Pattern Recognition 179 (2026) 113729



Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr





Boosting adversarial transferability of vision-language pre-trained models via optimal transport

Simeng Qin [a,b,c], Gang Li [a,b,c], Sensen Gao [d], Dongchen Han [e], Xiaojun Jia [f,*], Yang Bai [g], Jindong Gu [h], Xiaochun Cao [e,*]

[a] Northeastern University, Shenyang, China
[b] Hebei Key Laboratory of Data Science and Knowledge Management, China
[c] Key Laboratory of Financial Risk Intelligent Warning Prevention and Control, China
[d] Mohammed Bin Zayed University of Artificial Intelligence, United Arab Emirates
[e] Shenzhen Campus of Sun Yat-sen University, China
[f] Nanyang Technological University, Singapore
[g] ByteDance, USA
[h] University of Oxford, United Kingdom

ARTICLE INFO

Keywords:
Vision-language pre-training
Adversarial transferability
Multi-modal adversarial attack
Optimal transport

ABSTRACT

Vision-language pre-trained models (VLP) excel in processing both image and text data but are vulnerable to multi-modal adversarial examples. Notably, adversarial examples created for a specific model can deceive other models as well, a phenomenon known as adversarial transferability. This transferability presents a significant threat to practical applications, leading to increased interest in understanding and improving it. Recent works have indicated that leveraging data augmentation and image-text modal interactions can significantly enhance the transferability of adversarial examples for VLP models. However, these methods overlook the critical issue of finding the optimal alignment between data-augmented image-text pairs. This oversight generates adversarial examples overly customized to the source model, resulting in adversarial examples tailored to the source model and limiting their transferability. In this paper, we investigate the relationship between augmented image sets and their corresponding text sets, finding that augmented images align better with some texts and less with others. Based on this insight, we propose an Optimal Transport-based Adversarial Attack (OT-Attack), which formulates the features of image and text sets as two distinct distributions and employs optimal transport theory to determine the most efficient mapping between them, improving adversarial transferability. Extensive experiments across various network architectures and datasets demonstrate that the proposed method can effectively improve adversarial transferability and outperform state-of-the-art adversarial attack methods.

1. Introduction

Vision-Language Pre-trained (VLP) models have shown outstanding performance in various downstream tasks, including image-text matching [1], image captioning [2], visual question answering [3], and visual grounding [4]. As VLP models become foundational to modern multimodal applications and are increasingly embedded in practical systems, ensuring their robustness and security is critical—not only for reliable engineering performance, but also for broader societal impact.

Existing works have demonstrated that adversarial examples perturbed on white-box models remain effective on certain black-box models [5]. It indicates that adversarial examples generated via a proxy model can still mislead the prediction of black-box models due to their transferability [6,7]. An ideal attack scenario, in reality, is one where adversarial examples remain effective even in the absence of detailed knowledge about the model's inner workings, such as its model architecture, weights, and gradients, etc. In practice, the target VLP model is often only accessible via an API (black-box), so its parameters are unavailable. Meanwhile, attackers or auditors can usually train a functionally similar surrogate VLP model, generating adversarial examples on the surrogate and then transfer them to black-box targets. Moreover, improving adversarial transferability is practically important for (i) pre-deployment acceptance testing, (ii) third-party auditing and

* Corresponding authors.
E-mail addresses: qinsimeng@neuq.edu.cn (S. Qin), ligang@neuq.edu.cn (G. Li), Sensen.Gao@mbzuai.ac.ae (S. Gao), han.dongchen@mail.com (D. Han), jiaxiaojunqaq@gmail.com (X. Jia), baiyang0522@gmail.com (Y. Bai), jindong.gu@outlook.com (J. Gu), caoxiaochun@mail.sysu.edu.cn (X. Cao).

https://doi.org/10.1016/j.patcog.2026.113729
Received 26 October 2025; Received in revised form 3 March 2026; Accepted 10 April 2026
Available online 16 April 2026
0031-3203/© 2026 Published by Elsevier Ltd.

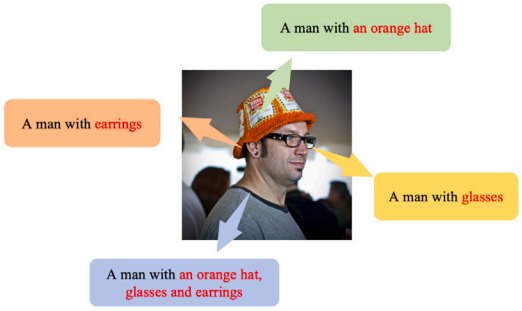


Fig. 1. An image example from the Flickr30K dataset where different captions tend to focus on different image contents.

compliance, and (iii) reliable defense development, making adversarial transferability particularly critical for VLP models in real-world multimodal systems. Consequently, understanding and enhancing the transferability of adversarial examples is particularly important for VLP models, given their widespread deployment in real-world multimodal systems and the prevalence of black-box access settings.

Motivated by the practical significance of transfer-based adversarial attacks and adversarial transferability, in this paper, we primarily study the transferability of adversarial examples across VLP models. Zhang et al. propose Co-Attack [8], which combines modalities using image-text pairs to improve transferability. And, Lu et al. develop the Set-level Guidance Attack (SGA) [9], advancing Co-Attack by employing data augmentation and multiple textual descriptions for set-level alignment and intermodal guidance to improve adversarial transferability. Further, Gao et al. propose expanding the diversity of adversarial examples by introducing diversification in the overlap region of adversarial trajectories (DRA) [10], achieving state-of-the-art results in improving adversarial transferability for VLP. However, as illustrated in Fig. 1, different captions of the same image may focus on different contents. This implies that a particular caption might exhibit a high degree of relevance to a specific region of the image while demonstrating limited correlation with other regions. Thus, previous attack methods adopt a uniform standard to assess this relationship to generate adversarial examples, which is inadequate. The primary limitation lies in averaging the alignment between sets of captions and images, without considering the crucial correspondence between specific captions and their corresponding image contents. This generalized matching strategy fails to ensure optimal alignment, especially after images have undergone data augmentation processes such as zooming, which can lead to significant misalignments with their captions. Specifically, as illustrated in Fig. 2, different captions focus on different fine-grained regions of the image (e.g., "a man with glasses", "a man with earrings", "a man with an orange hat"). After applying view augmentations (e.g., multi-scale resizing), a specific augmented view may preserve some attributes more clearly, while altering other attributes (thus aligning weakly with captions referring to those regions). Consequently, this average alignment may reduce the efficacy of data augmentation and modality interactions for improving adversarial transferability.

To address this issue, in this paper, we incorporate the theory of optimal transport [11] into the image-text matching loss to generate multimodal adversarial examples. Specifically, we treat the feature sets of augmented images and texts as two distinct distributions and aim to establish the optimal transport scheme between them. The distinction between our method and DRA, along with a comparative overview of the results, is shown in Fig. 2. In detail, we integrate optimal transport theory to analyze data-augmented image sets and text sets as distinct distributions. This holistic consideration allows us to incorporate similarity into the cost matrix and calculate the optimal transport scheme. Consequently, we compute the total transfer cost between these distributions, guiding the generation of adversarial examples. Our method achieves a more balanced matching relationship between the augmented image and text sets, leading to more effective alignment and improving the transferability of adversarial examples. Experiments conducted on various models including ALBEF [12], TCL [13], and CLIP [14], and utilizing well-known datasets like Flickr30K [15] and MSCOCO [16]. We also conduct a series of experiments on three vision-and-language downstream tasks: image-text retrieval (ITR), visual grounding (VG), and image captioning (IC). The experimental results indicate that our OT-Attack can significantly improve the transferability of multimodal adversarial examples, outperforming existing state-of-the-art attack methods. The key contributions of the work are summarized in three aspects:

- We propose OT-Attack that improves adversarial transferability of VLPs by ensuring a balanced match between image and text sets after data augmentation.
- We innovatively utilize Optimal Transport theory in boosting adversarial transferability in VLP models, promoting a more profound and thorough alignment between data-augmented images and textual descriptions.
- Extensive experiments establish that our method generates adversarial examples with superior transferability compared to existing state-of-the-art attack methods. Furthermore, our OT-Attack can successfully break current commercial models like GPT-4 and Bing Chat.

2. Background and related work

Vision-Language Pre-trained Models (VLP) are a pivotal technique in augmenting multimodal task performance, capitalizing on extensive pretraining with image-to-text pairs. Traditionally, much of the research in this area has relied on pre-trained object detectors, using region features to create vision-language representations. However, the advent of Vision Transformer (ViT) [17,18] has instigated a methodological shift. Increasingly, studies advocate adopting ViT in image encoding, which involves an end-to-end process of transforming inputs into patches. VLP models can be broadly classified into fused and aligned VLP models. Fused VLP models, as exemplified by architectures like ALBEF [12] and TCL [13], utilize individual unimodal encoders for processing token and visual feature embeddings. These models then employ a multimodal encoder to amalgamate image and text embeddings, crafting comprehensive multimodal representations. Conversely, aligned VLP models use unimodal encoders to process image and text modality embeddings independently.

Vision-Language Tasks. *Image-text retrieval* (ITR) [1,19] is a task that retrieves relevant instances from a database using one modality (image or text) to query the other. It splits into image-to-text retrieval (TR) and text-to-image retrieval (IR). Models like ALBEF and TCL calculate semantic similarity scores between image-text pairs for initial ranking, then employ a multimodal encoder for final ranking. Conversely, models like CLIP [14] directly rank based on similarity in an unimodal embedding space, showcasing varied ITR methodologies. *Image captioning* [2,20] generates textual captions for images and is crucial in VLP. This task requires converting visual content into coherent, contextually relevant text, which differs from image-text retrieval. *Visual grounding.* Visual Grounding [4,21] entails identifying and locating objects in an image per language descriptions, requiring precise text mapping to visual elements.

