

OT-Attack: Enhancing Adversarial Transferability of Vision-Language Models via Optimal Transport Optimization

Dongchen Han ^{1,†}, Xiaojun Jia ^{2,†}, Yang Bai ³, Jindong Gu ⁴, Yang Liu ² and Xiaochun Cao ^{1,*}

¹ Shenzhen Campus of Sun Yat-sen University

² Nanyang Technological University

³ Tsinghua University

⁴ University of Oxford

Abstract—Vision-language pre-training (VLP) models demonstrate impressive abilities in processing both images and text. However, they are vulnerable to multi-modal adversarial examples (AEs). Investigating the generation of high-transferability adversarial examples is crucial for uncovering VLP models’ vulnerabilities in practical scenarios. Recent works have indicated that leveraging data augmentation and image-text modal interactions can enhance the transferability of adversarial examples for VLP models significantly. However, they do not consider the optimal alignment problem between data-augmented image-text pairs. This oversight leads to adversarial examples that are overly tailored to the source model, thus limiting improvements in transferability. In our research, we first explore the interplay between image sets produced through data augmentation and their corresponding text sets. We find that augmented image samples can align optimally with certain texts while exhibiting less relevance to others. Motivated by this, we propose an Optimal Transport-based Adversarial Attack, dubbed OT-Attack. The proposed method 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. This optimal mapping informs our generation of adversarial examples to effectively counteract the overfitting issues. Extensive experiments across various network architectures and datasets in image-text matching tasks reveal that our OT-Attack outperforms existing state-of-the-art methods in terms of adversarial transferability.

1. Introduction

Vision-Language Pre-trained (VLP) models have shown outstanding performance in various downstream tasks, including image-text matching, image captioning, visual question answering, and visual grounding. Despite their impressive capabilities, these models encounter significant security challenges in real-world applications [1], [2], [3], [4]. Most of the prior researches [5], [6], [7], [8], [9] focused on white-box attacks, where there is complete access to the model’s

[†] The first two authors contributed equally to this work (me@handongchen.com & jiaxiaojunq@outlook.com).

* Corresponding author: caoxiaochun@mail.sysu.edu.cn

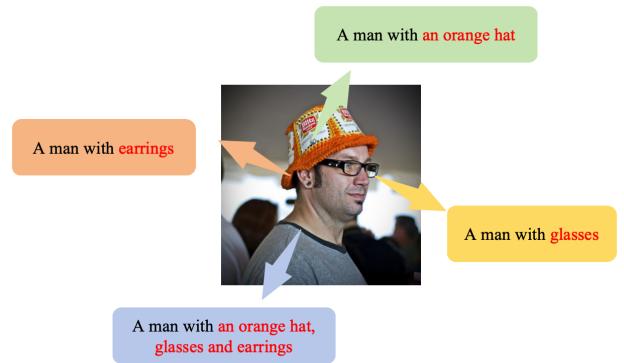


Figure 1: An image example, after undergoing various data augmentation strategies, tends to focus on different image contents. Consequently, it can better align with specific text content while maintaining limited relevance to others. Thus, using a uniform standard to assess this relationship is unsuitable, which highlights the limitation of the existing state-of-the-art SGA method.

structure, weights, and gradients. But they are not applicable to black-box models, such as GPT-4 [10].

Fortunately, existing works have demonstrated that adversarial examples perturbed on white-box models remain effective on certain black-box models [11], [12]. It indicates that adversarial examples generated via a proxy model can still mislead the judgment of black-box models due to their transferability [13], [14], [15], [16], [17], [18]. The most 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 [19], [20]. Motivated by the practical significance of black-box adversarial attacks and adversarial transferability [21], [22], [23], [24], in this paper, we primarily study the transferability of adversarial examples within Vision-Language Pre-trained (VLP) models.

Recent research in the field of VLP models has revealed that single-modality adversarial attacks, such as employing BERT-Attack [25] on text or PGD attack [26] on images,

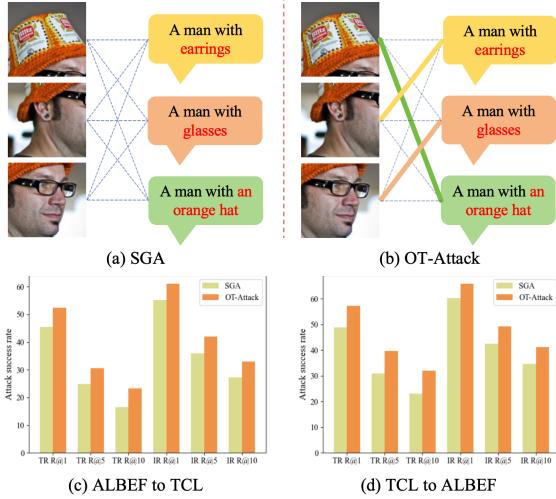


Figure 2: Comparative analysis of Set-level Guidance Attack (SGA) methods and their ITR attack success rates. Panel (a) illustrates the conventional SGA approach where image and text sets are averaged to establish pair-wise matches. Panel (b) showcases our proposed method, OT-Attack, where images are matched to texts based on optimal transport theory to enhance matching accuracy. Panels (c) and (d) depict the attack success rates for our method OT-Attack versus traditional SGA, with ALBEF and TCL models serving alternately as the source and target. **The bar charts indicate that our adversarial examples outperform SGA across all metrics**, demonstrating superior effectiveness in disrupting ITR performance.

yield limited attack success rates, even against white-box models. This insight led to the exploration of more integrated approaches. For instance, the Sep-Attack method, which simultaneously uses BERT-Attack on text and PGD on images without considering interactions between these modalities, has shown some effectiveness but lacks in fully exploiting the potential synergies of multi-modal attacks. Building upon these findings, Zhang *et al.* proposed the Co-Attack [6] method, which goes a step further by fusing information between modalities using individual image-text pairs. This approach underscored the importance of inter-modal guidance for improving the transferability of adversarial examples in VLP models. Further advancing this field, Lu *et al.* introduced the Set-level Guidance Attack (SGA) [27], which expands upon the concept of individual image-text pairings to a set-level alignment. The SGA methodology employs data augmentation to diversify the image set and pairs these images with multiple textual descriptions. This comprehensive approach, which integrates intermodal guidance, not only addresses the limitations observed in Sep-Attack but also achieves state-of-the-art results, highlighting the effectiveness in enhancing the transferability of adversarial examples against VLP models.

However, SGA has limitations in improving the transferability of adversarial examples for vision-language pre-

training models. Specifically, it aligns the features of images and texts at a collective level but does not adequately address the optimal matching of post-augmentation image examples with their corresponding texts. For example, when images undergo data augmentation, such as zooming, which enlarges certain parts of an image while omitting others, the resulting examples may not align equally well with their textual descriptions. SGA calculates its loss by averaging similarities across all pairs, which means that even if there is an optimal match for the augmented data, its influence on the final loss is obviously reduced. This process can undermine the benefits of data augmentation and modality interactions in enhancing adversarial transferability.

In this paper, we address this issue by incorporating the theory of optimal transport [28]. 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 SGA, along with a comparative overview of the results, is depicted in Figure 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 set and the text set. As depicted in Figure 1, an image example, after undergoing data augmentation, may align more closely with a particular text description. Applying optimal transport theory addresses this by balancing the effects of data augmentation and modality interactions, which in turn reduces overfitting and improves the transferability of adversarial examples.

Experiments conducted on various models including ALBEF [29], TCL [30], and CLIP [31], and utilizing well-known datasets like Flickr30K [32] and MSCOCO [33], quantitatively demonstrate the effectiveness of our approach. We focus on image-text and text-image matching as our primary downstream tasks, assessing the adversarial examples' transferability through their attack success rates on black-box models. The results exceed the current state-of-the-art on both datasets. Additionally, to assess cross-task transferability, we conduct experiments on image captioning and visual grouping, which confirm the efficacy of our OT-Attack method in varied downstream tasks. Additionally, we assess widely-used business models such as GPT-4 and Bing Chat under increased perturbation intensity, and we are able to effectively break them. Our research indicates that extra textural perturbations cause confusion in the decision-making processes of these models. The proposed method is positioned as a robust solution for uncovering and leveraging widespread vulnerabilities across the spectrum of VLP models. The key contributions of in this paper are summarized in three aspects:

- 1) Our proposed OT-Attack improves the SGA framework by ensuring a balanced match between image

- and text sets after data augmentation.
- 2) We innovatively utilize Optimal Transport theory in examining adversarial example transferability in VLP models, promoting a more profound and thorough alignment between data-augmented images and textual descriptions.
- 3) Extensive experiments establish that our method generates adversarial examples with superior transferability compared to existing state-of-the-art techniques. **Furthermore, our OT-Attack can successfully break current business models like GPT-4 and Bing Chat.**

2. Background And Related Work

2.1. Vision-Language Pre-training Models

Vision-language pre-training (VLP) [34] is a pivotal technique in augmenting multimodal task performance, capitalizing on extensive pre-training 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) [35], [36] has instigated a methodological shift. Increasingly, studies are advocating for the adoption of ViT in image encoding, which involves an end-to-end process of transforming inputs into patches. VLP models can be broadly classified into two categories: fused and aligned VLP models. Fused VLP models, as exemplified by architectures like ALBEF [29] and TCL [30], 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, such as CLIP, use unimodal encoders to independently process image and text modality embeddings.

