgetting direction would corrupt the learned semantic structure in the feature space. To address this, we propose to learn the hardest negative information that can provide explicit unlearning direction. As illustrated in Fig. 1, on the one hand, the negative information would directly serve as reliable supervision for forgetting false positives. On the other hand, it would facilitate the modeling of many-to-many relationships among unpaired data for forgetting false negatives. Then, we show that such twin goals unlearning process can be formalized as one unified optimal transport problem, which efficiently fine-tunes CLIP to resist both FP and FN.

Our main contributions are highlighted below:

- To the best of our knowledge, this work could be the first study to eliminate the harmful effects of noisy correspondence from pre-trained CLIP.
- We propose the NCU framework, which efficiently unlearns FP and FN with explicit direction derived from the hardest negative information.
- We demonstrate that NCU achieves significant improvements over CLIP on several downstream tasks and surpasses the previous robust pre-training method with lower computational overhead.

2. Related Work

Noisy Correspondence Learning. Noisy correspondence refers to the alignment error presented in multimodal data. The false positive is a typical NC problem, where irrelevant multimodal pairs are wrongly treated as matched. To alleviate this, several techniques have been developed in various multimodal applications, including cross-modal retrieval [16, 18, 20, 36], video temporal learning [17, 30], multimodal person re-identification [35, 47], question answering [23], and image captioning [10, 24]. In more complex scenarios, *e.g.*, vision-language pre-training [19, 21], models also suffer from false negatives caused by the training paradigm [12], where similar unpaired samples are forced to be distant. Considering the computational burden of large VLMs, this work presents a low-carbon solution to directly improve the robustness of pre-trained VLMs.

Contrastive Vision Language Models. Contrastive vision language models (VLMs) [8, 13, 14, 33, 44, 48] aim to learn visual representations by the corresponding textual supervision, which have attracted significant attention due to their simplicity and powerful representation capability. Pioneering works CLIP [37] and ALIGN [22] have shown great success via learning from massive image-text pairs. However, such web-crawled data are noisy [38, 42] and inevitably harm the efficacy of existing VLMs. To tackle this issue, a series of works attempted to train the VLM with refined soft image-text alignments by label smoothing [12], knowledge distillation [1], fine-grained intra-modal guidance [11], text rewriting [3], or positive-negative contrastive loss [21]. Besides, OT-based methods[40, 46] have also emerged as they naturally model such matching problems. Despite the success, previous works focus on training robust VLMs from scratch, which overlooks readily available pre-trained models and incurs unnecessary computational costs. To this end, this paper pioneers an efficient approach to enhance model robustness by unlearning noisy information from pre-trained models.

Machine Unlearning. Recent advances in MU mainly focus on practical approximate unlearning, which seeks to mimic the behavior of a model re-trained from scratch. Driven by privacy concerns, existing MU works [7, 9, 32] in computer vision focus on image classification that attempts to forget specific classes. In parallel, MU has also become a popular topic in large language models due to its capability to eliminate harmful responses [31, 50]. However, multimodal forgetting remains under-explored in the literature. Pioneering works [26, 27] studied data-free class removal for CLIP's downstream image classification. To date, none of the existing MU methods has explored the unlearning of noisy concepts from VLMs.

3. Preliminaries

3.1. Contrastive Language-Image Pre-training

CLIP is a vision-language model trained on millions of web-harvested image-text pairs. We consider a batch of N image-text pairs $\{v_i, t_i\}_{i=1}^{N}$ sampled from a cross-modal dataset $\mathbb{D}$, where v_i and t_i represent the raw image and corresponding text, respectively. The goal of CLIP is to train two modality-specific encoders that bring matched pairs closer while pushing unmatched ones apart. Specifically, image embedding $\boldsymbol{v}_i \in \mathbb{R}^d$ and text embedding $\boldsymbol{t}_i \in \mathbb{R}^d$ are obtained by passing v_i and t_i through the image encoder f_v and text encoder f_t, respectively, where d is the embedding dimension. The encoded l_2 normalized embeddings are then aligned in the feature space by minimizing the contrastive objective, *i.e.*, InfoNCE loss:

$$\mathcal{L}_{v\to t}^{CL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\langle \boldsymbol{v}_i, \boldsymbol{t}_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \boldsymbol{v}_i, \boldsymbol{t}_j\rangle/\tau\right)}, \quad (1)$$

where $\langle\cdot\rangle$ represents the inner product and τ is a trainable temperature parameter. As InfoNCE loss is symmetric, we can define $\mathcal{L}^{CL}_{t\to v}$ similarly. The complete CLIP training objective is formulated as: $\mathcal{L}_{CLIP} = \mathcal{L}^{CL}_{v\to t} + \mathcal{L}^{CL}_{t\to v}$.

Despite its promising performance, the standard contrastive learning can suffer from the noisy correspondence problem in two aspects. First, the web-collected pairs inevitably contain an unknown portion of mismatched data, *i.e.*, false positives. Second, hard target alignment neglects the potential semantic similarity among unpaired samples, *i.e.*, false negatives, especially under large batch settings.

3.2. Machine Unlearning

Given a CLIP model (also named *reference model*) $\{f_v, f_t\}$ that is already trained on a cross-modal dataset $\mathbb{D}$, machine unlearning aims to fine-tune the model to forget a specific subset $\mathbb{D}_{FG} \subseteq \mathbb{D}$ while maintaining effectiveness on the retained set $\mathbb{D}_{RT} = \mathbb{D} \setminus \mathbb{D}_{FG}$. Ideally, the model should behave as if it were trained without any sample from $\mathbb{D}_{FG}$. In principle, re-training the model from scratch on $\mathbb{D}_{RT}$ would serve as the gold standard. However, since CLIP is trained on massive-scale data, it is unrealistic to obtain a forget set that includes all noisy information, especially when some data are not publicly accessible. Therefore, we focus on an approximate unlearning approach in which $\mathbb{D} = \mathbb{D}_{FG} \cup \mathbb{D}_{RT}$ does not need to contain all training pairs, making the unlearning process more practical for real-world scenarios.