Transferability of Adversarial Examples. Transfer-based evaluation has been widely adopted to probe robustness when only surrogate models are available [5–7]. In the vision domain, prior work improves transferability through diversity-enhancing objectives and augmentation-based strategies, such as augmentation mixing and patch-wise transformations (*e.g.*, ADMIX [22], PAM [23]). For VLP models, recent studies explicitly leverage cross-modal interactions to

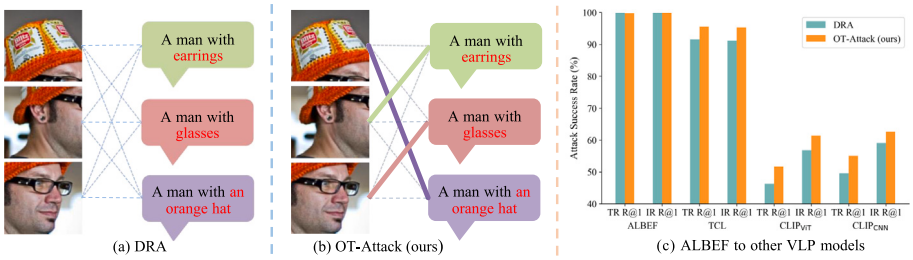


Fig. 2. Comparative analysis of Diversification Region Attack (DRA) methods and their ITR attack success rates. **(a)** illustrates the conventional DRA approach where image and text sets are averaged to establish pair-wise matches. **(b)** demonstrates OT-Attack, where images are matched to texts based on optimal transport theory to enhance matching accuracy. **(c)** indicates the attack success rates for our OT-Attack versus traditional DRA, with ALBEF serving as the source and TCL, CLIP_{ViT}, and CLIP_{CNN} serving as the target. The bar charts show that our adversarial examples outperform DRA in all metrics.

strengthen multimodal transferability. Co-Attack integrates unimodal adversarial signals from both image and text modalities to encourage cross-model transfer [8,24,25]. Building on this, SGA [9] performs set-level guidance by expanding a single image–text pair into augmented image sets and multiple textual descriptions, improving adversarial diversity and inter-modal guidance. More recently, DRA [10] and SA-AET [26] introduce trajectory-based diversification by encouraging diversity in overlapping regions of adversarial optimization paths, achieving strong transferability in VLP settings. Despite this progress, many set-based methods implicitly rely on uniform aggregation across augmented views and captions, which may under-emphasize fine-grained correspondences between specific textual descriptions and the most relevant image regions.

Optimal Transport. Optimal Transport (OT), a concept first introduced by Monge [11], its unique ability to match distributions leads to its widespread application in various theoretical and practical areas. OT provides a principled framework for matching two empirical distributions under a specified cost, and has been used broadly in distribution alignment problems across vision and language [27,28]. Compared with uniform averaging matching, OT explicitly optimizes a transport plan that captures structured correspondences between elements from two sets, making it well-suited for modeling set-to-set relationships (e.g., between augmented image features and multiple textual descriptions). Entropy-regularized OT [29] further improves scalability and differentiability, which facilitates its integration into modern learning and optimization pipelines.

3. The proposed method

3.1. Threat model

The adversarial attack methods can be divided into two scenarios: the white-box attack scenario and black-box attack scenario. In the white-box attack scenario, the adversary has complete access to the model's architecture, parameters, and gradients, which allows for direct optimization to generate adversarial examples. While in the black-box attack scenario, the adversary can only access the model's outputs or behavior. This work focuses on generating adversarial examples in the white-box scenario and leveraging these examples to attack other models in the black-box scenario. This approach is aimed at evaluating the transferability of adversarial examples and the effectiveness of attack strategies. A well-designed loss function $\mathcal{L}$ in the white-box setting plays a critical role in generating high-quality adversarial examples. In this regard, the proposed method explores the integration of optimal transport loss into adversarial attacks, which will be discussed in detail in the following sections. Given the input image $\mathbf{I}$ and the input text $\mathbf{T}$, the adversary seeks to maximize a loss function $\mathcal{L}$ on the VLP model $f(\cdot)$ under a constraint:

$$\begin{aligned} &\mathbf{I}_{adv}, \mathbf{T}_{adv} = \arg\max \ \mathcal{L}\left(f\left(\mathbf{I}_{adv}, \mathbf{T}_{adv}\right)\right), \\ &\text{subject to} \quad \mathbf{I}_{\text{adv}} \in \mathbb{B}\left[\mathbf{I}, \epsilon_v\right], \mathbf{T}_{\text{adv}} \in \mathbb{B}\left[\mathbf{T}, \epsilon_t\right], \end{aligned} \tag{1}$$

where $\mathbf{I}_{adv}$ and $\mathbf{T}_{adv}$ represent the adversarial image and text input, respectively. $\mathbb{B}[\mathbf{I}, \epsilon_v]$ represents the searching spaces for optimizing adversarial image, and ϵ_v represents the maximum range that input image can be modified. $\mathbb{B}[\mathbf{T}, \epsilon_t]$ represents the searching spaces for optimizing adversarial text, and ϵ_t represents the maximum range that input text can be modified. We use the same text attack method following [9]. The adversarial text can be formulated as:

$$\mathbf{T}_{adv} = \arg\max \left(- \frac{f_{\text{txt}}(\mathbf{T}_{adv})^{\top} f_{\text{img}}(\mathbf{I}_{adv})}{\left\| f_{\text{txt}}(\mathbf{T}_{adv}) \right\| \left\| f_{\text{img}}(\mathbf{I}_{adv}) \right\|} \right), \tag{2}$$

where $\mathbf{T}_{adv} \in \mathbb{B}[\mathbf{T}, \epsilon_t]$. For adversarial image generation, we adopt the proposed optimal transport loss to generate adversarial examples.

3.2. The proposed method

3.2.1. Symbol conventions

In this section, we describe the sources of image and text features utilized in our framework, followed by a discussion on how traditional attack methods define the loss function $\mathcal{L}$. Given an original set of images $\mathbf{I}$ and a set of image enhancement factors A, we construct the augmented image set $\mathbf{I}_{aug}$ by applying an image enhancement method f_{enh} to each image for every factor $a \in A$:

$$\mathbf{I}_{\text{aug}} = \bigcup_{a \in A} f_{\text{enh}}(\mathbf{I}, a), \tag{3}$$

where f_{enh} represents a generic image enhancement operation. Using the augmented image set $\mathbf{I}_{aug}$ and an original text set $\mathbf{T}$, we extract their feature representations via encoders. Specifically, the image encoder $f_{\text{img}}(\cdot)$ and text encoder $f_{\text{txt}}(\cdot)$ produce $\mathbf{F}_{img} = f_{\text{img}}(\mathbf{I}_{aug})$ and $\mathbf{X}_{txt} = f_{\text{txt}}(\mathbf{T})$, where $\mathbf{F}_{img}$ and $\mathbf{X}_{txt}$ are the image and text feature representations, respectively. The similarity matrix $\mathbf{S}$, representing the pairwise similarity between image and text features, is computed as:

$$\mathbf{S} = \mathbf{F}_{img} \mathbf{X}_{txt}^{\top}. \tag{4}$$

Traditional attack methods often define the loss function $\mathcal{L}$ using the similarity matrix $\mathbf{S}$. A commonly employed formulation is as follows:

$$\mathcal{L}_{ori} = -\frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{S}_{ij}, \tag{5}$$

where n and m denote the two dimensions of $\mathbf{S}$. This formulation encourages the adversarial examples to maximize dissimilarity between the features of augmented images and original texts, facilitating the generation of effective attacks on the white-box model. However, a significant limitation often hinders traditional attack methods: the generated adversarial examples tend to overfit the source (white-box) model. During optimization, the perturbations are excessively tailored to exploit the white-box model's specific features and decision boundaries. While this overfitting improves attack success on the source model, it severely reduces the transferability of adversarial examples to black-box models. This lack of transferability is a critical challenge, as it undermines the effectiveness of adversarial attacks in practical scenarios.

Algorithm 1 Sinkhorn Iteration for OT

Require: $\mathbf{C}$: cost matrix, $\mathbf{P}$: source measure, $\mathbf{Y}$: target measure, φ: convergence threshold
Ensure: $\widetilde{\Pi}$: transport matrix

$\mathbf{r} \leftarrow$ ones-like($\mathbf{P}$), $\mathbf{c} \leftarrow$ ones-like($\mathbf{Y}$) ▷ Initialize scaling vectors
$\mathbf{K} \leftarrow \exp\left(-\frac{\mathbf{C}}{\lambda}\right)$ ▷ Kernel matrix
$\phi \gg \varphi$
while $\phi > \varphi$ **do**
  $\mathbf{r}_0 \leftarrow \mathbf{r}$
  $\mathbf{r} \leftarrow \mathbf{P} \oslash (\mathbf{K}\mathbf{c})$ ▷ Row normalization
  $\mathbf{c} \leftarrow \mathbf{Y} \oslash (\mathbf{K}^\top \mathbf{r})$ ▷ Column normalization
  $\phi \leftarrow \mathbb{E}[|\mathbf{r} - \mathbf{r}_0|]$ ▷ Convergence check
end while
$\widetilde{\Pi} \leftarrow (\mathbf{r} \otimes \mathbf{c}) \times \mathbf{K}$ ▷ Recover transport plan
return $\widetilde{\Pi}$

3.2.2. Optimal transport

Defining Source (P) and Target (Y) Distributions. In the Optimal Transport framework, we begin by defining two fundamental distributions: the source distribution $\mathbf{P}$ and the target distribution $\mathbf{Y}$. These distributions represent the starting and ending points of the transportation process in the Optimal Transport problem. Specifically, the source distribution $\mathbf{P} = (p_1, p_2, \ldots, p_n)$ and the target distribution $\mathbf{Y} = (y_1, y_2, \ldots, y_m)$ describe the quantities to be transported from and to each respective location.