2.2. Vision-Language Tasks

Image-text Retrieval. Image-Text Retrieval (ITR) [37], [38] 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 [31] directly rank based on similarity in an unimodal embedding space, showcasing varied ITR methodologies.

Image Captioning. Image captioning [39], [40] involves generating textual captions for images and is crucial in Vision-Language Pre-training (VLP) models. This task requires converting visual content into coherent, contextually relevant text, differing from image-text retrieval which is about finding the most relevant content from a database based on the opposite modality queries.

Visual Grounding. Visual Grounding [41], [42], also termed referential expression comprehension, is key in both computer vision and NLP. It entails identifying and locating objects or regions in an image as per language descriptions, requiring a precise mapping of text to visual elements.

2.3. Transferability of Adversarial Examples

In the field of adversarial attacks, there are two main categories: white-box and black-box approaches. White-box attacks assume full knowledge of the target model's structure and parameters, a scenario often not feasible in real-world black-box settings. In natural language processing (NLP), predominant methods typically involve textual manipulation, such as altering or replacing specific tokens, exemplified by BERT-Attack, introduced by Li *et al.* [25]. In computer vision, white-box attacks have seen significant developments with methods like the Fast Gradient Sign Method (FGSM) by Goodfellow *et al.* [11], Carlini & Wagner (C&W) by Carlini and Wagner [26], and the Momentum Iterative Method (MIM) by Dong *et al.* [43]. These methods exploit the gradient information of the models to craft effective adversarial examples. Separately, Projected Gradient Descent (PGD), proposed by Madry *et al.* [26], marks a notable advancement in the field. PGD iteratively adjusts images in small steps along the gradient of the loss function, making it a potent tool for creating subtle yet effective adversarial images. This approach has proven particularly effective in attacking the visual components of models.

Building on the foundations laid by PGD and BERT-Attack, Co-Attack, introduced by Xie *et al.* [6], emerged as a method integrating both visual and textual modalities. This approach exploits the synergies between image and text within VLP models, targeting their multimodal nature. The recent innovation in this domain is the Set-level Guidance Attack (SGA), developed by Zhang *et al.* [27]. SGA uses data augmentation to generate multiple image sets, aligning them with various text captions. This approach not only complements the multimodal essence of VLP models but also significantly enhances the transferability of adversarial examples across a spectrum of black-box models. The progression from individual-focused PGD and BERT-Attack to the integrated Co-Attack, and finally to the comprehensive SGA, illustrates an evolving landscape of adversarial strategies against VLP models.

2.4. Optimal Transport

Optimal Transport (OT), a concept first introduced by Monge [28], was initially developed to minimize logistical costs in transporting goods. In modern times, especially within the realms of machine learning and computer vision, OT has become a prominent tool due to its effectiveness in comparing and aligning distributions that are represented as feature sets [44]. Its unique ability to match distributions has led to its widespread application in various theoretical and practical areas. This includes its use in generative models and structural alignments involving sequences [45],

graphs [46], and image matching [47], [48], [49]. The versatility of OT also extends to other distribution-centric tasks like clustering, distribution estimation, and causal discovery.

3. Approach

In this section, we first introduce the treated model in this work, which consists of the adversary’s goals and the adversary’s capabilities. Then we present the observation of the previous works and the corresponding analysis. Finally, we introduce our proposed method in detail.

3.1. Threat Model

Adversary’s goals. Let us consider a white-box model M_{white} with an encoding function f_{white} and an original image I_{orig} . The aim is to determine the optimal perturbation Δ^* that maximizes the loss function \mathcal{L} when applied to I_{orig} . The perturbation is restricted to a bounded domain to ensure the adversarial example I_{adv} remains within a permissible visual deviation from I_{orig} . The optimal perturbation Δ^* is found by solving:

$$\Delta^* = \arg \max_{\Delta} \mathcal{L}(f_{\text{white}}(I_{\text{orig}})), \quad (1)$$

subject to the constraint:

$$\|\Delta\|_p \leq \epsilon, \quad (2)$$

where $\|\cdot\|_p$ represents the p -norm, and ϵ is the perturbation bound. Once Δ^* is obtained, the adversarial example is computed as:

$$I_{\text{adv}} = I_{\text{orig}} + \Delta^*. \quad (3)$$

Our goal with respect to a black-box model M_{black} characterized by an encoding function f_{black} , is to ensure that the adversarial example I_{adv} deviates from the original image I_{orig} by at least a threshold θ in the feature space of f_{black} . This criterion for successful misclassification can be defined as follows:

$$\|f_{\text{black}}(I_{\text{adv}}) - f_{\text{black}}(I_{\text{orig}})\|_2 \geq \theta, \quad (4)$$

where $\|\cdot\|_2$ denotes the Euclidean norm, signifying the measure of dissimilarity between the feature representations of the adversarial and original examples within the black-box model’s domain. This condition implies that the adversarial example x_{adv} is sufficiently distinct from the original image x_{orig} to affect the black-box model’s output, potentially leading to incorrect classification.

Adversary’s capabilities. We postulate that the adversary possesses a comprehensive understanding of the white-box model, denoted as M_{white} . This encompasses an intimate familiarity with the model’s architectural framework, its parameter configurations, and the gradients that dictate learning. In stark contrast to the transparency afforded by M_{white} , the adversary’s purview over the black-box model, M_{black} , is notably opaque. Devoid of any direct knowledge regarding the internal mechanisms, parameter values, or gradient information of M_{black} , the adversary operates without the capacity to compute internal representations or gradients.

3.2. Observation & Analysis

The Set-level Guidance Attack (SGA) method employs a comprehensive approach, focusing on sets of image-text pairs and utilizing all textual descriptions alongside a wide array of image augmentations for generating adversarial examples. This process follows a cyclic methodology, where adversarial text is initially guided by original images, followed by the generation of adversarial images influenced by this text, and concludes with the creation of further adversarial text influenced by these images.

Figure 3 demonstrates the impact of varying scaling factors on the success rate of image-text matching attacks. In this setup, ALBEF is used as the source model and TCL as the target model. The observed trend, where success rates increase initially with scaling factors but subsequently decrease, suggests that while augmentation initially improves example diversity, excessive scaling leads to overfitting.

Our analysis of this phenomenon involves two key aspects. **First**, we examine overfitting during the phase where adversarial images are created based on text. In this phase, both original images and adversarial texts are encoded, resulting in unique feature representations. The success of adversarial examples relies on how closely these features match. However, using too much data augmentation can disrupt this match, leading to adversarial examples that either miss or exaggerate certain image features. This can make the attacks less effective, particularly if they focus too much on prominent features while neglecting finer details. **Second**, to address this, our approach balances data augmentation and intermodal guidance more effectively. We aim for a better match between modified examples and texts, enhancing the adaptability of adversarial examples to different models and reducing the risk of overfitting, especially in cases where subtle or less prominent image features are involved.

3.3. The proposed Method

3.3.1 Symbol Conventions

To facilitate expression, we initially agree upon certain primary notations as shown in TABLE 1.

Symbol	Description
\mathcal{I}	Original Image Set
\mathcal{I}_{aug}	Augmented Image Set
ϕ	Image Encoder
\mathbf{F}_{img}	Features of Augmented Images
\mathcal{T}	Text Set
φ	Text Encoder
\mathbf{X}_{txt}	Text Features

TABLE 1: Symbol conventions.