The most straightforward method to unlearn is *gradient ascent* or its variants, which optimizes the negative prediction loss over the forget set. Another typical approach is performing *forget loss* that encourages the model to relearn the modified form of undesired data. For example, we can update CLIP by minimizing InfoNCE loss in pair $\{v_i, \tilde{t}_i\} \sim \mathbb{D}_{FG}$ to forget the relation between v_i and t_i, where $\tilde{t}_i \neq t_i$ could be random or hand-crafted text to replace the original. Based on these, existing MU methods have shown promising progress in class forgetting and LLM privacy protection. However, applying these strategies to CLIP unlearning poses a key challenge: the ambiguous forgetting direction would corrupt the learned semantic structure in the feature space. In other words, the model forgets the undesired data by learning other meaningless patterns.

4. Methodology

To tackle the above issues, we introduce the Noisy Correspondence Unlearning (NCU) framework. In the following, we first introduce the division of forget and retained sets in Sec 4.1. Subsequently, we elaborate on learning the hardest negative information in Sec 4.2 and explain how to formalize the twin goals unlearning process into an optimal transport object for efficiently fine-tuning in Sec 2. The overall training pseudo-code is shown in Supplementary A.

4.1. Identifying the Forget Set

Unlike standard MU tasks with a predefined forget set, we need to manually identify mismatched samples from CLIP's training data to construct it. As pre-trained CLIP has shown strong representation capability, we propose using the basic similarity score to obtain $\mathbb{D}_{FG}$ and $\mathbb{D}_{RT}$, *i.e.*,

$$\omega_i = \frac{1}{2}\left[\frac{\exp\left(\langle \boldsymbol{v}_i, \boldsymbol{t}_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \boldsymbol{v}_i, \boldsymbol{t}_j\rangle/\tau\right)} + \frac{\exp\left(\langle \boldsymbol{t}_i, \boldsymbol{v}_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \boldsymbol{t}_i, \boldsymbol{v}_j\rangle/\tau\right)}\right]. \tag{2}$$

By comparing (v_i, t_i) with other cross-modal samples in the batch, ω_i serves as a clean confidence that measures the extent of semantic match. Then, we select pairs with the lowest $P\%$ of ω_i within the batch as false positives to construct the forget set $\mathbb{D}_{FG}$, while treating the remaining in-batch pairs as the retained set $\mathbb{D}_{RT}$.

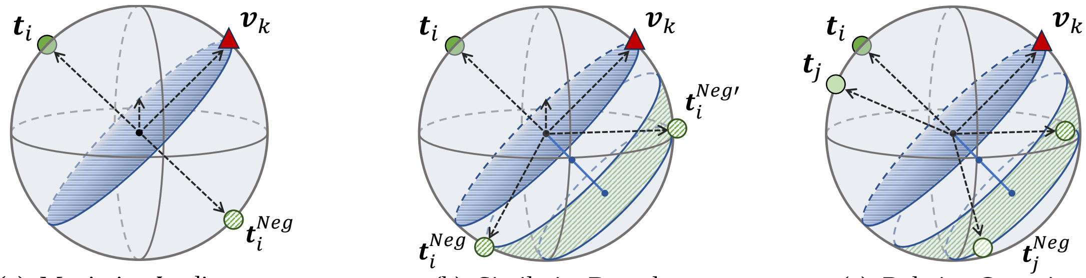
Note that $\mathbb{D}_{FG}$ and $\mathbb{D}_{RT}$ are dynamically selected at each batch, which enjoys two merits: 1) CLIP could be efficiently updated with one intra-batch optimization; 2) The forget-retain ratio could be flexibly adjusted through the predefined parameter P.

4.2. Learning Hardest Negative Semantics

To guide CLIP with an explicit unlearning direction, we aim to learn the hardest negative semantics as supervision. Intuitionally, for an irrelevant pair (v_i, t_i) that misleads the model with '*v_i and t_i are matched*', we encourage the model to forget this information by relearning that '*v_i and t_i are not matched*'. From a data utilization viewpoint, this paradigm is similar to negative learning[25] that supervises the model with complementary information [18], *i.e.*, pushing the candidate away from other unpaired samples. Differently, our method seeks the hardest negative information to avoid uncertain optimization directions.

To achieve this, we incorporate a set of learnable vectors to represent the textual negative semantics inspired by prompt learning [51]. Specifically, for any training pair (v_i, t_i), the token features of t_i are combined with m shared prompt vectors to present the corresponding negative semantics t_i^{Neg} in the feature space. While such prompt-driven semantic negation of CLIP has demonstrated success in out-of-distribution detection [29, 43], existing methods are confined to closed-set downstream tasks with limited category concepts. In contrast, our challenge lies in extending the semantic opposite operation into the open-set knowledge that CLIP pre-trained.