The Transportation Matrix Π. In Optimal Transport (OT), $\Pi \in \mathbb{R}_+^{n\times m}$ denotes the transportation plan, where Π_{ij} is the mass moved from source p_i to target y_j. It must satisfy the marginal constraints:

$$\sum_{j=1}^{m} \Pi_{ij} = p_i, \ \forall i \in \{1, \ldots, n\}, \qquad \sum_{i=1}^{n} \Pi_{ij} = y_j, \ \forall j \in \{1, \ldots, m\}, \tag{6}$$

and the non-negativity constraint $\Pi_{ij} \geq 0$.

Optimal Transport and Sinkhorn Regularization. Given the cost matrix $\mathbf{C}$ with entries $\mathbf{C}_{ij}$, representing the cost of transporting a unit from source p_i to target y_j, the OT problem is

$$OT(\mathbf{P}, \mathbf{Y}, \mathbf{C}) = \min_{\Pi \in \mathcal{U}(P,Y)} \sum_{i,j} \Pi_{ij} \mathbf{C}_{ij}, \tag{7}$$

where $\mathcal{U}(P, Y)$ is the set of feasible plans satisfying the above marginals. To improve scalability, we adopt the entropic OT (Sinkhorn distance) [30]:

$$OT_\lambda(\mathbf{P}, \mathbf{Y}, \mathbf{C}) = \min_{\Pi \in \mathcal{U}(P,Y)} \sum_{i,j} \Pi_{ij} \mathbf{C}_{ij} - \lambda H(\Pi), \tag{8}$$

where $H(\Pi)$ is the entropy term and λ controls the regularization strength. Following the Sinkhorn algorithm [30], we iteratively update

$$\begin{aligned} \widetilde{\Pi} &= \operatorname{diag}(\mathbf{P}^t) \exp\left(-\frac{\mathbf{C}}{\lambda}\right) \operatorname{diag}(\mathbf{Y}^t), \\ \text{where} \quad \mathbf{P}^t &= \mathbf{P} \oslash \left(\exp\left(-\frac{\mathbf{C}}{\lambda}\right) \mathbf{Y}^{t-1}\right), \\ \mathbf{Y}^t &= \mathbf{Y} \oslash \left(\left(\exp\left(-\frac{\mathbf{C}}{\lambda}\right)\right)^T \mathbf{P}^{t-1}\right), \end{aligned} \tag{9}$$

until convergence $\mathbb{E}[|\mathbf{P}^t - \mathbf{P}^{t-1}|] < \varphi$. Hence, the Sinkhorn algorithm iteratively normalizes the rows and columns of the transport matrix to satisfy the marginal constraints while minimizing the regularized objective function [30]. Here, $H(\Pi)$ is the entropy of the transport matrix, introducing regularization (controlled by λ) to ensure numerical stability and efficient computation. The complete Sinkhorn procedure used in OT-Attack is summarized in Algorithm 1.

3.2.3. Calculating loss through optimal transport

The Optimal Transport loss $\mathcal{L}_{OT}$ is computed using the feature representations of augmented images $\mathbf{F}_{img}$, original texts $\mathbf{X}_{txt}$, and the similarity matrix $\mathbf{S}$. First, the cost matrix $\mathbf{C}$ is defined as $\mathbf{C} = 1 - \mathbf{S}$, transforming similarity scores into a cost structure. Next, the exponentiated negative cost matrix $\mathbf{K}$ is computed for the Sinkhorn iterations, given by $\mathbf{K} = \exp\left(-\frac{\mathbf{C}}{\lambda}\right)$, where λ is a small positive regularization parameter. From Eq. (9), we can get the optimal transport plan $\widetilde{\Pi}$. The OT loss is calculated as:

$$\mathcal{L}_{OT} = \sum_{i,j} \widetilde{\Pi}_{ij} \mathbf{C}_{ij}, \tag{10}$$

where $\widetilde{\Pi}_{ij}$ in $\widetilde{\Pi}$ represents the optimal transport of features from the ith element in $\mathbf{F}_{img}$ to the jth element in $\mathbf{X}_{txt}$, and $\mathbf{C}_{ij}$ is the corresponding cost in $\mathbf{C}$. This formulation of $\mathcal{L}_{OT}$ captures the minimal cost required to align the feature representations of augmented images with those of the original texts. By leveraging the overall feature distribution, it facilitates the generation of more effective adversarial examples. Importantly, this method addresses potential overfitting issues inherent in relying solely on a similarity matrix as the loss metric. Hence, we employ the adversarial example generation method outlined in Eq. (10) to create adversarial examples. These samples are then used to mount attacks on black-box models. The process is shown in Algorithm 2.

Algorithm 2 Adversarial Image Generation Process

Require: $model$: source model, $\mathbf{I}$: the original image, A: adjustment factors, $\mathbf{T}$: the original text set.
Ensure: $\mathbf{I}_{adv}$: generated adversarial images

$model \rightarrow eval()$ ▷ Set model to evaluation mode
$\delta = 0$ ▷ Adversarial perturbation initialization
for $k = 1$ **to** N **do**
  $\mathbf{I}_{adv} = \mathbf{I} + \delta$
  $\mathbf{I}_{aug} \leftarrow \bigcup_{a\in A} f_{\text{enh}}(\mathbf{I}_{adv}, a)$ ▷ Construct enhanced view
  $\mathbf{F}_{img} \leftarrow f_{\text{img}}(\mathbf{I}_{aug})$, $\mathbf{X}_{txt} \leftarrow f_{\text{txt}}(\mathbf{T})$ ▷ Extract image/text features
  $\mathbf{S} \leftarrow \mathbf{F}_{img}\mathbf{X}_{txt}^\top$, $\mathbf{C} \leftarrow 1 - \mathbf{S}$ ▷ Build similarity and cost matrices
  $\widetilde{\Pi} \leftarrow \arg\min_{\Pi\in\mathcal{U}(P,Y)} \sum_{i,j} \Pi_{ij}\mathbf{C}_{ij} - \lambda H(\Pi)$ ▷ Solve OT with Sinkhorn
  $\mathcal{L}_{OT} \leftarrow \sum_{i,j} \widetilde{\Pi}_{ij}\mathbf{C}_{ij}$ ▷ Compute OT loss
  $\delta \leftarrow \text{clip}\left(\delta + \alpha \ \text{sign}(\nabla_\delta \mathcal{L}_{OT}), -\epsilon_v, \epsilon_v\right)$, ▷ Update adversarial perturbation
  $\mathbf{I}_{adv} = \mathbf{I} + \delta$ ▷ Update adversarial image
end for
return $\mathbf{I}_{adv}$

4. Experiments

4.1. Settings

VLP Models. To evaluate adversarial examples' transferability and our framework's performance, we examined two Vision-Language Pre-trained (VLP) model types: fused and aligned VLPs. Fused VLPs, like ALBEF [12] and TCL [13], process images and text together with shared layers, using a 12-layer ViT-B/16 [17] for visuals and two 6-layer transformers for image and text data. Aligned VLPs, such as CLIP [14] variants (CLIP_{ViT} with ViT-B/16 and CLIP_{CNN} with ResNet-101 [31]), process data separately before aligning it in later stages. We assessed cross-task attack success on image captioning using BLIP, with adversarial examples generated using TCL.

Datasets. For the image-text retrieval task, we use two datasets: Flickr30K [15] and MSCOCO [16]. A salient characteristic shared by both is the quintuple of descriptive captions accompanying each image, providing a valuable asset for the assessment of our image-text retrieval approach. For the task of Visual Grounding, we employed the RefCOCO+ [32] dataset, which further enriched our cross-task attack effectiveness analysis. The details of these datasets are summarized as follows:

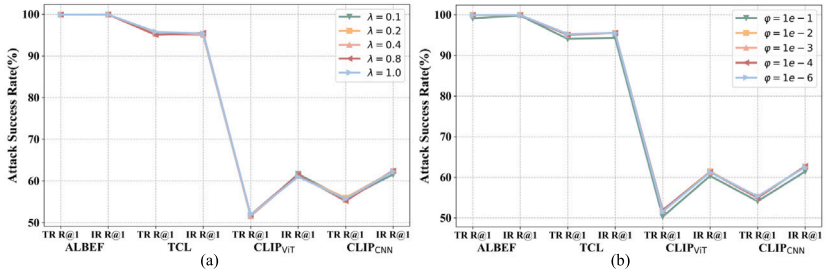


Fig. 3. Hyper-parameter selection on the Flickr30K dataset: (a) regularization coefficient λ; (b) convergence threshold φ.

- Flickr30K (ITR): we follow the common split of 29,783 train images, 1000 val images, and 1000 test images, and we conduct evaluation experiments on the 1000-image test set with 5000 captions.
- MSCOCO (ITR/IC): we follow the widely used split with 113,287 train images, 5000 val images, and 5000 test images, and we evaluate on the 5000-image test set with 25,000 captions.
- RefCOCO+ (VG): we use the standard split with 120,191 train images, 10,758 val images, 5726 testA images, and 4889 testB images, and report results on val, testA, and testB images.