Given an original set of images \mathcal{I} and a set of scaling factors \mathcal{A} , the augmented image set \mathcal{I}_{aug} can be obtained by first applying a horizontal flipping function f_{flip} to each image $I \in \mathcal{I}$, followed by a scaling function f_{scale} for each scaling factor $\alpha \in \mathcal{A}$:

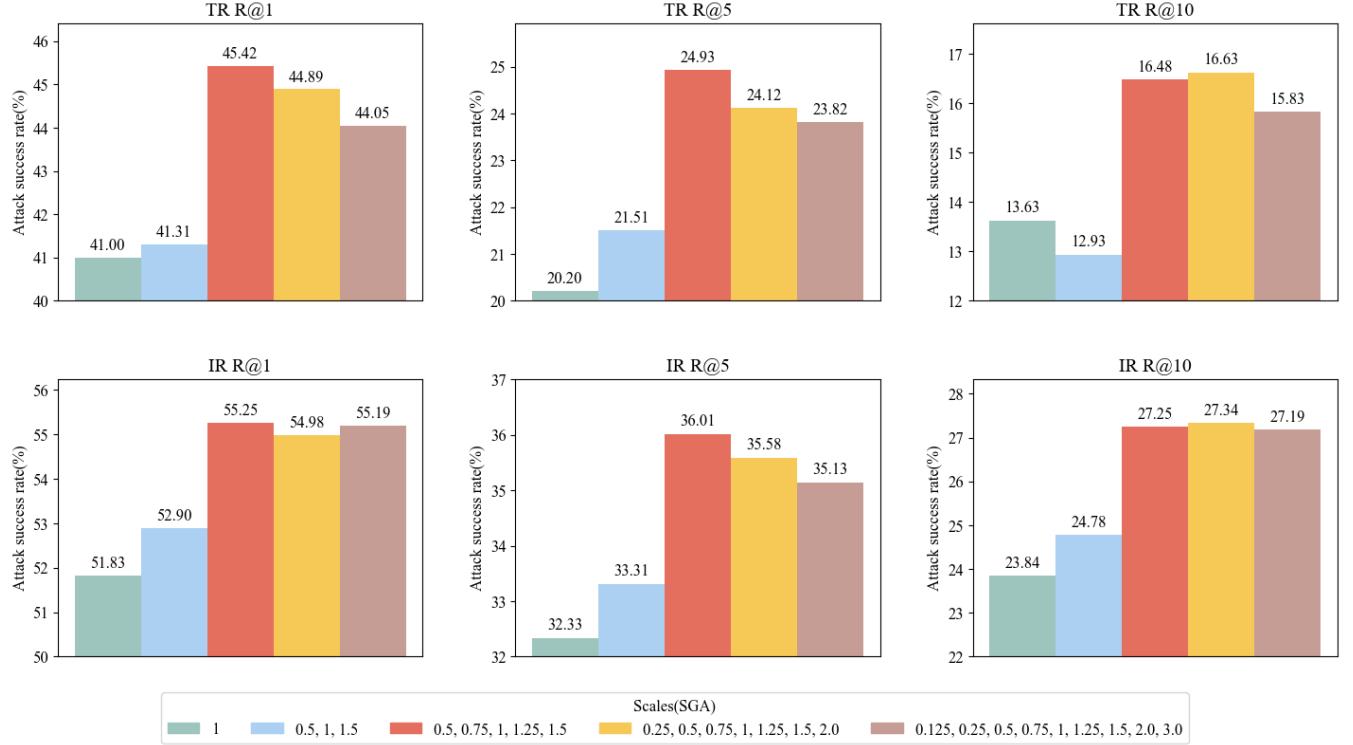


Figure 3: Utilizing the SGA method, this caption presents the attack success rates when the augmented image set, originating from the ALBEF source model and targeting the TCL model, contains 1, 3, and up to 9 images. The overall trend progresses from an increase to a decrease in success rates with the addition of examples, illustrating the effectiveness of the image set and the diminishing performance on the black-box model with an excessive number of examples.

$$\mathcal{I}_{aug} = \bigcup_{\alpha \in \mathcal{A}} (f_{scale}(f_{flip}(I), \alpha)) \quad (5)$$

where f_{flip} horizontally flips the image, and f_{scale} applies the scaling transformation to the image, for each α . Given the augmented image set \mathcal{I}_{aug} and the original text set \mathcal{T} , we apply the image encoder ϕ and the text encoder φ to obtain the corresponding feature representations:

$$\mathbf{F}_{img} = \phi(\mathcal{I}_{aug}) \quad (6)$$

$$\mathbf{X}_{txt} = \varphi(\mathcal{T}) \quad (7)$$

where \mathbf{F}_{img} represents the features of the augmented images, and \mathbf{X}_{txt} represents the features of the original texts. The function ϕ symbolizes the operation of the image encoder and φ the operation of the text encoder. Then the similarity matrix \mathbf{S} can be calculated as:

$$\mathbf{S} = \mathbf{F}_{img} \odot \mathbf{X}_{txt} \quad (8)$$

where \mathbf{S} represents the similarity matrix. The operation \odot denotes matrix multiplication, which computes the similarity between each pair of image and text features.

In previous research, the direct application of the matrix \mathbf{S} to guide the generation of adversarial examples has,

after several iterations, led to the reinforcement of the most feature-similar regions within the purview of the white-box model during the adversarial example creation process, with minimal perturbation occurring in other regions. When these adversarial examples encounter a new black-box model, the focal points of concern differ from those of the white-box model. Since the regions of interest to the black-box model have undergone only limited perturbation, they fail to disrupt the black-box model, resulting in an unsuccessful attack.

$$loss_{ori} = - \left(\sum_i \mathbf{S}_i \right)_{\text{mean}} \quad (9)$$

Here, the summation $\sum_i \mathbf{S}_i$ is taken over the last dimension of the similarity matrix \mathbf{S} , and the mean of this sum is computed to obtain the final loss value $loss_{ori}$.

3.3.2 Optimal Transport

Defining Source and Target Distributions. Initially, we define two pivotal distributions within the Optimal Transport framework: the source distribution \mathbf{P} and the target distribution \mathbf{Y} . These distributions represent the starting and ending points of the transportation journey in the Optimal Transport problem. The source distribution $\mathbf{P} = (p_1, p_2, \dots, p_n)$ and the target distribution $\mathbf{Y} = (y_1, y_2, \dots, y_m)$ encapsulate the

quantities to be transported from and to each respective location.

The Transportation Matrix \mathbf{T} . In the context of Optimal Transport, the matrix $\mathbf{T} = [T_{ij}]$ of size $n \times m$ is referred to as the transportation matrix. Each element T_{ij} of this matrix represents the amount of a commodity or resource transported from the i -th source in \mathbf{P} to the j -th target in \mathbf{Y} . The matrix \mathbf{T} effectively captures the transportation scheme between the sources and targets.

The matrix \mathbf{T} must satisfy specific constraints to ensure an optimal transportation plan. Firstly, the Marginal Constraints:

$$\sum_{j=1}^m T_{ij} = p_i \quad \forall i \in \{1, \dots, n\}, \quad (10)$$

$$\sum_{i=1}^n T_{ij} = y_j \quad \forall j \in \{1, \dots, m\}. \quad (11)$$

These constraints require that the total transported amount from each source i and to each target j matches the respective supply p_i and demand y_j .

Additionally, the Non-Negativity Constraint is imposed:

$$T_{ij} \geq 0 \quad \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, m\}. \quad (12)$$

This condition ensures that all transport amounts T_{ij} in the matrix \mathbf{T} are non-negative, reflecting the practical impossibility of negative transportation.

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

$$OT(\mathbf{P}, \mathbf{Y}, \mathbf{C}) = \min_{\mathbf{T} \in \Pi(\mathbf{r}, \mathbf{c})} \sum_{i,j} T_{ij} C_{ij} \quad (13)$$

Here, \mathbf{C} denotes the cost matrix, with each element C_{ij} representing the cost of transporting a unit from source p_i to target y_j . The matrix \mathbf{T} signifies the transportation scheme, while $\Pi(\mathbf{r}, \mathbf{c})$ encompasses all feasible transportation schemes that satisfy the marginal constraints.