Intuitively, the hardest negative satisfies two constraints in the feature space: 1) t_i^{Neg} needs to maximize its distance from v_i and t_i; 2) t_i^{Neg} should maintain certain similarity to those unpaired images, as it is not a wrong description [43] despite being semantically irrelevant to v_j. Furthermore, we only use $\mathbb{D}_{RT}$ to learn the prompt tokens to avoid overfitting caused by noisy correspondence. For notation simplicity,



(a). Maximize L_2-distance (b). Similarity Bound (c). Relation Opposite

Figure 2. **Illustration of the optimization objective to learn the hardest negative semantics.** (a) Previous attempts that directly maximize the L_2 distance prevent $\boldsymbol{t}_i^{Neg}$ from providing certain guidance for unpaired images, *e.g.*, $\boldsymbol{v}_k$. (b) We bound the similarity with margins for a more relaxed semantic separation, but it may lead to uncertain targets, *e.g.*, $\boldsymbol{t}_i^{Neg}$ and $\boldsymbol{t}_i^{Neg\prime}$. (c) We further preserve relation structures for precise objectives. The intuition is that the opposite text should also maintain semantic relationships, *e.g.*, $\langle \boldsymbol{t}_i, \boldsymbol{t}_j^{Neg} \rangle \approx \langle \boldsymbol{t}_j, \boldsymbol{t}_i^{Neg} \rangle$.

we denote $\tilde{N}$ as the size of $\mathbb{D}_{RT}$ within each batch. Based on the above insights, we propose the following intra-modal and cross-modal training objectives.

Text Relation Opposite. It encourages semantic separation between the embeddings of negative text and its original. Most previous works [29, 43] typically reduce the per-instance similarity gap among textual pairs, *e.g.*, $\|\boldsymbol{t}_i - \boldsymbol{t}_i^{Neg}\|_2 \to 2$ [43] to directly maximize its L_2 distance. However, such rigid constraint incurs a crucial limitation in the open-set semantic space—enforcing maximal distance pushes t_i^{Neg} away from unpaired images (Fig. 2(a)), which is contrary to our objective. To this end, we propose a relaxed similarity bound to constrain the semantic separation:

$$\mathcal{L}^{sep} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left([\alpha - \langle \boldsymbol{t}_i, \boldsymbol{t}_i^{Neg} \rangle]_+ + [\langle \boldsymbol{t}_i, \boldsymbol{t}_i^{Neg} \rangle - \beta]_+ \right), \quad (3)$$

where $\alpha < 0$ and $\beta < 0$ are the margin parameters to locate $\langle \boldsymbol{t}_i, \boldsymbol{t}_i^{Neg} \rangle \in [\alpha, \beta]$, and $[x]_+ = max(x, 0)$ is the hinge function. As illustrated in Fig. 2(b), although minimizing Eq.(3) enables t_i^{Neg} to be distant from t_i while remaining relatively similar to unpaired images, the broad range of variations makes training convergence difficult. To address this issue, we propose to perform semantic opposite at the relation level instead of the instance level, which is achieved by preserving the geometrical structures among all negative and original text within the batch:

$$\mathcal{L}^{rel} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \left(\langle \boldsymbol{t}_i, \boldsymbol{t}_j^{Neg} \rangle - \langle \boldsymbol{t}_j, \boldsymbol{t}_i^{Neg} \rangle \right)^2. \quad (4)$$

As shown in Fig. 2(c), regularizing the negative-original relation consistency can guide t_i^{Neg} toward a precise location in the feature space.

Image-text Matching Opposite. It aims to model the alignment between the embeddings of negative text and images. As discussed, t_i^{Neg} provides positive supervision to unpaired images while separating from its paired image, which presents opposite matching patterns to the normal contrastive objective. To achieve this, we take inspiration from Sigmoid loss [49] that efficiently supports such multi-positive alignment. Specifically, it guides per cross-modal pair independently by the binary matching target:

$$\begin{aligned}\mathcal{L}^{itm} = &\frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \Big(\log \frac{1}{1 + \exp(m_{ij}(-\langle \boldsymbol{t}_i, \boldsymbol{v}_j \rangle / \tau))} \\ &+ \log \frac{1}{1 + \exp(-m_{ij}(-\langle \boldsymbol{t}_i^{Neg}, \boldsymbol{v}_j \rangle / \tau))} \Big),\end{aligned} \quad (5)$$

where m_{ij} equals 1 for $i = j$ and -1 for $i \neq j$. In Eq.(5), the first part follows the standard one-to-one matching to retain the original CLIP knowledge, while the second part utilizes the opposite binary target, *i.e.*, $-m_{ij}$, to bring t_i^{Neg} closer to multiple images.

With the visual encoder frozen, the overall loss for learning the hardest negative semantics is balanced by a scaling factor λ and given by:

$$\mathcal{L}^{HN} = \lambda(\mathcal{L}^{sep} + \mathcal{L}^{rel}) + \mathcal{L}^{itm}. \quad (6)$$

4.3. Hardest-Negative Guided Noise Unlearning

The hardest negative semantics serve dual purposes in erasing the learned noisy correspondence: 1) guide CLIP to unlearn the false positive pair (v_i, t_i) by matching v_i with t_i^{Neg}. 2) While for the well-matched pair (v_i, t_i), t_i^{Neg} assists in inferring the soft alignment among unpaired data to unlearn the false negative pattern. To efficiently fine-tune CLIP, we formalize such twin goals unlearning process into one unified Optimal Transport problem.

Optimal Transport. OT seeks to establish a flexible alignment between images and captions by computing a

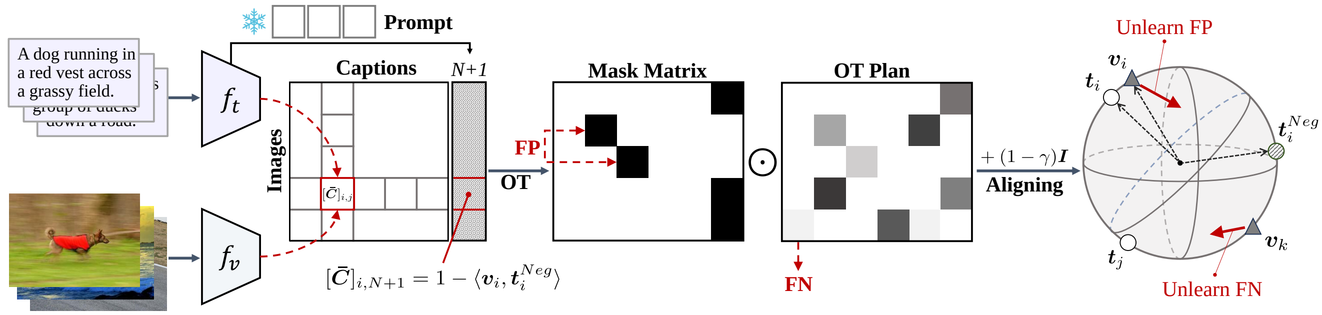


Figure 3. **Overview of the Noisy Correspondence Unlearning process.** With the learned negative prompt frozen, we formulate an optimal transport problem guided by the hardest negative and then use the solved transport plan to robustly fine-tune the model f_t and f_v.