Baselines. In this work involving Vision-Language Pre-trained (VLP) models, we implement several prevalent adversarial attack methods as baselines, which include PGD [24], BERT-Attack [25], Sep-Attack [8], Co-Attack [8], SGA [9], and DRA [10]. We adopt the default setting of their original paper to generate adversarial examples for comparison.

Adversarial Attack Configuration. To validate our framework's effectiveness, following SGA [9] and DRA [10], the augmentation is multi-scale resizing to enlarge the image set: the original image is resized into five scales $\{0.50, 0.75, 1.00, 1.25, 1.50\}$ using bicubic interpolation, forming an augmented image set for computing the image–text matching for adversarial example generation. We generate adversarial visual examples using the Projected Gradient Descent (PGD) method [24] with specific settings: a perturbation limit of $\epsilon_v = 8/255$, a step size of $\alpha = 2/255$, and $N = 10$ iterations. For textual examples, we use BERT-Attack [25] with a disturbance limit of $\epsilon_t = 1$ and a vocabulary size of $W = 10$. The hyper-parameters λ and φ are set to 0.1 and $1e-2$, respectively.

Evaluation Criteria. The transferability of the adversarial attacks is quantitatively assessed using the Attack Success Rate (ASR). ASR is a crucial metric that measures the proportion of successful adversarial examples out of the total number of attacks conducted. A higher ASR is indicative of increased transferability of the adversarial examples, signifying the effectiveness of the attack in compromising the model under various conditions.

4.2. Hyper-parameter selection

The proposed OT-Attack includes two hyper-parameters: regularization coefficient λ and convergence threshold φ. Following [10,26], using different hyper-parameters, we adopt ALBEF to generate adversarial examples on the Flickr30K dataset and evaluate adversarial transferability with other models. We use the 1000-image test set with 5000 captions of the Flickr30K dataset for the selection of the hyper-parameters. The results are shown in Fig. 3. Our OT-Attack with different λ can achieve similar performance of adversarial transferability, which also indicates that our method is not sensitive to the exact choice of λ. We randomly selected $\lambda = 0.1$ for subsequent experiments. Since $\varphi = 1e-2$ performs better than others. Hence, we set φ to $1e-2$.



Fig. 4. Visualization of adversarial examples from Flickr30K. In the task of image-text matching, adversarial examples for both images and texts are generated and utilized for image-to-text and text-to-image matching tasks, respectively. We highlight the distinctions in the text adversarial examples compared to the original samples and also quantified the pixel differences between the image adversarial examples and the original images.

4.3. Comparative experimental results

In our experiments, the primary focus is on Image-Text Retrieval (ITR) tasks. Adversarial examples are generated on various white-box VLP models, and their effectiveness is evaluated by measuring the attack success rates on both the white-box models and three additional black-box models.

Our analysis spans two widely recognized datasets: Flickr30K, which includes 1000 images and 5000 captions, and MSCOCO, which provides a larger set of 5000 images and 25,000 captions. This comprehensive dataset selection enables a robust evaluation of our attack methods in image-text matching tasks, effectively quantifying the ability of adversarial examples to mislead these models. The results are shown in Tables 1 and 2.

Our results show that our OT-Attack achieves significant advancements in generating adversarial examples that are not only effective within models of the same type but also demonstrate strong cross-type attack success. This is particularly evident from the R@1 success rates in TR and IR tasks, where the adversarial examples consistently maintain high effectiveness across various models, including ALBEF, TCL, CLIP_{ViT}, and CLIP_{CNN}.

For example, when using ALBEF to target TCL, our method achieves a 4.09% increase in the TR R@1 attack success rate on Flickr30K and a 3.53% increase on MSCOCO compared to the state-of-the-art results obtained by DRA. Conversely, in scenarios where TCL is employed to target ALBEF, OT-Attack outperforms DRA with improvements of 3.05% on Flickr30K and 2.62% on MSCOCO in the TR R@1 attack success rate. These results highlight the effectiveness of enhancing

Table 1
Comparison with SOTA methods on the Flickr30K dataset for the image-text retrieval task. The source column shows VLP models used to generate multimodal adversarial examples. For Image Retrieval and Text Retrieval, we provide an R@1 attack success rate.

Source	Attack	ALBEF		TCL		$CLIP_{ViT}$		$CLIP_{CNN}$	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	PGD [24]	93.74	94.43	24.03	27.90	10.67	15.82	14.05	19.11
	BERT-Attack [25]	11.57	27.46	12.64	28.07	29.33	43.17	32.69	46.11
	Sep-Attack [8]	95.72	96.14	39.30	51.79	34.11	45.72	35.76	47.92
	Co-Attack [8]	97.08	98.36	39.52	51.24	29.82	38.92	31.29	41.99
	SGA [9]	99.90	99.98	87.88	88.05	36.69	46.78	39.59	49.78
	DRA [10]	99.90	99.93	91.57	91.17	46.26	56.80	49.55	59.01
	OT-Attack (ours)	**99.79**	**99.93**	**95.57**	**95.36**	**51.66**	**61.44**	**55.04**	**62.54**
TCL	PGD [24]	35.77	41.67	99.37	99.33	10.18	16.3	14.81	21.10
	BERT-Attack [25]	11.89	26.82	14.54	29.17	29.69	44.49	33.46	46.07
	Sep-Attack [8]	52.45	61.44	99.58	99.45	37.06	45.81	37.42	49.91
	Co-Attack [8]	49.84	60.36	91.68	95.48	32.64	42.69	32.06	47.82
	SGA [9]	92.49	92.77	100.0	100.0	36.81	46.97	41.89	51.53
	DRA [10]	95.20	95.58	100.0	99.98	47.24	57.28	52.23	62.23
	OT-Attack (ours)	**98.75**	**98.13**	**99.89**	**99.98**	**53.13**	**61.37**	**55.43**	**64.01**
$CLIP_{ViT}$	PGD [24]	3.13	6.48	4.43	8.83	69.33	84.79	13.03	17.43
	BERT-Attack [25]	9.59	22.64	11.80	25.07	28.34	39.08	30.40	37.43
	Sep-Attack [8]	7.61	20.58	10.12	20.74	76.93	87.44	29.89	38.32
	Co-Attack [8]	8.55	20.18	10.01	21.29	78.53	87.50	29.50	38.49
	SGA [9]	22.42	34.59	25.08	36.45	100.0	100.0	53.26	61.10
	DRA [10]	27.84	42.84	27.82	44.60	100.0	100.0	64.88	69.50
	OT-Attack (ours)	**35.04**	**49.95**	**39.83**	**52.0**	**100.0**	**100.0**	**70.75**	**74.75**
$CLIP_{CNN}$	PGD [24]	2.29	6.15	4.53	8.88	5.40	12.08	89.78	93.04
	BERT-Attack [25]	8.86	23.27	12.33	25.48	27.12	37.44	30.40	40.10
	Sep-Attack [8]	9.38	22.99	11.28	25.45	26.13	39.24	93.61	95.30
	Co-Attack [8]	10.53	23.62	12.54	26.05	27.24	40.62	95.91	96.50
	SGA [9]	15.64	28.60	18.02	33.07	39.02	51.45	99.87	99.90
	DRA [10]	19.50	34.59	21.60	37.88	48.47	59.12	99.87	99.90
	OT-Attack (ours)	**22.42**	**35.69**	**23.92**	**39.1**	**50.18**	**61.37**	**99.62**	**99.76**

Table 2
Comparison with SOTA methods on the MSCOCO dataset for the image-text retrieval task. The source column shows VLP models used to generate multimodal adversarial examples. For Image Retrieval and Text Retrieval, we provide an R@1 attack success rate.

Source	Attack	ALBEF		TCL		$CLIP_{ViT}$		$CLIP_{CNN}$	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	PGD [24]	94.35	93.26	34.15	36.86	21.71	27.06	23.83	30.96
	BERT-Attack [25]	24.39	36.13	24.34	33.39	44.94	52.28	47.73	54.75
	Sep-Attack [8]	97.01	96.59	61.06	66.13	57.19	65.82	58.81	68.61
	Co-Attack [8]	65.22	72.41	94.95	97.87	55.28	62.33	56.68	66.45
	SGA [9]	**99.95**	**99.94**	87.46	88.17	63.72	69.71	63.91	70.78
	DRA [10]	99.90	99.93	88.81	90.06	69.25	75.31	68.53	75.09
	OT-Attack (ours)	99.87	99.87	**92.62**	**93.32**	**72.95**	**78.08**	**72.7**	**78.47**
TCL	PGD [24]	40.81	44.09	98.54	98.20	21.79	26.92	24.97	32.17
	BERT-Attack [25]	35.32	45.92	38.54	48.48	51.09	58.80	52.23	61.26
	Sep-Attack [8]	66.38	72.19	99.15	98.98	59.94	65.95	60.77	69.37
	Co-Attack [8]	49.84	60.36	91.68	95.48	32.64	42.69	32.06	47.82
	SGA [9]	92.70	92.99	100.0	100.0	59.79	65.31	60.52	67.34
	DRA [10]	94.72	95.89	**100.0**	**100.0**	70.51	74.95	70.29	76.99
	OT-Attack (ours)	**97.86**	**97.99**	99.95	99.97	**72.91**	**77.44**	**74.17**	**78.46**
$CLIP_{ViT}$	PGD [24]	10.26	13.69	12.72	15.81	82.91	90.51	21.62	28.78
	BERT-Attack [25]	20.34	29.74	21.08	29.61	45.06	51.68	44.54	53.72
	Sep-Attack [8]	25.91	36.84	28.20	38.47	88.36	97.09	47.57	57.79
	Co-Attack [8]	26.35	36.69	28.23	38.42	88.78	96.72	47.36	58.45
	SGA [9]	43.75	51.08	44.05	51.02	100.0	100.0	70.66	75.58
	DRA [10]	52.69	61.50	51.88	61.06	100.0	**100.0**	80.18	84.11
	OT-Attack (ours)	**57.7**	**66.4**	**58.99**	**66.72**	**100.0**	99.98	**83.29**	**86.7**
$CLIP_{CNN}$	PGD [24]	8.38	12.73	11.90	15.68	13.66	20.62	92.68	94.71
	BERT-Attack [25]	23.38	34.64	24.58	29.61	51.28	57.49	54.43	62.17
	Sep-Attack [8]	29.13	40.64	31.40	42.99	52.23	59.73	96.16	97.54
	Co-Attack [8]	29.49	41.50	31.83	43.44	53.15	60.15	97.79	98.54
	SGA [9]	36.94	46.79	38.81	48.90	62.19	67.73	**99.92**	**99.97**
	DRA [10]	41.40	52.25	43.62	54.15	70.43	74.14	99.80	99.92
	OT-Attack (ours)	**41.94**	**53.37**	**45.48**	**56.64**	**71.23**	**75.69**	99.39	99.53