The Sinkhorn distance is utilized in Optimal Transport (OT) for its effectiveness in high-dimensional spaces. Traditional OT approaches, based on linear programming, face challenges with computational intensity and scaling with data dimensionality. In contrast, the Sinkhorn distance applies entropy regularization to the OT calculation, enhancing tractability and differentiability. This approach uses a regularization parameter λ , which balances accuracy and computational efficiency. Higher λ values lead to results closer to traditional OT but at increased computational costs, while lower values of λ expedite calculations at the expense of some bias. Therefore, the Sinkhorn distance, often computed using the Sinkhorn-Knopp algorithm, presents a more feasible solution for OT in machine-learning scenarios that demand scalability and stability in computations. The Sinkhorn Optimization Process can be defined as:

$$OT_\lambda(\mathbf{P}, \mathbf{Y}, \mathbf{C}) = \min_{\mathbf{T} \in \Pi(\mathbf{r}, \mathbf{c})} \sum_{i,j} T_{ij} C_{ij} + \lambda H(\mathbf{T}) \quad (14)$$

The Sinkhorn algorithm iteratively normalizes the rows and columns of the transport matrix to satisfy the marginal constraints while minimizing the regularized objective function [50]. Here, $H(\mathbf{T})$ is the entropy of the transport matrix, introducing regularization (controlled by λ) to ensure numerical stability and efficient computation. Regarding the computation of Sinkhorn, the algorithm of the proposed OT-Attack is summarized in Algorithm 1.

Algorithm 1 Sinkhorn Iteration for Optimal Transport

Require: K : cost matrix, u : source measure, v : target measure

Ensure: T : transport matrix

```

1:  $r \leftarrow \text{ones\_like}(u)$ 
2:  $c \leftarrow \text{ones\_like}(v)$ 
3:  $thresh \leftarrow 1e-2$ 
4: for  $i = 1, \dots, 100$  do
5:    $r_0 \leftarrow r$ 
6:    $r \leftarrow u / (\text{MatMul}(K, c))$ 
7:    $c \leftarrow v / (\text{MatMul}(K^\top, r))$ 
8:    $err \leftarrow \text{Mean}(\text{Abs}(r - r_0))$ 
9:   if  $err < thresh$  then
10:    break
11:   end if
12: end for
13:  $T \leftarrow \text{Outer}(r, c) \times K$ 
14: return  $T$ 
```

3.3.3 Calculating Loss through Optimal Transport

The Optimal Transport loss loss_{OT} is computed using the feature representations of augmented images \mathbf{F}_{img} , original texts \mathbf{X}_{txt} , and the similarity matrix \mathbf{S} .

Firstly, the cost matrix \mathbf{C} is defined as $\mathbf{C} = 1 - \mathbf{S}$, where \mathbf{S} is the similarity matrix. This transformation converts similarity scores into a cost structure. Subsequently, we compute the exponentiated negative cost matrix \mathbf{K} for the Sinkhorn iterations, defined as $\mathbf{K} = \exp(-\frac{\mathbf{C}}{\lambda})$, with λ being a small positive regularization parameter. The Optimal Transport loss is calculated as:

$$\text{loss}_{OT} = \sum_{i,j} T_{ij} C_{ij} \quad (15)$$

where T_{ij} in \mathbf{T} represents the optimal 'transport' of features from the i -th element in \mathbf{F}_{img} to the j -th element in \mathbf{X}_{txt} , and C_{ij} is the corresponding cost in \mathbf{C} .

This formulation of loss_{OT} captures the minimal cost of aligning the feature representations of the augmented images with the original texts, facilitating a more effective generation of adversarial examples. In essence, this approach for computing OT loss in adversarial example generation accounts for the overall feature distribution, addressing potential overfitting issues inherent in using a similarity matrix as the sole loss metric. It ensures that adversarial examples generated through this method exhibit better transferability to novel black-box models. The process of generating adversarial images is detailed in Algorithm 2.

Algorithm 2 Adversarial Image Generation

Require: $model$: model for attack, $imgs$: input images, $device$: computation device, α : scaling factors, X_{txt} : text embeddings

Ensure: I_{adv} : adversarial images

```
1:  $model.eval()$ 
2: Initialize  $I_{adv} \leftarrow imgs.detach() + \text{Uniform}(-\epsilon, \epsilon)$  and clamp to  $[0.0, 1.0]$ 
3: for each iteration  $i = 1$  to  $N$  do
4:   for each  $img \in I_{adv}$  do
5:     Apply transformations and data augmentation to  $img$ 
6:     Select text embeddings for operation
7:     Compute similarity and Wasserstein distance
8:     Perform Sinkhorn optimization to obtain  $T$ 
9:     if  $\text{isnan}(T)$  then
10:      return None
11:    end if
12:    Compute and backpropagate loss  $loss_{OT}$ 
13:    Update adversarial image using gradient sign
14:    Clamp  $I'_{adv}$  within perturbation limits
15:    Update  $I_{adv}$  with the new adversarial image
16:  end for
17: end for
18: return  $I_{adv}$ 
```

We employed the adversarial example generation method outlined in Equation 1 to create adversarial samples. These samples were then used to mount attacks on black-box models.

4. Experiments

4.1. Settings

VLP Models. To assess the transferability of adversarial examples and the efficacy of our proposed framework, we employed two distinct categories of Vision-Language Pre-training (VLP) models: fused VLP and aligned VLP. These were respectively designated as the source and target models in our experiments. Fused VLP models are characterized by their early integration of visual and linguistic information within the processing pipeline. Specifically, these models concurrently handle images and text, extracting and processing both types of data through shared layers. Within this category, we selected ALBEF [29] and TCL [30] as representative fused models. Both models incorporate a 12-layer ViT-B/16 [35] visual transformer and two separate 6-layer transformers for image and text encoding, with their primary differences arising from distinct pre-training objectives. Conversely, aligned VLP models process visual and textual data independently at the initial stages and subsequently align these representations in the deeper layers of the model. This approach facilitates the learning of intricate relationships between the two modalities. In this category, we opted for two variants of CLIP [31] as our representatives:

CLIP_{ViT}, which employs ViT-B/16 for image encoding, and CLIP_{CNN}, which uses a ResNet-101 [51] architecture for the same purpose. For the targeted task of image captioning within our cross-task attack efficacy study, we incorporated BLIP as the target model, assessing adversarial examples' performance in the black-box setting when transitioning from TCL as the source model to ALBEF.

Datasets. For the image-text retrieval task, our study utilized two datasets renowned for their breadth and depth: Flickr30K [32] and MSCOCO [33]. Flickr30K boasts a diverse corpus of 31,783 images, while MSCOCO expands the dataset considerably with 123,287 images. 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+ [52] dataset, which further enriched our cross-task attack effectiveness analysis.

Baselines In our research involving Vision-Language Pre-training (VLP) models, we implemented several prevalent adversarial attack methods as baselines. These included using PGD [26] exclusively on images, applying BERT-Attack [25] only to texts, and separately utilizing PGD and BERT-Attack on both images and texts without integrating inter-modality interactions, a technique designated as Sep-Attack. Additionally, we employed Co-Attack [6], which integrates information between individual image-text pairs, and Set-level Guidance Attack (SGA) [27], which utilizes guidance information across modalities between sets. Each baseline was tested under identical conditions for a consistent comparative analysis.

Adversarial Attack Configuration. In order to validate the effectiveness of our proposed framework, we employed the Projected Gradient Descent (PGD) [26] method for generating adversarial visual examples. The configuration settings for PGD were as follows: a perturbation limit of $\epsilon_v = \frac{2}{255}$, a step magnitude of $\alpha = \frac{0.5}{255}$, and a total of $T = 10$ iterations. For textual attacks, we utilized the BERT-Attack [25] method, setting a disturbance limit to $\epsilon_t = 1$ and defining a vocabulary list length of $W = 10$. In our experimentation with Sep-Attack and Co-Attack, we adhered to the previously mentioned settings. Specifically for Co-Attack, in addition to these settings, we also incorporated the similarity between individual image pairs as a loss metric, guiding the generation of adversarial examples through inter-modality interactions. In the case of SGA, we followed the experimental conditions outlined in its original publication. Specifically, we enhanced the images by rescaling them to five distinct sizes 0.5, 0.75, 1.0, 1.25, 1.5. To validate that our method effectively mitigates overfitting, we enhanced the image set by rescaling it to six different sizes: 0.5, 0.75, 1.0, 1.25, 1.5, 2, utilizing bicubic interpolation. Additionally, we applied horizontal flipping to the images. This form of data augmentation, which led to significant overfitting in SGA, was employed to rigorously test the robustness of our approach against such overfitting tendencies. In parallel, the text corpus was expanded by selecting and refining the top five most closely matched