minimal-cost transport plan, where the cost refers to the expense of transporting mass from source to target distribution and is generally set to a distance measure [15]. Let $\boldsymbol{C} \in \mathbb{R}_+^{N\times N}$ denotes the cost matrix for the mini-batch, where $[\boldsymbol{C}]_{i,j} = 1 - \langle \boldsymbol{v}_i, \boldsymbol{t}_j \rangle$ is the cosine distance of v_i and t_j. $\boldsymbol{\Gamma} \in \mathbb{R}_+^{N\times N}$ denotes the corresponding transport plan that $[\boldsymbol{\Gamma}]_{i,j}$ represents the alignment probability between v_i and t_j. Formally, the objective of OT is defined as follows:

$$\begin{gathered} \min_{\boldsymbol{\Gamma}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \boldsymbol{\Gamma}, \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{\Gamma}) \\ \text{s.t. } \Pi(\boldsymbol{\mu},\boldsymbol{\nu}) = \{\boldsymbol{\Gamma}\in\mathbb{R}_+^{N\times N} | \boldsymbol{\Gamma}\mathbb{1}_N = \boldsymbol{\mu}, \boldsymbol{\Gamma}^\top\mathbb{1}_N = \boldsymbol{\nu}\}, \end{gathered} \tag{7}$$

where $\mathbb{1}_N$ denotes a N-dimensional all-one vector, $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^N$ are probability measures representing the relative importance of each image and caption. Without prior knowledge, $\boldsymbol{\mu} = \frac{1}{N}\mathbb{1}_N$ and $\boldsymbol{\nu} = \frac{1}{N}\mathbb{1}_N$ are considered to be uniformly distributed since each pair is sampled independently. $H(\boldsymbol{\Gamma})$ is an additional entropy regularizer controlled by the smooth parameter ϵ, which enables the OT objective to be solved by the rapid Sinkhorn-Knopp algorithm [6].

Boosting OT via Hardest Negatives. To endow the transport plan with dual forgetting purposes, we reformulate Eq.(7) by imposing guidance from the hardest negative information. Specifically, for each image v_i, we extend its transport target from $\{t_i\}_{i=1}^N$ to include its paired negative text t_i^{Neg}. As shown in Fig. 3, the negative text composes a new alignable column for the transport objective, which append the cost matrix $\boldsymbol{C}$ to $\bar{\boldsymbol{C}} \in \mathbb{R}_+^{N\times(N+1)}$, *i.e.*,

$$[\bar{\boldsymbol{C}}]_{i,N+1} = 1 - \langle \boldsymbol{v}_i, \boldsymbol{t}_i^{Neg} \rangle, [\bar{\boldsymbol{C}}]_{i,j} = [\boldsymbol{C}]_{i,j}, \forall i,j \in [1,N].$$

For the two parts $\mathbb{D}_{FG}$ and $\mathbb{D}_{RT}$ within each batch, the hardest negative should impose different guidance for distinct unlearning goals. To this end, we propose a mask-base constraint to the corresponding transport plan $\bar{\boldsymbol{\Gamma}}$ that regulates the effect of t_i^{Neg}. Specifically, the mask matrix $\boldsymbol{M} \in \mathbb{R}_+^{N\times(N+1)}$ satisfies that

$$[\boldsymbol{M}]_{i,j} = \begin{cases} 0, & \text{if } (v_i,t_i) \in \mathbb{D}_{FG} \text{ and } j = i, \\ 0, & \text{if } (v_i,t_i) \in \mathbb{D}_{RT} \text{ and } j = N+1, \\ 1, & \text{otherwise.} \end{cases} \tag{8}$$

For the toy example illustrated in Fig. 3, if the pair is considered to be mismatched, the transport mass between v_i and t_i should be constrained to zero. Conversely, for the well-matched pair, t_i^{Neg} acts as a lower limit where the transport mass between v_i and t_j should be higher than it. Following the solver from [15], we model the mask constraint as the Hadamard product form that $\hat{\boldsymbol{\Gamma}} = \boldsymbol{M} \odot \bar{\boldsymbol{\Gamma}}$, and the optimal alignment is formulated as (detailed Sinkhorn solution is presented in Supplementary B):

$$\hat{\boldsymbol{\Gamma}}^* = \mathop{\arg\min}_{\hat{\boldsymbol{\Gamma}}\in\Pi(\boldsymbol{\mu},\bar{\boldsymbol{\nu}})} \langle \hat{\boldsymbol{\Gamma}}, \bar{\boldsymbol{C}} \rangle - \epsilon H(\hat{\boldsymbol{\Gamma}}), \tag{9}$$

where $\bar{\boldsymbol{\nu}} = \frac{1}{N+1}\mathbb{1}_{N+1}$ to satisfy the additional column.