Table 3
Computational cost (minutes) compared with DRA.

Method	ALBEF	TCL	$CLIP_{ViT}$	$CLIP_{CNN}$
DRA	**77**	**76**	**31**	**21**
OT-Attack (ours)	108	107	42	27

adversarial transferability. Complementing the numerical analysis, Fig. 4 provides a detailed visual representation of the impact of adversarial examples. It contrasts the original images and texts with their perturbed counterparts, illustrating how even subtle, nearly imperceptible modifications can significantly degrade a model's performance in image-text matching tasks. The visual differences, particularly the nuanced texture alterations introduced in the adversarial images, are highlighted through difference masks, making the deceptive potency of these adversarial examples more evident. These perturbations effectively mislead VLP systems by subtly altering key features that influence their decision-making processes. Following the default setting of SGA [9], we also adopt 1000 images from the Flickr30K dataset for our experiments. To evaluate performance, we compute the computational cost (minutes) and compare the OT-Attack with DRA across four models (ALBEF, TCL, $CLIP_{ViT}$, and $CLIP_{CNN}$). The results are shown in Table 3. OT-Attack consistently requires more time than DRA, which is approximately 1.29-−1.40 times that of SGA, reflecting the added complexity of the proposed method.

4.4. Cross-task transferability

4.4.1. Visual grounding

To thoroughly evaluate the effectiveness of our adversarial attack strategies, the RefCOCO+ [32] dataset is utilized, specifically curated for visual grounding tasks. This dataset comprises various subsets designed to assess different aspects of model performance, including:

- **RefCOCO+ val**: Offers a broad range of scenarios for a comprehensive evaluation.
- **RefCOCO+ testA**: Focuses on the model's ability to identify and localize human figures, testing its precision in distinguishing human subjects within images.
- **RefCOCO+ testB**: Targets the model's efficacy in recognizing and localizing non-human elements, challenging the model's versatility beyond human-centric tasks.

By leveraging the diverse testing scenarios provided by RefCOCO+, we aim to demonstrate the broad adaptability and transferability of our method across a wide array of visual grounding challenges, highlighting its potential for robust performance in varied contexts. The quantitative analysis in Table 4 evaluates the effectiveness of adversarial examples against the ALBEF model pre-trained on the Visual Grounding Task, using ALBEF pre-trained on the Image-Text Retrieval Task as the source (following the setting in SGA [9] and DRA [10]). The baseline scores, representing unmodified examples, establish the study's benchmark. The OT-Attack strategy outperforms DRA, reducing ALBEF's scores by 1.02 in Val, 1.05 in TestA, and 0.88 in TestB, demonstrating our method's superior disruption of visual grounding. Additionally, Fig. 5 provides a visual analysis using the Flickr30K dataset, illustrating how minor perturbations significantly impair object recognition and localization in the ALBEF model, highlighting the impact of adversarial attacks on model reliability.

4.4.2. Image captioning

In our research, adversarial examples are generated using the ALBEF model [33] to target the BLIP framework in a white-box scenario. BLIP is recognized for its advanced multi-modal encoder–decoder structure, trained on a diverse dataset with synthetic captions and noise reduction techniques. Our experiments are conducted on the MSCOCO dataset, examining both original and adversarially altered images. To evaluate the impact of adversarial actions, a set of metrics designed for image captioning tasks is utilized, including BLEU [34], METEOR [35], ROUGE [36], CIDEr [37], and SPICE [38]. These metrics evaluate various aspects of caption quality, from precision and semantic accuracy to recall, uniqueness, relevance, and the depiction of properties.

The metrics used in our study offer varied insights into the text quality and relevance, providing a view of the adversarial effects, as shown in Table 4. Our OT-Attack, compared to DRA, demonstrates lower scores across metrics: BLEU-4 decreased by 3.6, METEOR by 2.1, ROUGE-L by 3.3, CIDEr by 13.4, and SPICE by 2.0. These score reductions suggest our method's higher cross-task attack efficacy, with more significant decreases indicating better performance. Fig. 6 shows original versus adversarial image-caption pairs. These comparisons starkly exhibit how minor perturbations can drastically alter the model's interpretation, deviating from the intended meaning, thus highlighting our findings' practical significance. Further delving into the realm of large-scale models, the experiments are designed with specific parameters to gauge the extent of adversarial impact. We set the perturbation intensity at a subtle yet effective level of 16/255 and run our adversarial process for 500 iterations.

To evaluate the broader applicability and transfer effectiveness of our attacks, we test them on both closed-source and open-source MLLMs, including GPT-4, GPT-4o, Bing Chat, Claude−4.5-Sonnet, and Qwen3-VL-8B. Following [10,26], we adopt a CLIP-based ranking criterion to quantify transfer success. Specifically, for each adversarial image, we query the target MLLM with a fixed prompt: "Please describe this image, limited to 50 words". We then rank the generated description together with 100 irrelevant texts using CLIP similarity. If the generated description is not ranked first, we count it as a successful transfer attack. We randomly sample images from Flickr30K and generate adversarial examples using SGA [9], DRA [10], and our OT-Attack. The transfer results are summarized in Table 5. On the 100-image setting (covering GPT-4, Bing Chat, Qwen3-VL-8B, and Claude−4.5-Sonnet), OT-Attack consistently achieves the highest ASR, reaching 24% on GPT-4 and 30% on Bing Chat, and further improving to 18% on Qwen3-VL-8B and 15% on Claude−4.5-Sonnet, compared to SGA (6%/13%/8%/7%) and DRA (16%/22%/14%/11%). To further verify robustness at a larger scale, we additionally evaluate on 1000 images. OT-Attack remains superior across all tested models, achieving 21.6% on GPT-4, 16.4% on GPT-4o, 17.5% on Qwen3-VL-8B, and 13.6% on Claude−4.5-Sonnet, substantially outperforming SGA (5.3%/5.6%/7.3%/7.1%) and DRA (14.7%/12.5%/14.9%/10.4%). These results demonstrate that OT-guided reweighting yields consistently stronger cross-model transferability, and the advantage persists when scaling the evaluation from 100 to 1000 samples. The findings, illustrated in Fig. 7, reveal a notable level of success in our adversarial attacks, with these models showing susceptibility to being misled.

4.5. More comparative experiments

Comparison of Similarity-based Loss Variants: we conduct more experiments by comparing plain similarity-based losses (DotSim, CosSim, and KL) with their OT-augmented counterparts under the same DRA framework. As shown in Table 6, the analyses are as follows. First, without OT, DotSim and CosSim perform very similarly, especially on cross-model transferability. For example, on $CLIP_{ViT}$, DotSim achieves 46.26/56.80 while CosSim achieves 46.28/56.35; on $CLIP_{CNN}$, DotSim gives 49.55/59.01 and CosSim gives 50.08/59.34. This indicates that simply changing the similarity function does not produce a meaningful gain. Second, adding OT consistently improves all base losses, and the improvement is particularly evident on transfer models. For instance, DotSim + OT improves over DotSim from 46.26/56.80 to 51.66/61.44 on $CLIP_{ViT}$, and from 49.55/59.01 to 55.04/62.54 on $CLIP_{CNN}$. Likewise, CosSim + OT improves over CosSim from 46.28/56.35 to 52.27/61.18 on $CLIP_{ViT}$, and from 50.08/59.34

Table 4
Cross-task transferability. We utilize ALBEF to generate multi-modal adversarial examples for attacking both Visual Grounding (VG) on the RefCOCO+ dataset and Image Captioning (IC) on the MSCOCO dataset. The baseline represents the performance of each task without any attack, where a lower value indicates better effectiveness of the adversarial attack for both tasks.