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	52.45	58.65	3.06	6.79	8.69	13.21	10.34	14.65
	BERT-Attack	11.57	27.46	12.64	28.07	29.33	43.17	32.69	46.11
	Sep-Attack	65.69	73.95	17.60	32.95	31.17	45.23	32.82	45.49
	Co-Attack	77.16	83.86	15.21	29.49	23.60	36.48	25.12	38.89
	SGA	97.24	97.28	45.42	55.25	33.38	44.16	34.93	46.57
	OT-Attack (Ours)	95.93	95.86	52.37	61.05	34.85	47.10	42.33	53.03
TCL	PGD	6.15	10.78	77.87	79.48	7.48	13.72	10.34	15.33
	BERT-Attack	11.89	26.82	14.54	29.17	29.69	44.49	33.46	46.07
	Sep-Attack	20.13	36.48	84.72	86.07	31.29	44.65	33.33	45.80
	Co-Attack	23.15	40.04	77.94	85.59	27.85	41.19	30.74	44.11
	SGA	48.91	60.34	98.37	98.81	33.87	44.88	37.74	48.30
	OT-Attack (Ours)	57.32	65.83	97.81	98.01	34.72	47.16	43.44	54.12
CLIP _{VIT}	PGD	2.50	4.93	4.85	8.17	70.92	78.61	5.36	8.44
	BERT-Attack	9.59	22.64	11.80	25.07	28.34	39.08	30.40	37.43
	Sep-Attack	9.59	23.25	11.38	25.60	79.75	86.79	30.78	39.76
	Co-Attack	10.57	24.33	11.94	26.69	93.25	95.86	32.52	41.82
	SGA	13.40	27.22	16.23	30.76	99.08	98.94	38.76	47.79
	OT-Attack (Ours)	14.29	29.28	16.58	33.49	98.65	98.52	43.55	50.50
CLIP _{CNN}	PGD	2.09	4.82	4.00	7.81	1.10	6.60	86.46	92.25
	BERT-Attack	8.86	23.27	12.33	25.48	27.12	37.44	30.40	40.10
	Sep-Attack	8.55	23.41	12.64	26.12	28.34	39.43	91.44	95.44
	Co-Attack	8.79	23.74	13.10	26.02	28.79	40.03	94.76	96.89
	SGA	11.42	24.80	14.91	28.82	31.24	42.12	99.24	99.49
	OT-Attack (Ours)	11.57	26.24	14.91	30.52	35.63	48.20	99.39	99.32

TABLE 2: Adversarial Attack Success Rates on Image-Text Retrieval. **The best results are boldfaced.** The table presents the attack success rate at Rank 1 (ASR @ R1) for text-image retrieval (IR) and text-image retrieval (TR) tasks using the Flickr30K dataset. The originating models for generating adversarial samples are listed under the ‘Source’ column. White-box attacks are distinguished by a gray background, emphasizing their distinct context in the evaluation. **The results demonstrate the proposed OT-Attack’s impressive performance in achieving SOTA transferability of adversarial samples across different VLP models.**

caption pairs for each image in the dataset. Furthermore, we incorporated the Sinkhorn [50] algorithm for calculating the optimal transport plan. This algorithm was chosen for its efficient use of matrix operations, which accelerates the process and ensures convergence to an approximate solution of the original transport problem. To avoid infinite iterations, we set a convergence threshold $\text{thresh} = 1e-2$. The iteration process was terminated once the average absolute difference of the vector r between two consecutive iterations fell below this threshold, indicating the achievement of algorithmic convergence.

Evaluation Criteria. In our study, the robustness and transferability of the adversarial attacks, in both white-box and black-box scenarios, are 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. The ASR is computed as follows:

$$ASR = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100\% \quad (16)$$

where ASR denotes the Attack Success Rate, N_{success} represents the number of successful attacks, and N_{total} is the total number of attacks conducted. The formula calculates the percentage of successful attacks, providing a quantitative measure of the attack’s effectiveness.

4.2. Comparative Experimental Results

In our experiments, we primarily focused on Image-Text Retrieval (ITR) tasks. We generated adversarial examples on various white-box models and then evaluated their effectiveness by calculating the attack success rates on both the white-box models and three additional black-box models. Specifically, we computed the success rates for both image-to-text and text-to-image matching attacks to demonstrate the comprehensiveness of our method’s impact on adversarial sample generation. This deliberate selection of models and tasks was instrumental in enabling an exhaustive examination of the adversarial examples’ ability to

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	76.70	86.30	12.46	17.77	13.96	23.1	17.45	23.54
	BERT-Attack	24.39	36.13	24.34	33.39	44.94	52.28	47.73	54.75
	Sep-Attack	82.6	89.88	32.83	42.92	44.03	54.46	46.96	55.88
	Co-Attack	79.87	87.83	32.62	43.09	44.89	54.75	47.3	55.64
	SGA	96.75	96.95	58.56	65.38	57.06	62.25	58.95	66.52
	OT-Attack (Ours)	95.41	95.8	63.44	68.9	58.79	65.87	63.56	72.16
TCL	PGD	10.83	16.52	59.58	69.53	14.23	22.28	17.25	23.12
	BERT-Attack	35.32	45.92	38.54	48.48	51.09	58.8	52.23	61.26
	Sep-Attack	41.71	52.97	70.32	78.97	50.74	60.13	51.9	61.26
	Co-Attack	46.08	57.09	85.38	91.39	51.62	60.46	52.13	62.49
	SGA	65.93	73.3	98.97	99.15	56.34	63.99	59.44	65.7
	OT-Attack (Ours)	71.64	78.38	98.69	98.78	58.64	65.75	63.45	72.01
CLIP _{VIT}	PGD	7.24	10.75	10.19	13.74	54.79	66.85	7.32	11.34
	BERT-Attack	20.34	29.74	21.08	29.61	45.06	51.68	44.54	53.72
	Sep-Attack	23.41	34.61	25.77	36.84	68.52	77.94	43.11	49.76
	Co-Attack	30.28	42.67	32.84	44.69	97.98	98.8	55.08	62.51
	SGA	33.41	44.64	37.54	47.76	99.79	99.79	58.93	65.83
	OT-Attack (Ours)	35.11	46.48	38.52	50.32	99.69	99.75	62.16	68.96
CLIP _{CNN}	PGD	7.01	10.62	10.08	13.65	4.88	10.7	76.99	84.2
	BERT-Attack	23.38	34.64	24.58	29.61	51.28	57.49	54.43	62.17
	Sep-Attack	26.53	39.29	30.26	41.51	50.44	57.11	88.72	92.49
	Co-Attack	29.83	41.97	32.97	43.72	53.1	58.9	96.72	98.56
	SGA	31.61	43	34.81	45.95	56.62	60.77	99.61	99.8
	OT-Attack (Ours)	32.9	44.03	36.07	48.17	61.14	67.79	99.16	99.59

TABLE 3: Visualization of Adversarial Examples in Image-Text Matching on the MSCOCO Dataset. **The best results are boldfaced.** It illustrates the effects of adversarial attacks on both images and their associated captions. The top row displays the original, clean images with their corresponding accurate captions. The middle row presents the adversarial images and the modified captions that resulted from the attacks. Key alterations in the text are indicated in red, showcasing the change in the description of the visual content due to the adversarial manipulation. The bottom row quantifies the pixel differences between the original and adversarial images, highlighting the subtlety of the visual perturbations.

generalize across varying architectures, which is a critical factor in the real-world application of these models.

Our analysis spanned two widely recognized datasets: Flickr30K, with a sample of 1,000 images and 5,000 captions, and MSCOCO, which provided a larger pool of 5,000 images and 25,000 captions. This broad dataset coverage allowed us to conduct a robust evaluation of our attack methods in image-text matching tasks, quantifying the success of adversarial examples in misleading these complex models. The detailed outcomes are methodically presented in TABLE 2 and TABLE 3.

Our results demonstrated that the OT-Attack method made significant strides in the creation of adversarial examples that were not only effective within models of the same type but also exhibited impressive cross-type attack success. This is particularly evident from the R@1 success rates in TR and IR tasks, where our adversarial examples maintained high effectiveness across varied models including ALBEF, TCL, CLIP_{VIT}, and CLIP_{CNN}. For example, when using ALBEF to target TCL, our method improved the TR R@1 attack success rate by 6.95% on Flickr30K and 4.88% on MSCOCO, compared with the state-of-the-art results

obtained by SGA. Conversely, in scenarios where TCL was employed to target ALBEF, our approach showed significant improvements over SGA, with increases of 8.41% on Flickr30K and 5.71% on MSCOCO in the TR R@1 attack success rate. The results demonstrate the effectiveness of our proposed method in improving adversarial transferability.