The Unlearning Objective. Although $\hat{\boldsymbol{\Gamma}}^*$ provides the more refined alignment, we suggest further incorporating an identity-like matrix $\boldsymbol{I}$ for two merits. First, diagonal elements are set as 1 for true positives to retain the initial alignment. Second, $[\boldsymbol{I}]_{i,N+1} = 1$ to enhance the unlearning for the possible false positive $(v_i,t_i) \in \mathbb{D}_{FG}$. Thus, the overall alignment balanced by the factor γ is defined as:

$$\boldsymbol{T} = \gamma\hat{\boldsymbol{\Gamma}}^* + (1-\gamma)\boldsymbol{I}. \tag{10}$$

To fine-tune CLIP with this soft alignment, we use the KL divergence to optimize the matching distribution. Formally, we denote the batched similarity matrix as $\boldsymbol{P} \in \mathbb{R}^{N\times(N+1)}$ where $\boldsymbol{P_i} = [\langle \boldsymbol{v}_i, \boldsymbol{t}_1 \rangle, \ldots, \langle \boldsymbol{v}_i, \boldsymbol{t}_N \rangle, \langle \boldsymbol{v}_i, \boldsymbol{t}_i^{Neg} \rangle]^\top$. We obtain $\boldsymbol{P}_i^{v2t}$ and $\boldsymbol{P}_i^{t2v}$ by applying row-wise and column-wise softmax operation to $\boldsymbol{P}$, respectively. Correspondingly, let $\boldsymbol{T}_i^{v2t}$ and $\boldsymbol{T}_i^{t2v}$ be the row-wise and column-wise normalized refined alignment for the i-th sample, respectively. The

OT-guided re-aligning is defined as:

$$\mathcal{L}^{otr} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}(\boldsymbol{T}_i^{v2t}\|\boldsymbol{P}_i^{v2t}) + \frac{1}{N+1}\sum_{i=1}^{N+1}\mathrm{KL}(\boldsymbol{T}_i^{t2v}\|\boldsymbol{P}_i^{t2v}). \tag{11}$$

Moreover, we empirically observe that preserving the textual semantic separation term can make the unlearning more stable. Thus, the final unlearning objective is defined as:

$$\mathcal{L}^{UL} = \mathcal{L}^{otr} + \mathcal{L}^{sep}. \tag{12}$$

5. Experiment

In this section, we experimentally analyze the effectiveness of NCU in unlearning the NC knowledge from CLIP.

5.1. Setup

Datasets. Our experiments are conducted on three vision-language datasets at different scales and noise: Conceptual Captions 3M (CC3M) [39], Conceptual Captions 12M (CC12M) [4], and YFCC15M-R (provided by [14], an LLM-recaptioned subset from the YFCC100M [42]). All datasets are web-crawled and contain an unknown portion of NC pairs, *e.g.*, CC3M is estimated to include at least 3% false positives. We evaluate NCU on ImageNet and 15 common downstream datasets for classification performance and on MSCOCO and Flickr30K for retrieval capability. Details for datasets are shown in Supplementary C.

Unlearning Details. Following CLIP, we consider two architectures for the image encoder, *i.e.*, ViT/B16 and ViT/B32, while the text encoder adopts the transformer architecture. We consider a CLIP pre-trained on dataset $\mathbb{D}$, *e.g.*, CC3M, CC12M, or YFCC15M-R, as our reference model, then we perform the NC unlearning on $\mathbb{D}$ or its subset to enhance CLIP's robustness. For all experiments, we allocate 2 epochs for learning negative semantics and 8 epochs for noise unlearning. All models are trained with a batch size of 2,048 on 16 NVIDIA V100 GPUs. Detailed training settings are presented in Supplementary D.

Evaluation Protocol. We evaluate NCU's transferability with Zero-Shot (ZS) classification accuracy and Linear Probing (LP) accuracy. For ZS classification, we follow CLIP's [37] prompt templates to compute distances between class text embeddings and image features. For LP, we follow the mainstream setting [8, 37] that trains a linear classifier using L-BFGS on features extracted from the frozen image encoder. Besides, we evaluate the retrieval performance with the Recall at rank K (R@K) metric.

5.2. Evaluation on Diverse Downstream Tasks

To verify the generalization of NCU, we compare it with CLIP on three different types of downstream tasks.

Zero-Shot Transfer. We compare the zero-shot performance of CLIP and NCU on 16 popular image classification datasets. We follow the prompt templates suggested in the CLIP paper [37] to form each class name into a natural sentence. As demonstrated in Tab. 1, our NCU approach significantly outperforms the baseline CLIP model on both ImageNet and other downstream datasets. Specifically, across all fine-tuning datasets and all model architectures, NCU gains in the range of $2.8\% \sim 4.1\%$ in top-1 accuracy on ImageNet and $2.5\% \sim 4.0\%$ on average over the other downstream datasets. This reveals that NCU can successfully eliminate the impact of NC on CLIP by robust fine-tuning with the same dataset.

Image-Text Retrieval. We present the zero-shot cross-modal retrieval performance on the testing set of Flickr30K (1K) and MSCOCO (5K) in Tab. 2. Our method considerably outperforms the vanilla CLIP in almost all cases. For instance, when fine-tuning CLIP (ViT-B/32) pre-trained on the CC3M dataset, our NCU method achieves a 7.7% improvement in average recall scores on Flickr30K and 4.8% improvement in average recall scores on MSCOCO. This finding indicates that NCU can remarkably enhance the alignment of images and text in the embedding space.

Linear Probing. Tab. 3 reports the linear probing performance on 4 representative downstream datasets. Our NCU consistently surpasses CLIP in the vast majority of cases, suggesting that the visual embeddings learned by our NCU are more effective and transferable than CLIP.