Attack	ITR → VG			ITR → IC				
	Val ↓	TestA ↓	TestB ↓	B@4 ↓	METEOR ↓	ROUGE-L ↓	CIDEr ↓	SPICE ↓
Clean	58.46	65.89	46.25	39.7	31.0	60.0	133.3	23.8
SGA [9]	50.56	57.42	40.66	28.0	24.6	51.2	91.4	17.7
DRA [10]	49.70	56.32	40.54	27.2	24.2	50.7	88.3	17.2
OT-Attack (ours)	**48.68**	**55.27**	**39.66**	**23.6**	**22.1**	**47.4**	**74.9**	**15.2**



Fig. 5. Visualization results for Visual Grounding. We employed TCL as the source model and ALBEF as the target model, with images and captions sourced from the Flickr30K dataset. The adversarial examples exhibit limited visual differences from the original samples; however, they disrupt the model's judgment of visual elements in the Visual Grounding task. Compared to clean data, the localization results for the same elements may have shifted or dispersed. The visualizations of Visual Grounding vividly demonstrate the disruptive impact of adversarial examples on the model.

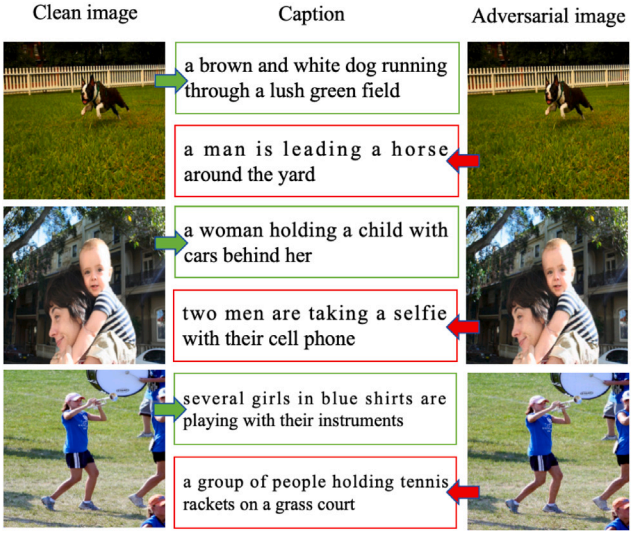


Fig. 6. Comparison of Clean and Adversarial Image Captions.

to 56.07/62.68 on CLIP_{CNN}. Similar trends are also observed for the KL-based variant. These results show that the improvement is mainly brought by the proposed OT-based alignment mechanism, rather than by a specific choice of similarity function such as dot-product or cosine similarity.

Comparison of Different Attack Hyper-parameters: we conduct a sensitivity study on the key attack hyper-parameters and summarize the results in Table 7. Hence, the proposed OT-Attack is reasonably stable under moderate changes of the hyper-parameters, and the trends are consistent with standard iterative attacks: stronger budgets or more iterations generally lead to stronger transferability. Increasing the visual perturbation limit ϵ_v can improve transfer performance. For example, when attacking from ALBEF, raising ϵ_v from 4/255 to 16/255 improves TR@1 on TCL from 94.94 to 96.73, and on CLIP_{CNN} from 55.43 to 57.47; Increasing the text perturbation limit ϵ_t can substantially strengthen adversarial transferability. For example, when attacking from ALBEF, raising ϵ_t from 1 to 2 improves TR@1 on CLIP_{ViT} from 51.66 to 65.28, and on CLIP_{CNN} from 55.04 to 68.71; Results are relatively stable around the default the vocabulary size $W = 10$. Varying W within $\{5, 10, 20\}$ leads to moderate changes overall. For instance, on TCL, TR@1 changes from 94.42 ($W = 5$) to 95.57 ($W = 10$) and 95.89 ($W = 20$); on CLIP_{CNN}, TR@1 changes from 48.91 ($W = 5$) to 55.04 ($W = 10$) and 55.43 ($W = 20$); Results are stable around the default step size α=2/255. Using a smaller step (1/255) reduces transfer noticeably, while a larger step (3/255) can improve transfer to some targets (e.g., CLIP_{ViT} TR@1 51.66 to 57.42); Increasing the number of iterations N also can strengthen adversarial transferability. Compared with $N = 5$, the default $N = 10$ yields higher transfer, and $N = 20$ provides further gains (e.g., TCL TR@1 95.57 to 95.89, CLIP_{CNN} IR@1 62.54 to 63.23).

Table 5
Attack transferability performance on open-source and closed-source MLLMs.

Models	GPT-4	Bing	Qwen3-VL-8B	Claude-4.5-Sonnet
100 Samples				
SGA [9]	6%	13%	8%	7%
DRA [10]	16%	22%	14%	11%
OT-Attack (ours)	**24%**	**30%**	**18%**	**15%**
Models	GPT-4	GPT-4o	Qwen3-VL-8B	Claude-4.5-Sonnet
1000 Samples				
SGA [9]	5.3%	5.6%	7.3%	7.1%
DRA [10]	14.7%	12.5%	14.9%	10.4%
OT-Attack (ours)	**21.6%**	**16.4%**	**17.5%**	**13.6%**

Comparison of OT-Loss Variant: We conduct experiments on the Flickr30K with more OT-Loss variants, including Partial OT and Unbalanced OT, evaluated under the same attack budget ($\epsilon_v = 8/255$, $\alpha = 2/255$, $N = 10$). As shown in Table 8, the results demonstrate that OT-based matching consistently enhances transferability, and also

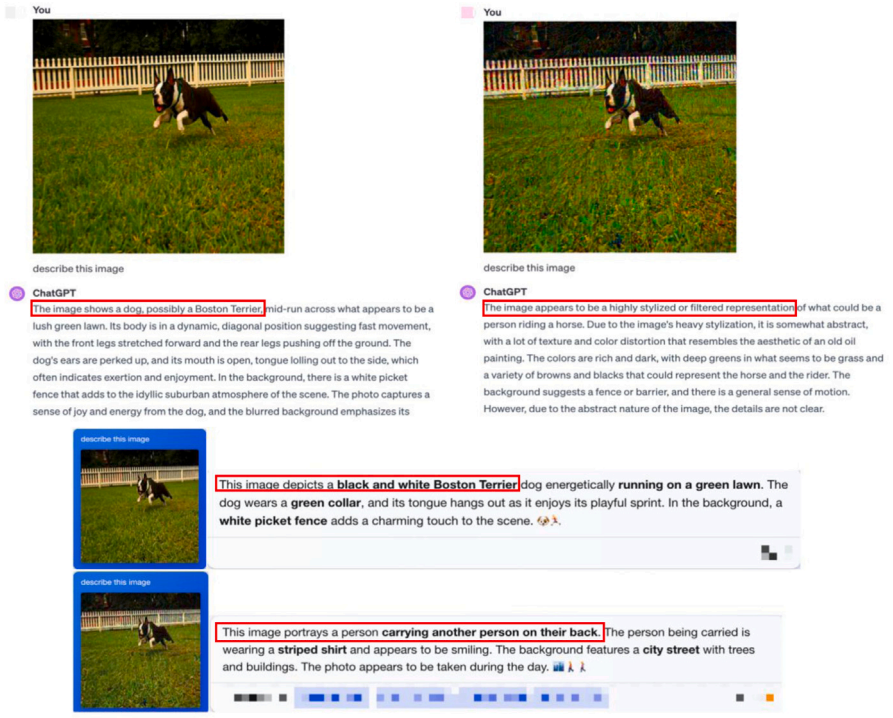


Fig. 7. Impact of Adversarial Attacks on GPT-4 and Bing Chat Descriptions. This figure showcases the alterations in image descriptions by GPT-4 and Bing Chat before and after adversarial attacks. Original descriptions are compared to those generated from manipulated images, with increased perturbation strength and iteration count to mislead the AI models. The stark contrast in the outputs highlights the susceptibility of these models to adversarial examples, reflecting the effectiveness of the perturbations in altering the perceived content of the images.

Table 6
Comparison of similarity-based loss variants for adversarial attacks. The perturbation budget is $\epsilon_v = 8/255$, with step size 2/255 for 10 steps. We report text retrieval (TR R@1) and image retrieval (IR R@1) under attacks crafted on ALBEF and evaluated on ALBEF, TCL, CLIP_{ViT}, and CLIP_{CNN}.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
					Similarity-Based Loss				
ALBEF	Dot Sim (DRA)	99.90	99.93	91.57	91.17	46.26	56.80	49.55	59.01
	Dot Sim (DRA) + OT	**99.79**	**99.93**	**95.57**	**95.36**	**51.66**	**61.44**	**55.04**	**62.54**
	Cos Sim	99.78	99.92	91.25	91.82	46.28	56.35	50.08	59.34
	Cos Sim + OT	**99.79**	**99.91**	**95.26**	**94.81**	**52.27**	**61.18**	**56.07**	**62.68**
	KL Loss	84.47	85.71	77.18	84.32	39.87	53.94	44.03	56.38
	KL Loss + OT	**91.09**	**89.86**	**81.93**	**88.79**	**44.72**	**58.16**	**48.57**	**61.05**

reveal that certain OT variants can further improve performance in some target-model settings. This indicates additional room for exploration in designing a stronger OT strategy, which we will investigate in future work. Moreover, we additionally include several structured matching methods beyond uniform averaging for comparative experiments, including Top-k hard matching and Softmax-attention weighted aggregation. These comparisons were designed to clarify whether the observed gain comes merely from moving away from uniform averaging, or from the OT-based matching mechanism itself. As shown in Table 9, the experiment results show that replacing uniform averaging with a simple structured weighting strategy can indeed bring some improvement over the DRA (Uniform) baseline. However, these gains are still limited and do not match the performance of OT-Attack, especially on cross-model transfer. For example, on CLIP_{ViT}, Softmax Attention improves over Uniform from 46.26/56.80 to 47.47/57.74, while OT-Attack further improves the results to 51.66/61.44. On CLIP_{CNN}, the corresponding numbers are 49.55/59.01 (Uniform), 51.01/60.11 (Softmax Attention), and 55.04/62.54 (OT-Attack). Similar trends are also observed on TCL. These results suggest that the improvement is not simply due to using non-uniform aggregation, but rather comes from the stronger structured correspondence enforced by OT-based matching.