Complementing our numerical analysis, Figure 4 offers a visual representation of the impact of our adversarial examples. It contrasts the original images and texts with their modified versions, illuminating how subtle perturbations can drastically alter a model’s performance in image-text matching tasks. The visual differences, particularly the nuanced texture changes introduced in the adversarial images, are made evident through difference masks, underscoring the deceptive potency of the adversarial examples and their potential to misguide VLP systems. This visual depiction reinforces the quantitative data, providing a holistic view of our adversarial attack’s efficacy. Hence, compared with previous works, our proposed OT-Attack can achieve better adversarial transferability.

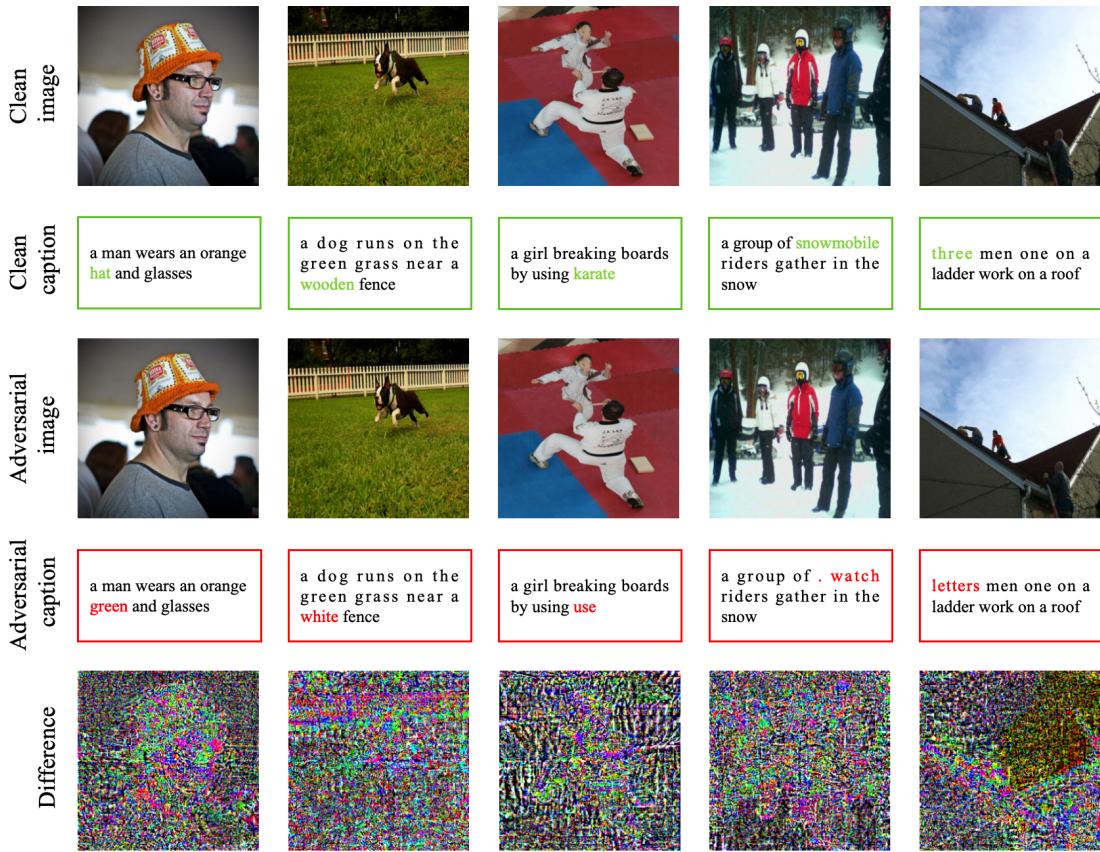


Figure 4: Visualization of adversarial examples from Flickr30K. In the task of image-text matching, adversarial examples for both images and texts were generated and utilized for image-to-text and text-to-image matching tasks, respectively. We have highlighted the distinctions in the text adversarial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

4.3. Cross-Task Transferability

4.3.1 Image Captioning

In our research, we employed the ALBEF model in a white-box setting to generate adversarial examples specifically designed to target the BLIP [53] framework. Renowned for its innovative multimodal encoder-decoder model, BLIP is pre-trained on a dataset enriched with diverse synthetic captions and noise reduction techniques. We chose the MSCOCO dataset for our evaluations, focusing on both original and adversarially manipulated image samples. To quantify the effectiveness of our adversarial approach, we utilized a suite of metrics for image captioning tasks, each offering distinct insights: **BLEU** [54] assesses the precision of generated captions by comparing them with reference captions, focusing on the exactness of word usage and sequence. **METEOR** [55] expands upon BLEU by including synonym matching and stemming, providing a more comprehensive view of semantic accuracy. **ROUGE** [56] measures recall, evaluating how much of the reference content is captured in the generated captions. **CIDEr** [57] focuses on the distinctiveness and

Attack	B@4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline	39.7	31.0	60.0	133.3	23.8
Co-Attack	37.4	29.8	58.4	125.5	22.8
SGA	34.8	28.4	56.3	116.0	21.4
OT-Attack (Ours)	34.1	27.9	55.7	112.6	20.9

TABLE 4: Adversarial Impact on Image Captioning Metrics. This table displays the results of adversarial attacks on the image captioning task, where 10,000 MSCOCO dataset images were used. Adversarial samples were crafted using the ALBEF model in a white-box attack scenario, and captions were subsequently generated with the BLIP model. The performance of these attacks was assessed by measuring the BLEU-4 (B@4), METEOR, ROUGE-L, CIDEr, and SPICE metrics. Lower scores on these metrics indicate a more effective attack, revealing a significant deviation of the generated captions from the expected, accurate descriptions. The results underscore the effectiveness of our proposed method in comparison to other attacks, demonstrating its capability to disrupt the caption generation process by inducing semantic errors.

Clean image				
Caption	a brown and white dog running through a lush green field	a woman holding a child with cars behind her	several girls in blue shirts are playing with their instruments	a person is playing with water in the ocean
Adversarial image				
Caption	a man is leading a horse around the yard	two men are taking a selfie with their cell phone	a group of people holding tennis rackets on a grass court	a man in the water with a kite and no head

Figure 5: Comparison of Clean and Adversarial Image Captions. This figure juxtaposes the original clean images with their accurate captions against adversarial images and the resulting captions generated by the BLIP model. The adversarial examples were created using the ALBEF model as a white-box framework on the Dataset Flickr30K. Despite the perturbations being subtle, and limited to a magnitude of 2, the adversarial examples show minimal visual deviation from the original images. However, these slight alterations are significant enough to mislead the captioning model, leading to discrepancies in the generated captions, as evidenced by the erroneous and sometimes nonsensical descriptions.

relevance of captions by assessing consensus with reference captions. Lastly, SPICE [58] evaluates the semantic content of captions, assessing their fidelity to the actual objects, attributes, and relationships in the image.

Each metric provides a different perspective on the quality and relevance of the generated text, offering a comprehensive understanding of the adversarial impact. The detailed results of these evaluations are compiled in TABLE 4. Compared to SGA, our method exhibited a decrease in performance across various metrics: a reduction of 0.7 in BLEU-4, 0.5 in METEOR, 0.6 in ROUGE-L, 3.4 in CIDEr, and 0.5 in SPICE. Notably, these declines in scores are indicative of enhanced effectiveness in cross-task attacks, as a greater decrease signifies better performance of our method in this context.

To convey the practical implications of our findings, Figure 5 provides a visual comparison of select experimental results. This figure juxtaposes the original images and their corresponding captions with the adversarial counterparts, highlighting the stark differences. These visual examples compellingly illustrate how subtle and seemingly innocuous perturbations can lead to significant deviations in the model’s interpretation, often diverging entirely from the original context or meaning of the image.

Further delving into the realm of large-scale models,

our experiments were conducted 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 ran our adversarial process for 500 iterations. To assess the broader applicability and effectiveness of our attacks, we tested them on advanced models like GPT-4 and Bing Chat, posing the query “Describe this image” to these systems. The findings, illustrated in Figure 6, reveal a notable level of success in our adversarial attacks, with these sophisticated models showing susceptibility to being misled.