5.3. Compared to Robust Methods

In this section, we compare NCU with other robust-designed techniques against NC on zero-shot ImageNet1K classification task, *i.e.*, gradient ascent (GA), and SoftCLIP [11]. Specifically, we evaluate GA as a standalone method, where $-\mathcal{L}_{CLIP}$ is performed on $\mathbb{D}_{FG}$ for handling FPs and $\mathcal{L}_{CLIP}$ with label smoothing is performed on $\mathbb{D}_{RT}$ for FNs. SoftCLIP is a noise-robust SOTA method that trains CLIP from scratch by additional intra-modal guided alignment, *i.e.*, ROI features. As shown in Tab. 4, although GA is a naive unlearning strategy, it still achieves observable performance gains. Meanwhile, SoftCLIP's self-similarity modeling fails to excavate supervision from false positives, which may explain why it performs worse than GA in some cases, *e.g.*, CC12M with ViT-B/32. By contrast, our NCU achieves solid improvements by forgetting both false positives and false negatives, outperforming SoftCLIP by $1.1\% \sim 2.3\%$ without external guidance.

5.4. Ablation Study

To investigate the effectiveness of specific components in our method, we carry out some ablation studies on Ima-

Dataset	Model	Caltech101	CIFAR-10	CIFAR-100	DTD	Aircraft	SST2	Flowers102	Food101	GTSRB	OxfordPets	RESISC45	SUN397	EuroSAT	StanfordCars	STL10	**Average**	**ImageNet1K**
Model Architecture: ViT-B/16																		
CC3M	CLIP	52.3	**55.2**	24.1	10.9	1.0	**50.1**	11.9	11.1	6.9	12.9	19.5	25.0	13.5	0.8	81.7	25.1	16.0
	NCU	**59.1**	54.3	**28.8**	**12.3**	**1.1**	**50.1**	**14.1**	**14.8**	**7.4**	**16.3**	**22.8**	**32.3**	**21.7**	**1.5**	**86.3**	**28.2**$\uparrow_{3.1}$	**20.0**$\uparrow_{4.0}$
CC12M	CLIP	77.0	66.5	38.3	21.2	2.5	47.7	**33.4**	51.9	7.3	64.2	39.0	44.7	21.2	25.5	91.4	42.1	40.6
	NCU	**80.9**	**79.3**	**49.1**	**23.2**	**2.7**	**48.0**	31.7	**52.7**	**10.1**	**66.5**	**41.9**	**52.6**	**28.6**	**29.0**	**93.2**	**46.0**$\uparrow_{3.9}$	**43.4**$\uparrow_{2.8}$
Model Architecture: ViT-B/32																		
CC3M	CLIP	47.7	54.2	18.0	7.6	1.2	**50.1**	9.3	9.1	6.0	7.4	16.2	16.0	15.5	0.8	77.7	22.5	11.8
	NCU	**53.0**	**56.7**	**25.9**	**10.4**	**1.7**	**50.1**	**10.2**	**10.5**	**6.5**	**10.5**	**19.0**	**22.2**	**16.7**	**1.4**	**80.1**	**25.0**$\uparrow_{2.5}$	**14.6**$\uparrow_{2.8}$
CC12M	CLIP	76.3	68.2	35.2	16.1	**2.8**	50.1	**29.3**	37.6	6.4	54.1	30.1	39.2	22.5	14.8	90.8	38.2	33.8
	NCU	**80.4**	**68.5**	**41.4**	**19.3**	2.6	**52.8**	28.6	**43.4**	**7.2**	**62.4**	**35.7**	**48.3**	**31.3**	**18.1**	**92.5**	**42.2**$\uparrow_{4.0}$	**36.7**$\uparrow_{2.9}$
YFCC15M-R	CLIP	53.7	67.0	34.4	13.1	1.1	49.3	22.1	18.6	11.0	13.5	20.3	29.3	23.0	1.7	83.7	29.5	17.8
	NCU	**58.2**	**69.5**	**37.8**	**15.3**	**1.8**	**49.9**	**29.2**	**23.7**	**11.2**	**16.1**	**23.1**	**34.0**	**23.3**	**1.8**	**86.7**	**32.1**$\uparrow_{2.6}$	**21.9**$\uparrow_{4.1}$

Table 1. Zero-shot transfer evaluation of different models.

Dataset	Architecture	Model	Flickr30K 1K Testing							MSCOCO 5K Testing						
			Image-to-Text			Text-to-Image			**Average**	Image-to-Text			Text-to-Image			**Average**
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
CC3M	ViT-B/16	CLIP	27.6	54.2	65.8	19.0	40.5	51.3	43.1	12.8	30.9	42.7	9.7	25.4	35.2	26.1
		NCU	**32.1**	**60.7**	**71.9**	**25.2**	**49.5**	**61.0**	**50.1**$\uparrow_{7.0}$	**15.8**	**36.9**	**48.4**	**12.1**	**29.7**	**40.1**	**30.5**$\uparrow_{4.4}$
	ViT-B/32	CLIP	14.0	35.2	47.7	11.5	27.9	37.8	29.0	7.0	20.1	28.9	6.0	16.7	24.0	17.1
		NCU	**21.3**	**44.5**	**55.7**	**15.7**	**36.3**	**46.4**	**36.7**$\uparrow_{7.7}$	**10.7**	**25.7**	**35.4**	**8.1**	**21.3**	**30.2**	**21.9**$\uparrow_{4.8}$
CC12M	ViT-B/32	CLIP	50.3	77.2	**85.9**	37.9	64.8	74.2	65.1	26.8	54.0	65.9	**20.0**	42.4	54.2	43.9
		NCU	**53.0**	**77.3**	85.0	**38.4**	**66.6**	**76.7**	**66.2**$\uparrow_{1.1}$	**28.2**	**54.7**	**66.7**	**20.0**	**42.8**	**54.3**	**44.5**$\uparrow_{0.6}$
YFCC15M-R	ViT-B/32	CLIP	57.4	81.6	89.1	40.8	66.4	75.4	68.5	**35.0**	60.9	71.8	22.6	46.0	58.0	49.1
		NCU	**58.0**	**83.2**	**89.7**	**42.5**	**69.9**	**78.8**	**70.4**$\uparrow_{1.9}$	34.3	**62.0**	**73.5**	**24.6**	**49.0**	**60.8**	**50.7**$\uparrow_{1.6}$

Table 2. Zero-shot cross-modal retrieval evaluation of different models.