Comparison of The Latest Method: We conduct more comparison experiments on the Flickr30K beyond DRA by including a more recent transfer attack method SA-AET [26], and additionally evaluate a combined variant SA-AET + OT-Attack. As shown in Table 10, the results show that integrating our OT module into SA-AET yields consistent additional gains over both DRA and SA-AET across multiple target models, supporting that the proposed OT-based alignment loss is complementary and plug-and-play rather than tied to a single baseline.

4.6. Analysis of the effectiveness of OT-attack

Prior VLP transfer attacks typically aggregate image–text alignment by uniformly averaging over augmented images and multiple texts, implicitly assuming consistent image–text correspondence. However, common augmentations (e.g., multi-scale resizing) often break this assumption, creating many semantically mismatched image–text pairs that inject noisy, surrogate-specific gradients and thus weaken adversarial transferability. The proposed OT-Attack addresses this by reformulating image–text matching as an optimal transport problem and using an OT-guided alignment loss to perform globally constrained

Table 7
Ablation settings of attack hyperparameters for OT-Attack. We study different configurations of visual and textual perturbation parameters. The original OT-Attack setting uses $\epsilon_v = 8/255$, $\alpha = 2/255$, $N = 10$ for visual attacks, and $\epsilon_t = 1$, $W = 10$ for textual attacks. Only rows corresponding to the original configuration are populated with results.

	Setting	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
	(1) Visual perturbation limit ϵ_v								
ALBEF	$\epsilon_v = 4/255$	**99.79**	**99.93**	94.94	94.93	51.53	60.66	55.43	63.84
	$\epsilon_v = 8/255$	**99.79**	**99.93**	95.57	95.36	51.66	61.44	55.04	62.54
	$\epsilon_v = 16/255$	**99.79**	99.86	**96.73**	**95.67**	**53.37**	**62.08**	**57.47**	**63.88**
	(2) Step size α								
ALBEF	$\alpha = 1/255$	99.48	99.63	86.93	87.98	46.38	56.15	48.66	57.7
	$\alpha = 2/255$	**99.79**	**99.93**	95.57	95.36	51.66	61.44	55.04	62.54
	$\alpha = 3/255$	**99.79**	99.91	**97.26**	**97.02**	**57.42**	**63.72**	**57.47**	**66.07**
	(3) Number of iterations N								
ALBEF	$N = 5$	99.27	99.56	88.09	89.24	49.33	58.76	51.09	59.66
	$N = 10$	**99.79**	99.93	95.57	95.36	51.66	61.44	55.04	62.54
	$N = 20$	**99.79**	**99.97**	**95.89**	**95.83**	**52.67**	**62.47**	**55.9**	**63.23**
	(4) Textual perturbation limit ϵ_t								
ALBEF	$\epsilon_t = 1$	99.79	99.93	**95.57**	95.36	51.66	61.44	55.04	62.54
	$\epsilon_t = 2$	**100.0**	**99.95**	95.05	**96.29**	**65.28**	**73.81**	**68.71**	**76.36**
	(5) Vocabulary size W								
ALBEF	$W = 5$	**100.0**	99.91	94.42	94.81	49.69	57.18	48.91	60.03
	$W = 10$	99.79	**99.93**	95.57	95.36	51.66	61.44	55.04	62.54
	$W = 20$	99.79	99.65	**95.89**	**96.14**	**54.48**	**64.01**	**55.43**	**64.84**

Table 8
A variety of attempts with different variations of OT-Loss. The ϵ_v is set to 8/255, with a per-step size of 2/255, and a total of 10 steps.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	OT-Attack (ours)	99.79	99.93	95.57	95.36	51.66	61.44	55.04	62.54
	Partial OT	99.9	99.91	95.57	95.24	52.02	61.4	**56.07**	62.61
	Unbalanced OT	**99.9**	**99.95**	**95.79**	**95.52**	**52.88**	**61.69**	55.3	**63.02**

Table 9
Comparing structured aggregation baselines against OT-based matching under the same attack budget. We evaluate Uniform aggregation (DRA), Top-k hard matching, Softmax-attention weighted aggregation, and OT-Attack. The perturbation constraint is $\epsilon_v = 8/255$, with step size 2/255 for 10 steps.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	DRA (Uniform)	**99.90**	99.93	91.57	91.17	46.26	56.80	49.55	59.01
	Top-K	98.73	98.90	88.43	87.15	41.27	52.19	44.36	55.27
	Softmax Attention	99.53	99.67	92.91	92.22	47.47	57.74	51.01	60.11
	OT-Attack (ours)	99.79	**99.93**	**95.57**	**95.36**	**51.66**	**61.44**	**55.04**	**62.54**

Table 10
Comparison with the latest method SA-AET [26]. We evaluate the adversarial transferability of SA-AET and a variant that incorporates OT-Attack into SA-AET.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	DRA	99.90	99.93	91.57	91.17	46.26	56.80	49.55	59.01
	OT-Attack (ours)	99.79	99.93	95.57	95.36	51.66	61.44	55.04	62.54
	SA-AET	99.90	99.98	96.42	96.02	55.58	63.89	57.22	65.59
	SA-AET+OT-Attack	**99.91**	**99.98**	**96.73**	**96.41**	**57.71**	**65.37**	**60.03**	**68.14**

many-to-many reweighting, emphasizing compatible pairs while down-weighting mismatches. Moreover, overfitting adversarial examples to the source model can significantly reduce the attack transferability. To quantify the risk of overfitting, we leverage the PAC-Bayes theorem to measure the information stored in the network's weights (IIW) [39], a promising indicator of generalization ability. Lower IIW values indicate reduced overfitting risks. For each AE generated by SGA or our method during optimization iterations, we compute its IIWs by feeding it into four VLPs. We evaluate the IIWs of 1000 AEs throughout the optimization process and present the averaged results in Fig. 8. During optimization, the IIW of AEs from the SOTA baseline (SGA) initially decreases but then sharply rises. In contrast, our method maintains consistently low IIW values for generated AEs, effectively mitigating overfitting risks. Consequently, our method enhances attack transferability.

We also provide a theoretical justification from two complementary angles: (i) OT-based matching is *provably no worse* than uniform set-level averaging in the alignment objective under the same marginal constraints, hence suppressing augmentation-induced mismatches in a principled way; and (ii) under the adversarial transferability view, reducing mismatch-induced gradient noise increases cross-model gradient alignment, which improves transferability.

(i) OT yields no worse matching than uniform averaging. Let $S \in \mathbb{R}^{n\times m}$ be the similarity matrix between the n augmented views and

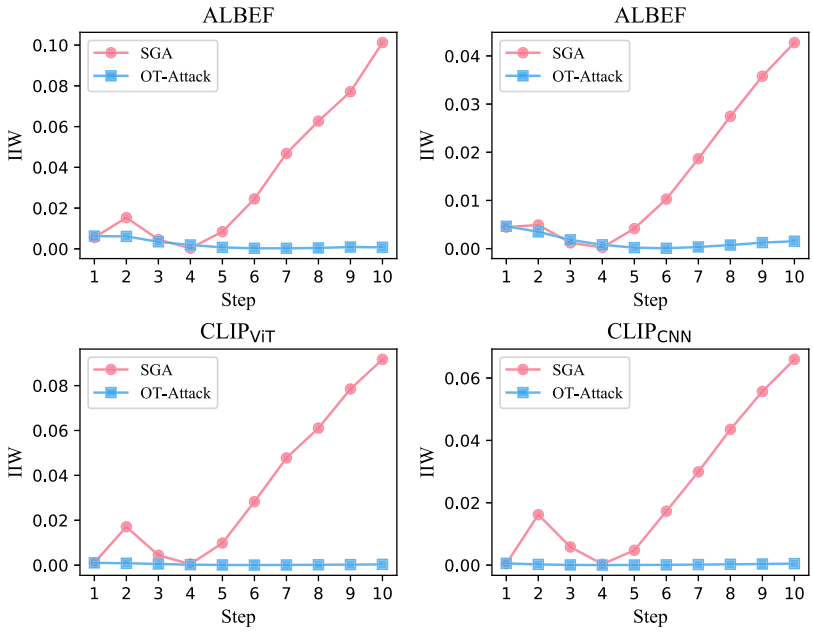


Fig. 8. Overfitting analysis of SGA & our OT-Attack via IIWs.

m texts, and define the cost $C = \mathbf{1} - S$. Let $P \in \Delta^n$ and $Y \in \Delta^m$ be fixed marginals, and denote the coupling feasible set by

$$\mathcal{U}(P,Y) = \left\{ \Pi \in \mathbb{R}_+^{n\times m} \;\middle|\; \Pi \mathbf{1}_m = P,\ \Pi^\top \mathbf{1}_n = Y \right\}. \tag{11}$$