4.3.2 Visual Grounding

To thoroughly evaluate the efficacy of our adversarial attack strategies, we utilized the RefCOCO+ [52] dataset, meticulously designed for the visual grounding task. This dataset is composed of several subsets, each tailored to assess different facets of model performance. These subsets include RefCOCO+ val, RefCOCO+ testA, and RefCOCO+ testB. Specifically, the RefCOCO+ val subset presents a wide array of scenarios for a comprehensive assessment, RefCOCO+ testA is dedicated to gauging the model’s proficiency in identifying and localizing human figures, and RefCOCO+ testB concentrates on non-human elements like inanimate objects, animals, and various environmental features. By

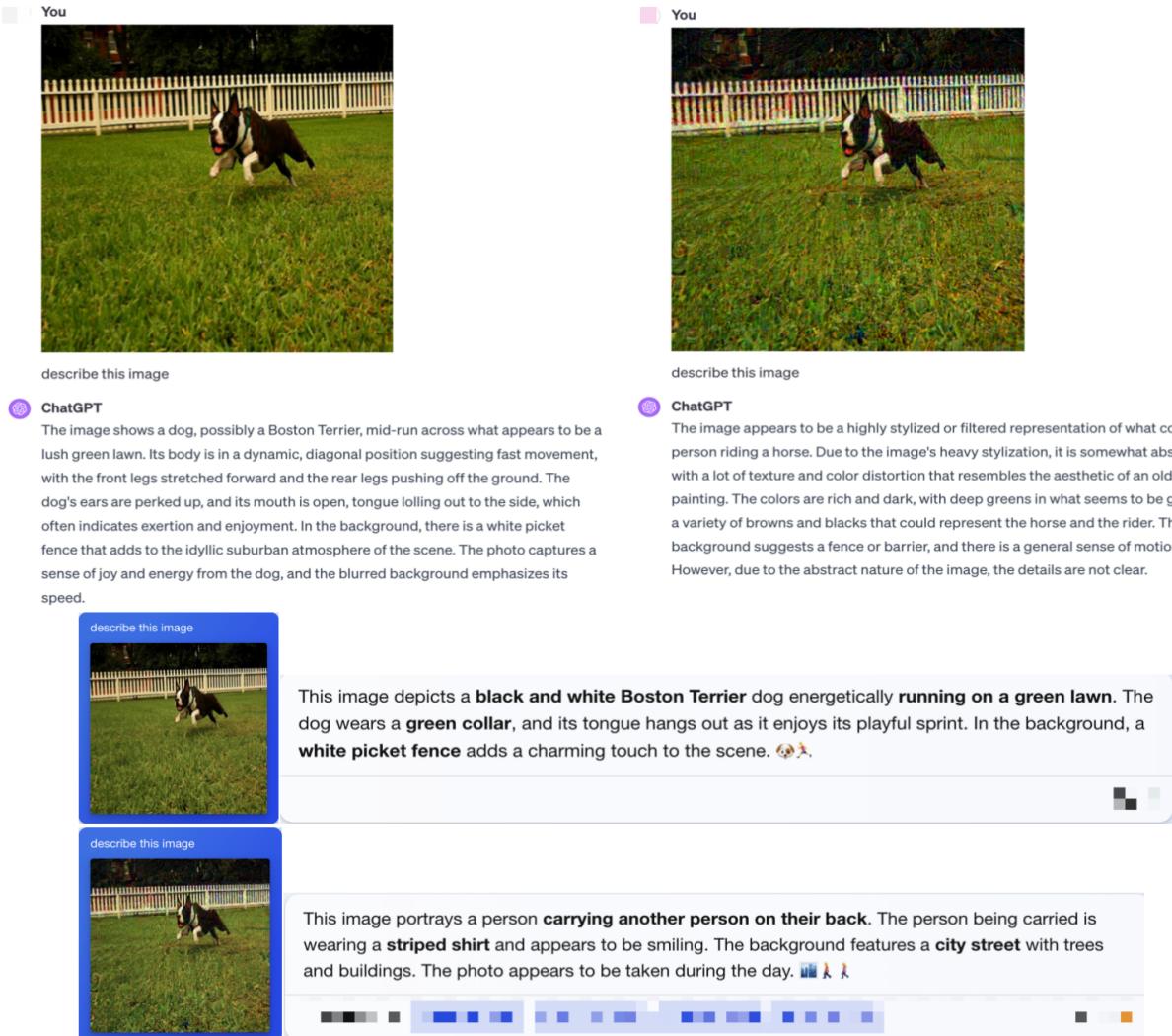


Figure 6: Impact of Adversarial Attacks on GPT-4 and Bing Chat Descriptions. This figure showcases the alterations in image descriptions by GPT-4 and Bing Chat before and after adversarial attacks. Original descriptions are compared to those generated from manipulated images, with increased perturbation 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.

leveraging the diverse testing environments provided by ReFCOCO+, our goal is to showcase the extensive adaptability and transferability of our method across a multitude of visual grounding challenges.

The quantitative analysis presented in TABLE 5 assesses the performance of our adversarial examples across different tasks. In this evaluation, TCL served as the source model from which adversarial examples were generated, while the ALBEF model was the target for these attacks. The baseline in the table refers to the scores obtained when the original, unaltered samples were evaluated, providing a control benchmark for our study. Alongside the baseline, both SGA and OT-Attack strategies were employed. Compared to SGA, our attack method proved to be more effective, as

evidenced by the decrease in scores for the ALBEF model due to our adversarial examples: a reduction of 0.2 on the Val subset, 0.2 on TestA, and 0.3 on TestB. This outcome highlights the effectiveness of our method in compromising the Visual Grounding capabilities of the target model.

Beyond the numerical assessments, we also employed a visualization technique to further elucidate the impact of our attacks on Visual Grounding tasks, as depicted in Figure 7. The base samples for this analysis were derived from the Flickr30K dataset, with adversarial examples being crafted using the ALBEF model in a white-box configuration. Our visual representations reveal that even subtle perturbations can significantly disrupt the model's ability to accurately recognize and localize objects. This underscores the sub-

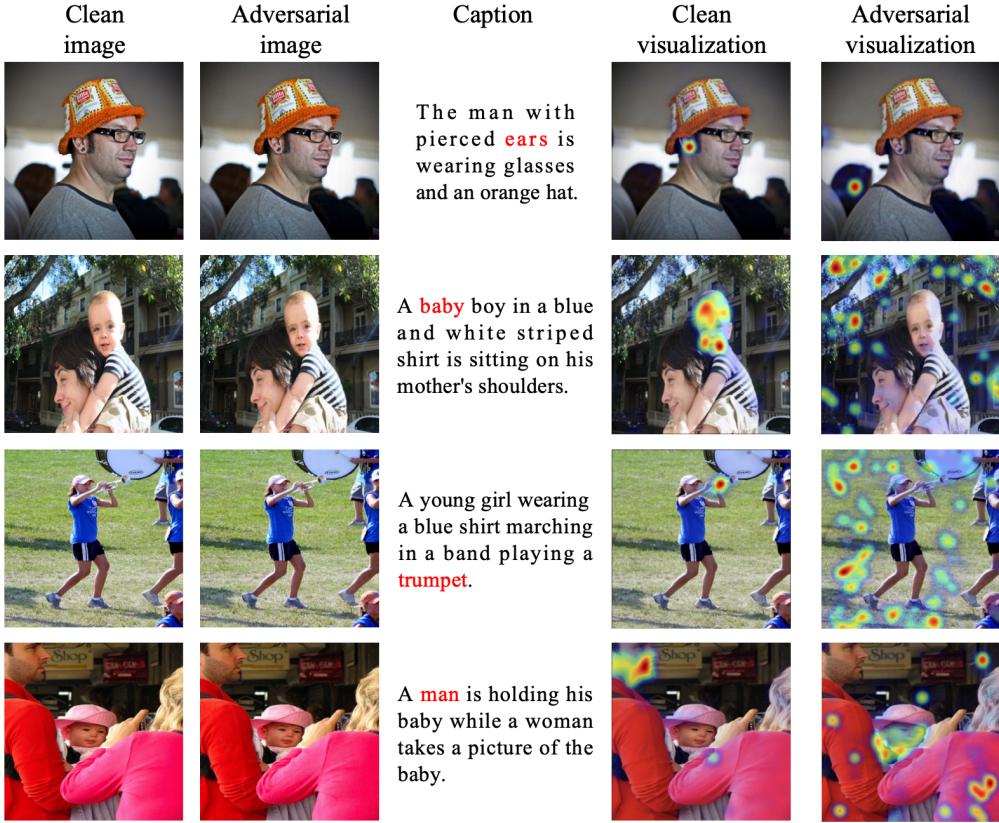


Figure 7: Visualization results for Visual Grounding. We employed TCL as the source model and ALBEF as the target model, with captions sourced from the Flickr30K dataset. The adversarial examples exhibit limited visual differences from the original samples; however, they disrupt the model’s judgment 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.