Dataset	Architecture	Model	SUN397	OxfordPets	Food101	ImageNet
CC3M	ViT-B/16	CLIP	54.38	62.20	54.36	48.13
		NCU	**55.60**	**62.85**	**54.74**	**49.90**
	ViT-B/32	CLIP	46.96	52.30	**46.96**	40.20
		NCU	**48.20**	**52.82**	46.85	**41.49**
CC12M	ViT-B/16	CLIP	70.94	82.83	78.99	**66.70**
		NCU	**71.36**	**84.76**	**79.18**	66.65
	ViT-B/32	CLIP	66.05	78.41	70.12	59.19
		NCU	**66.63**	**78.63**	**70.76**	**60.34**
YFCC15M-R	ViT-B/32	CLIP	60.46	61.90	59.09	51.07
		NCU	**60.75**	**62.85**	**60.46**	**52.29**

Table 3. Linear probing comparison of different models.

Dataset	Model	Model Architecture	ImageNet1K ZS top-1
CC3M	CLIP	ViT-B/16	16.0
	Gradient Ascent		16.7 $\uparrow_{0.7}$
	SoftCLIP		18.9 $\uparrow_{2.9}$
	NCU		**20.0** $\uparrow_{4.0}$
CC3M	CLIP	ViT-B/32	11.8
	Gradient Ascent		12.1 $\uparrow_{0.3}$
	SoftCLIP		13.3 $\uparrow_{1.5}$
	NCU		**14.6** $\uparrow_{2.8}$
CC12M	CLIP	ViT-B/16	40.6
	Gradient Ascent		41.6 $\uparrow_{1.0}$
	SoftCLIP		42.1 $\uparrow_{1.5}$
	NCU		**43.4** $\uparrow_{2.8}$
CC12M	CLIP	ViT-B/32	33.8
	Gradient Ascent		35.1 $\uparrow_{1.3}$
	SoftCLIP		34.4 $\uparrow_{0.6}$
	NCU		**36.7** $\uparrow_{2.9}$

Table 4. Zero-shot top-1 performance on ImageNet1K.

geNet1K with models unlearned on CC3M. We first ablate the contributions of two key components of NCU, *i.e.*, negative prompt and text relation opposite. Specifically, in the variant $\mathcal{V}_1$, we replace the learnable prompt tokens by prepending some textual negative prefixes to raw captions, *e.g.*, *'the image has no'* or *'this picture lacks'*. In the variant $\mathcal{V}_2$, we use maximal L_2 distance loss [43] as a substitute for our text relation opposite. Besides, we validate the impact of different NC unlearning by intervening with the refined alignment, *i.e.*, $\mathcal{V}_3$ and $\mathcal{V}_4$. As shown in Tab. 5, we observe that: 1) Using negative textual prefixes also shows competitive results, demonstrating the generalization of our method. However, we argue that the learnable prompt is preferable, except for performance gains, operating to features makes

Model	ImageNet1K ZS top-1	
	ViT-B/16	ViT-B/32
NCU	**20.0**	**14.6**
$\mathcal{V}_1$ (w/o Hardest Negative Prompts)	19.3 $\downarrow_{0.7}$	14.4 $\downarrow_{0.2}$
$\mathcal{V}_2$ (w/o Text Relation Opposite)	18.8 $\downarrow_{1.2}$	13.9 $\downarrow_{0.7}$
$\mathcal{V}_3$ (w only False Negatives Unlearning)	17.7 $\downarrow_{2.3}$	13.5 $\downarrow_{1.1}$
$\mathcal{V}_4$ (w only False Positives Unlearning)	19.2 $\downarrow_{0.8}$	14.3 $\downarrow_{0.3}$

Table 5. Ablation studies on zero-shot transfer task (ImageNet1K) of models unlearned on CC3M.

it possible for NCU to extend to modalities beyond text. 2) Simply maximizing the distance between negative and original text embeddings leads to suboptimal performance, which aligns with our analysis in Fig. 2. 3) Both types of NC impair CLIP's performance, among which the false positive causes a more severe impact. While NCU achieves the best performance by forgetting such noisy knowledge.

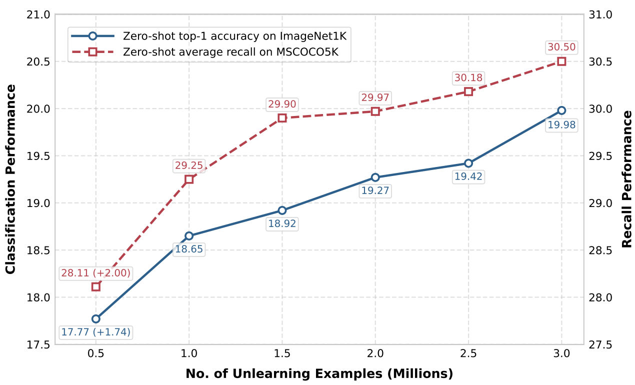
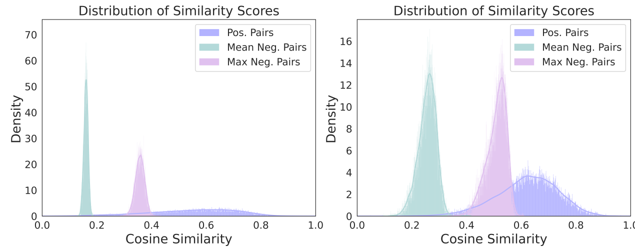


Figure 4. Effect of NCU with varying fine-tuning dataset sizes on zero-shot image classification and cross-modal retrieval.