Uniform averaging corresponds to the independent coupling

$$\Pi_{\text{avg}} = PY^\top, \quad (\Pi_{\text{avg}})_{ij} = P_i Y_j, \tag{12}$$

which assigns equal weights to all view–text pairs (up to a constant factor). Our OT-Attack computes the entropic OT (Sinkhorn) coupling

$$\Pi_e = \underset{\Pi\in\mathcal{U}(P,Y)}{\arg\min}\ J_\lambda(\Pi), \qquad J_\lambda(\Pi) \triangleq \langle \Pi, C\rangle - \lambda H(\Pi),$$
$$H(\Pi) = -\sum_{i,j} \Pi_{ij} \log \Pi_{ij}. \tag{13}$$

Proof (detailed). Step 1 (Optimality). Since $\Pi_{\text{avg}} \in \mathcal{U}(P,Y)$ is feasible and Π_e minimizes $J_\lambda(\cdot)$ over $\mathcal{U}(P,Y)$,

$$J_\lambda(\Pi_e) \le J_\lambda(\Pi_{\text{avg}}) \iff \langle \Pi_e, C\rangle - \lambda H(\Pi_e) \le \langle \Pi_{\text{avg}}, C\rangle - \lambda H(\Pi_{\text{avg}}), \tag{14}$$

which yields

$$\langle \Pi_e, C\rangle \le \langle \Pi_{\text{avg}}, C\rangle + \lambda\big(H(\Pi_e) - H(\Pi_{\text{avg}})\big). \tag{15}$$

Step 2 (Maximum-entropy of the independent coupling). Among all couplings with fixed marginals, the independent coupling $\Pi_{\text{avg}} = PY^\top$ maximizes entropy:

$$H(\Pi) \le H(\Pi_{\text{avg}}), \quad \forall\, \Pi \in \mathcal{U}(P,Y). \tag{16}$$

This follows from the identity $H(X,Y) = H(X) + H(Y) - I(X;Y)$, where $I(X;Y) \ge 0$ and $I(X;Y) = 0$ iff $X \perp Y$ (i.e., $\Pi = PY^\top$). Hence $H(\Pi_e) - H(\Pi_{\text{avg}}) \le 0$.

Step 3 (Combine). Plugging the above into (15) gives

$$\langle \Pi_e, C\rangle \le \langle \Pi_{\text{avg}}, C\rangle, \quad \Leftrightarrow \quad \langle \Pi_e, S\rangle \ge \langle \Pi_{\text{avg}}, S\rangle, \tag{17}$$

i.e., OT achieves no worse (and typically better) view–text matching than uniform averaging under the same marginals.

(ii) Better matching improves adversarial transferability. Transferability in a black-box target model depends on how well the perturbation direction learned on the surrogate aligns with the target gradients. Under uniform averaging, many augmentation-induced mismatched view–text pairs contribute to the loss and inject mismatch-induced noise into the gradient, making it more surrogate-specific and reducing cross-model gradient alignment. In contrast, the guarantee above shows that OT re-weights the set objective toward more compatible pairs and down-weights mismatches, thereby producing a cleaner, more semantically consistent gradient signal that is more likely shared across VLP models, which improves transferability.

5. Limitations

The proposed method still leaves considerable room for exploration. For example, the augmentation strategy could be expanded beyond the current multi-scale resizing protocol, and more adaptive alignment costs could be investigated to further improve robustness. Although our method achieves strong transfer attack performance, there is still room to improve transferability to more heterogeneous black-box VLP models by integrating additional transferability-enhancing strategies (e.g., ensemble-based surrogates). In addition, the proposed OT-based alignment introduces extra computation compared to simple set averaging; future work could explore more efficient approximations to reduce runtime while maintaining transferability.

6. Conclusion

We propose an Optimal Transport-based Adversarial Attack (OT-Attack) for VLPs to generate adversarial examples, which formulates the features of image and text sets as two distinct distributions and leverages optimal transport theory to determine the most effective mapping between them. By using mutual similarity as the cost matrix, OT-Attack derives an optimal mapping that guides the generation of adversarial examples, significantly enhancing adversarial transferability. Extensive experiments across various network architectures and datasets demonstrate that OT-Attack achieves superior transferability compared to SOTA attack methods. Notably, OT-Attack excels in cross-task attacks like image captioning and visual grounding and can compromise commercial models such as GPT-4 and Bing.

CRediT authorship contribution statement

Simeng Qin: Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology,

Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Gang Li:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Investigation, Data curation. **Sensen Gao:** Writing – review & editing, Visualization, Validation, Software, Investigation, Data curation. **Dongchen Han:** Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation. **Xiaojun Jia:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Data curation. **Yang Bai:** Visualization, Validation, Software, Resources, Project administration, Investigation, Data curation. **Jindong Gu:** Writing – review & editing, Visualization, Validation, Resources, Formal analysis, Conceptualization. **Xiaochun Cao:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Investigation, Formal analysis.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, We only use Generative AI to improve the language and grammar of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China General Project (No. 72371067), was supported by the National Social Science Fund of China General Project (No. 23BJY214), and was also supported by the 2026 Liaoning Provincial Economic and Social Development Research Project (Youth Unfunded Project, Project No.20261slqnwzzkt-033).

Data availability

Data will be made available on request.

References

[1] W. Li, S. Yang, Q. Li, X. Li, A.-A. Liu, Commonsense-guided semantic and relational consistencies for image-text retrieval, IEEE Trans. Multimed. (2023).
[2] T. Ghandi, H. Pourreza, H. Mahyar, Deep learning approaches on image captioning: A review, ACM Comput. Surv. 56 (3) (2023) 1–39.
[3] L. Fan, X. Gong, C. Zheng, X. Tan, J. Li, Y. Ou, Cycle-VQA: A cycle-consistent framework for robust medical visual question answering, Pattern Recognit. 165 (2025) 111609.
[4] Z. Yang, K. Kafle, F. Dernoncourt, V. Ordonez, Improving visual grounding by encouraging consistent gradient-based explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19165–19174.
[5] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint arXiv:1412.6572.
[6] Y. Xiao, J. Zhou, K. Chen, Z. Liu, Revisiting the transferability of adversarial examples via source-agnostic adversarial feature inducing method, Pattern Recognit. 144 (2023) 109828.
[7] J. Liu, W. Zhang, H. Fang, W. Zhou, E.-C. Chang, S.-K. Lam, Improving adversarial transferability and imperceptibility with loss landscape and diffusion model, Pattern Recognit. 170 (2026) 112076.
[8] J. Zhang, Q. Yi, J. Sang, Towards adversarial attack on vision-language pre-training models, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5005–5013.
[9] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, F. Zheng, Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 102–111.
[10] S. Gao, X. Jia, X. Ren, I. Tsang, Q. Guo, Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory, in: European Conference on Computer Vision, Springer, 2024, pp. 442–460.
[11] C. Villani, et al., Optimal Transport: Old and New, vol. 338, Springer, 2009.
[12] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Adv. Neural Inf. Process. Syst. 34 (2021) 9694–9705.
[13] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, J. Huang, Vision-language pre-training with triple contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15671–15680.
[14] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[15] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2641–2649.
[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[18] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 87–110.
[19] M. Cao, S. Li, J. Li, L. Nie, M. Zhang, Image-text retrieval: A survey on recent research and development, 2022, arXiv preprint arXiv:2203.14713.
[20] M.Z. Hossain, F. Sohel, M.F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Comput. Surv. (CsUR) 51 (6) (2019) 1–36.
[21] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, M. Tan, Visual grounding via accumulated attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7746–7755.
[22] X. Wang, X. He, J. Wang, K. He, Admix: Enhancing the transferability of adversarial attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16158–16167.
[23] J. Zhang, J.-t. Huang, W. Wang, Y. Li, W. Wu, X. Wang, Y. Su, M.R. Lyu, Improving the transferability of adversarial samples by path-augmented method, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8173–8182.
[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2017, arXiv preprint arXiv:1706.06083.
[25] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, Bert-attack: Adversarial attack against bert using bert, 2020, arXiv preprint arXiv:2004.09984.
[26] X. Jia, S. Gao, Q. Guo, S. Qin, K. Ma, Y. Huang, Y. Liu, I. Tsang, X. Cao, Semantic-aligned adversarial evolution triangle for high-transferability vision-language attack, IEEE Trans. Pattern Anal. Mach. Intell. (2025).
[27] Q. Yu, Z. Zeng, Y. Yan, L. Ying, R. Srikant, H. Tong, Joint optimal transport and embedding for network alignment, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 2064–2075.
[28] K. Zhou, F. Wu, Y. Lv, B. Zhang, Optimal transport-based prompt alignment for unsupervised domain adaptation, in: International Conference on Intelligent Computing, Springer, 2025, pp. 340–351.
[29] G. Li, Y. Chen, Y. Huang, Y. Chi, H.V. Poor, Y. Chen, Fast computation of optimal transport via entropy-regularized extragradient methods, SIAM J. Optim. 35 (2) (2025) 1330–1363.
[30] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, Adv. Neural Inf. Process. Syst. 26 (2013).
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[32] L. Yu, P. Poirson, S. Yang, A.C. Berg, T.L. Berg, Modeling context in referring expressions, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 69–85.
[33] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.
[34] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[35] S. Banerjee, A. Lavie, et al., An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/Or Summarization, 2005, pp. 65–72.
[36] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
[37] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[38] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part V 14, Springer, 2016, pp. 382–398.

[39] Z. Wang, S. Huang, E. Kuruoglu, J. Sun, X. Chen, Y. Zheng, PAC-BAYES information bottleneck, 2022, Publisher Copyright: © 2022 ICLR 2022 - 10th International Conference on Learning Representations. All rights reserved.; 10th International Conference on Learning Representations, ICLR 2022 ; Conference date: 25-04-2022 Through 29-04-2022.