Attack	Val	TestA	TestB
Baseline	58.4	65.9	46.2
SGA	56.5	63.7	45.4
OT-Attack (Ours)	56.3	63.5	45.0

TABLE 5: Performance on Visual Grounding Task Across RefCOCO+ Subsets. This table evaluates the success of different adversarial attacks across the validation (Val), TestA, and TestB subsets of the RefCOCO+ dataset. The TCL model served as the source for generating adversarial attacks, while the ALBEF model was the target for evaluating their effectiveness. Lower scores indicate a more successful adversarial attack, highlighting the inverse relationship between the evaluation metrics and the attack’s performance. The results demonstrate the comparative effectiveness of our proposed methodology in reducing the ALBEF model’s accuracy in the visual grounding task, as evidenced by the decreased performance across all subsets.

stantial and nuanced impact that these adversarial attacks can have on a model’s performance in visual grounding tasks, challenging its reliability and precision in real-world applications.

5. Conclusion

In this paper, we focus on improving the adversarial transferability of vision-language pre-training models. In detail, recent works have found that using inter-modality interaction and data augmentation can significantly enhance the transferability of adversarial examples for vision-language pre-training models. Unfortunately, previous works ignore the optimal alignment problem between data-augmented image-text pairs, which leads the generated adversarial examples overfit to the source model and achieves limited adversarial transferability improvement. To address the issue, we propose an Optimal Transport-based Adversarial Attack, *dubbed* OT-Attack. The proposed OT-Attack formulates the features of image and text sets as two distinct distributions, leveraging optimal transport theory to identify the most efficient mapping between them. It utilizes their mutual

similarity as the cost matrix. The derived optimal mapping guides the generation of adversarial examples, effectively mitigating overfitting issues and improving adversarial transferability. Extensive experiments across diverse network architectures and datasets in image-text matching tasks demonstrate the superior performance of the proposed OT-Attack compared to existing methods in terms of adversarial transferability. Significantly, our results also show that OT-Attack is also effective in cross-task attacks, including image captioning and visual grounding, and poses a considerable challenge to commercial models such as GPT-4 and Bing Chat, highlighting the evolving landscape of adversarial threats in advanced AI applications. This underscores the need for robust defenses against sophisticated attacks.

References

- [1] C. Lei, S. Luo, Y. Liu, W. He, J. Wang, G. Wang, H. Tang, C. Miao, and H. Li, “Understanding chinese video and language via contrastive multimodal pre-training,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2567–2576.
- [2] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and vqa,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.
- [3] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, “Vlmo: Unified vision-language pre-training with mixture-of-modality-experts,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 897–32 912, 2022.
- [4] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, “Scaling up vision-language pre-training for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 980–17 989.
- [5] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, “Las-at: adversarial training with learnable attack strategy,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 398–13 408.
- [6] J. Zhang, Q. Yi, and 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.
- [7] Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, and F. Ma, “Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [8] Z. Zhou, S. Hu, M. Li, H. Zhang, Y. Zhang, and H. Jin, “Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6311–6320.
- [9] X. Jia, Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao Sr, “Improving fast adversarial training with prior-guided knowledge,” *arXiv preprint arXiv:2304.00202*, 2023.
- [10] OpenAI, “GPT-4 technical report.” *CoRR*, vol. abs/2303.08774, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2303.08774>
- [11] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [12] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples,” *arXiv preprint arXiv:1605.07277*, 2016.
- [13] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, “Improving transferability of adversarial examples with input diversity,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2730–2739.
- [14] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, “Nesterov accelerated gradient and scale invariance for adversarial attacks,” in *International Conference on Learning Representations*, 2019.
- [15] Y. Dong, T. Pang, H. Su, and J. Zhu, “Evasion defenses to transferable adversarial examples by translation-invariant attacks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4312–4321.
- [16] X. Jia, X. Wei, X. Cao, and X. Han, “Adv-watermark: A novel watermark perturbation for adversarial examples,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1579–1587.
- [17] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, “Frequency domain model augmentation for adversarial attack,” in *European Conference on Computer Vision*. Springer, 2022, pp. 549–566.
- [18] X. Jia, Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao, “Prior-guided adversarial initialization for fast adversarial training,” in *European Conference on Computer Vision*. Springer, 2022, pp. 567–584.
- [19] S. Han, C. Lin, C. Shen, Q. Wang, and X. Guan, “Interpreting adversarial examples in deep learning: A review,” *ACM Computing Surveys*, 2023.
- [20] J. Gu, X. Jia, P. de Jorge, W. Yu, X. Liu, A. Ma, Y. Xun, A. Hu, A. Khakzar, Z. Li et al., “A survey on transferability of adversarial examples across deep neural networks,” *arXiv preprint arXiv:2310.17626*, 2023.
- [21] M. Gubri, M. Cordy, M. Papadakis, Y. L. Traon, and K. Sen, “Lgv: Boosting adversarial example transferability from large geometric vicinity,” in *European Conference on Computer Vision*. Springer, 2022, pp. 603–618.
- [22] Z. Qin, Y. Fan, Y. Liu, L. Shen, Y. Zhang, J. Wang, and B. Wu, “Boosting the transferability of adversarial attacks with reverse adversarial perturbation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 845–29 858, 2022.
- [23] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim, “Improving the transferability of targeted adversarial examples through object-based diverse input,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 244–15 253.
- [24] F. Waseda, S. Nishikawa, T.-N. Le, H. H. Nguyen, and I. Echizen, “Closer look at the transferability of adversarial examples: How they fool different models differently,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1360–1368.
- [25] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, “Bert-attack: Adversarial attack against bert using bert,” *arXiv preprint arXiv:2004.09984*, 2020.
- [26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.
- [27] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, and 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.
- [28] C. Villani et al., *Optimal transport: old and new*. Springer, 2009, vol. 338.
- [29] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.
- [30] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and 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. 15 671–15 680.

- [31] 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.
- [32] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and 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.
- [33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and 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.
- [34] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, “Vlp: A survey on vision-language pre-training,” *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023.
- [35] 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,” *arXiv preprint arXiv:2010.11929*, 2020.
- [36] 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 transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.
- [37] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, “Image-text retrieval: A survey on recent research and development,” *arXiv preprint arXiv:2203.14713*, 2022.
- [38] W. Li, S. Yang, Q. Li, X. Li, and A.-A. Liu, “Commonsense-guided semantic and relational consistencies for image-text retrieval,” *IEEE Transactions on Multimedia*, 2023.
- [39] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.
- [40] T. Ghandi, H. Pourreza, and H. Mahyar, “Deep learning approaches on image captioning: A review,” *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023.
- [41] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, “Visual grounding via accumulated attention,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7746–7755.
- [42] Z. Yang, K. Kafle, F. Dernoncourt, and 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. 19 165–19 174.
- [43] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, “Boosting adversarial attacks with momentum,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.
- [44] G. Peyré, M. Cuturi *et al.*, “Computational optimal transport,” *Center for Research in Economics and Statistics Working Papers*, no. 2017-86, 2017.
- [45] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*. PMLR, 2017, pp. 214–223.
- [46] H. Xu, D. Luo, H. Zha, and L. C. Duke, “Gromov-wasserstein learning for graph matching and node embedding,” in *International conference on machine learning*. PMLR, 2019, pp. 6932–6941.
- [47] C. Zhang, Y. Cai, G. Lin, and C. Shen, “Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213.
- [48] B. Liu, Y. Rao, J. Lu, J. Zhou, and C.-J. Hsieh, “Multi-proxy wasserstein classifier for image classification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8618–8626.
- [49] W. Zhao, Y. Rao, Z. Wang, J. Lu, and J. Zhou, “Towards interpretable deep metric learning with structural matching,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9887–9896.
- [50] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” *Advances in neural information processing systems*, vol. 26, 2013.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [52] L. Yu, P. Poirson, S. Yang, A. C. Berg, and 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.
- [53] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.
- [54] K. Papineni, S. Roukos, T. Ward, and 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.
- [55] 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.
- [56] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [57] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [58] P. Anderson, B. Fernando, M. Johnson, and 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.