5.5. NC Unlearning with Partial Data

In this section, we conduct an interesting study to verify whether CLIP can improve robustness by only unlearning NC with a portion of the pre-trained data. To this end, given a CLIP pre-trained on CC3M as the reference model, we evaluate NCU using different data portions ranging from 0.5 million to 3 million image-text pairs. Fig. 4 plots the ZS top-1 accuracy on ImageNet1K and the average of recalls on MSCOCO 5K. Remarkably, even when unlearning on less than 20% of the original data (0.5M), NCU achieves significant performance gains while preserving overall knowledge learned in CC3M. With the accessible data increasing, NCU shows consistent improvements on both zero-shot downstream tasks. This phenomenon indicates NCU's flexibility in enhancing the robustness of models with limited data, which is valuable to handling VLMs pre-trained with partially private or proprietary data.



(a) NCU (b) CLIP

Figure 5. Similarity scores distribution of positive and negative pairs from CLIP and NCU. Both models are based on ViT/B16 and learning from the CC3M training set.

5.6. Visualization and Analysis

To intuitively show the robust embedding space that is refined by our approach, we plot the distribution of normalized similarity for CLIP and NCU on the validation set of CC3M. In Fig. 5, we illustrate similarity scores for positive pairs, mean of negative pairs, and top 5% maximum of negative pairs. First, we observe that NCU produces a wider distribution of positive similarity scores, capturing more fine-grained matching degrees among positive pairs. Second, NCU improves the feature discrimination, which leads to a more significant separation between positive and negative pairs. Lastly, NCU provides more appropriate measures for hard negatives, which maintains separation from both positive and other negative pairs.

6. Limitations and Future Works

Our work still has certain limitations due to the finite computing capability, including 1) This work only uses CLIP to explore the efficacy of NCU. Further research is needed to confirm its applicability in other VLMs, such as BLIP-2 [28], and even larger VLMs like VisionLLM [45] or InternVL [5]. 2) The current experiments are mainly conducted on million-scale data, and we plan to extend it to larger-scale datasets to verify NCU's generalization.

7. Conclusion

This work provides a new thinking in robust vision-language learning. Instead of re-training models from scratch, we suggest eliminating the harmful effects of noisy correspondence from pre-trained models. To this end, we propose NCU, a robust fine-tuning framework that efficiently unlearns noisy correspondence in CLIP. Our key concept is to learn the hardest negative information that can provide explicit unlearning direction to resist both FP and FN. We formalize such twin goals unlearning process into one unified OT problem for fast fine-tuning. Extensive experiments are conducted to verify that NCU can endow CLIP with strong robustness against noisy correspondence.

References

[1] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441, 2022. 1, 2

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 2

[3] Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Fff: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182, 2024. 1, 2

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 6

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 8

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 5

[7] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024. 2

[8] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6

[9] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12043–12051, 2024. 2

[10] Zhongtian Fu, Kefei Song, Luping Zhou, and Yang Yang. Noise-aware image captioning with progressively exploring mismatched words. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12091–12099, 2024. 2

[11] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1860–1868, 2024. 1, 2, 6

[12] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022. 1, 2

[13] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 1, 2

[14] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. Rwkv-clip: A robust vision-language representation learner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4799–4812, 2024. 2, 6

[15] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. *Advances in Neural Information Processing Systems*, 35:14972–14985, 2022. 5

[16] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26679–26688, 2024. 2

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2

[18] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3

[19] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 926–934, 2023. 2

[20] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 1, 2

[21] Zhenyu Huang, Mouxing Yang, Xinyan Xiao, Peng Hu, and Xi Peng. Noise-robust vision-language pre-training with positive-negative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2

[23] Liang Jiang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Robust domain adaptation for machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8060–8069, 2023. 2

[24] Wooyoung Kang, Jonghwan Mun, Sungjun Lee, and Byungseok Roh. Noise-aware learning from web-crawled image-text data for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2942–2952, 2023. 2

[25] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *Proceedings of the IEEE/CVF conference on computer vi-*

sion and pattern recognition, pages 9442–9451, 2021. 3

[26] Alexey Kravets and Vinay Namboodiri. Zero-shot class unlearning in clip with synthetic samples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2

[27] Alexey Kravets and Vinay P. Namboodiri. Zero-shot clip class forgetting via text-image space adaptation. *Transactions on Machine Learning Research*, 2025. 2

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 8

[29] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024. 3, 4

[30] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Xi Peng, et al. Multi-granularity correspondence learning from long-term noisy videos. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[31] Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[32] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431, 2022. 2

[33] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 2

[34] Maitreya Patel, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, Yezhou Yang, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[35] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. 2

[36] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[38] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021. 1, 2

[39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1, 6

[40] Liangliang Shi, Jack Fan, and Junchi Yan. Ot-clip: Understanding and generalizing clip via optimal transport. In *Forty-first International Conference on Machine Learning*, 2024. 2

[41] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022. 1

[42] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2, 6

[43] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 3, 4, 7

[44] Jue Wang, Haofan Wang, Weijia Wu, Jincan Deng, Yu Lu, Xiaofeng Guo, and Debing Zhang. Eclip: Efficient contrastive language-image pretraining via ensemble confidence learning and masked language modeling. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. 2

[45] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023. 8

[46] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E Gonzalez, and Peter Vajda. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *International Conference on Learning Representations*, 2022. 2

[47] Mouxing Yang, Zhenyu Huang, and Xi Peng. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, pages 1–19, 2024. 2

[48] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. 2

[49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 4

[50] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024. 2

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3