Semantic-Aligned Adversarial Evolution Triangle for High-Transferability Vision-Language Attack

Xiaojun Jia *, Sensen Gao *, Qing Guo [⊠], Ke Ma, Yihao Huang, Simeng Qin, Yang Liu, *Senior Member, IEEE*, Ivor Tsang *Fellow, IEEE*, and Xiaochun Cao [⊠] *Senior Member, IEEE*

Abstract-Vision-language pre-training (VLP) models excel at interpreting both images and text but remain vulnerable to multimodal adversarial examples (AEs). Advancing the generation of transferable AEs, which succeed across unseen models, is key to developing more robust and practical VLP models. Previous approaches augment image-text pairs to enhance diversity within the adversarial example generation process, aiming to improve transferability by expanding the contrast space of image-text features. However, these methods focus solely on diversity around the current AEs, yielding limited gains in transferability. To address this issue, we propose to increase the diversity of AEs by leveraging the intersection regions along the adversarial trajectory during optimization. Specifically, we propose sampling from adversarial evolution triangles composed of clean, historical, and current adversarial examples to enhance adversarial diversity. We provide a theoretical analysis to demonstrate the effectiveness of the proposed adversarial evolution triangle. Moreover, we find that redundant inactive dimensions can dominate similarity calculations, distorting feature matching and making AEs modeldependent with reduced transferability. Hence, we propose to generate AEs in the semantic image-text feature contrast space, which can project the original feature space into a semantic corpus subspace. The proposed semantic-aligned subspace can reduce the image feature redundancy, thereby improving adversarial transferability. Extensive experiments across different datasets and models demonstrate that the proposed method can effectively improve adversarial transferability and outperform state-of-the-art adversarial attack methods. The code is released at https://github.com/jiaxiaojunQAQ/SA-AET.

Index Terms—Adversarial transferability, vision-language pretraining, adversarial evolution triangle, semantic-aligned.

I. INTRODUCTION

VISION-language pre-training (VLP) models achieve excellent performance across various downstream Visionand-Language tasks, such as visual entailment [51], visual

Xiaojun Jia , Yihao Huang, Yang Liu are with Nanyang Technological University, Singapore (e-mail: jiaxiaojunqaq@gmail.com, huangyihao22@gmail.com, yangliu@ntu.edu.sg)

Sensen Gao is with the Department of Computer Vision, Mohamed Bin Zayed University of Artificial Intelligence. (e-mail: Sensen.Gao@mbzuai.ac.ae)

Qing Guo and Ivor Tsang are with IHPC and CFAR, Agency for Science, Technology and Research, Singapore. (e-mail: tsingqguo@ieee.org and ivor tsang@cfar.a-star.edu.sg)

Ke Ma is with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: make@ucas.ac.cn)

Simeng Qin is with Northeastern University, Shenyang, Liaoning, China (e-mail: qinsimeng670@gmail.com)

Xiaochun Cao is with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China. (e-mail: caoxiaochun@mail.sysu.edu.cn)

* Xiaojun Jia and Sensen Gao contribute equally to this work. $^{\bowtie}$ Qing Guo and Xiaochun Cao are corresponding authors

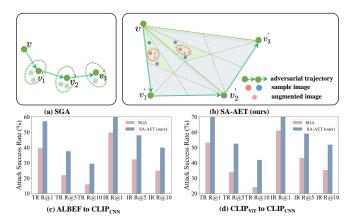


Fig. 1. Comparison of Our Method and Set-Level Guided Attack (SGA) [42]. (a) illustrates the main concept of SGA, which involves performing data augmentations around online adversarial examples. (b) demonstrates the core idea of our SA-AET, where data augmentations are applied within the adversarial sub-triangle. The red and blue dots represent images sampled from this sub-triangle, with red dots highlighting the optimal samples chosen through a text-guided adversarial example selection strategy. The surrounding light red dots represent resized augmentations applied to these optimal samples, similar to the strategy used in SGA. (c) and (d) compare the adversarial transferability of our SA-AET against SGA using adversarial examples from ALBEF [34] and CLIP_{VIT} [48] to attack CLIP_{CNN} [48], respectively.

grounding [32], image captioning [25], and image-text retrieval [30]. However, they have been found to be vulnerable to adversarial examples [66; 42; 21; 22; 9; 43; 16]. Exploring adversarial vulnerabilities can inspire additional research dedicated to developing more robust and applicable VLP models.

Previous works mainly concentrate on exploring generating adversarial examples for VLP models in a white-box setting, in which the attacker can access model internal information, such as model parameters, etc. Some studies [18] have shown that adversarial examples generated on a victim model can successfully attack unseen target models, a phenomenon known as adversarial transferability. Given the limited access to detailed model structures in real-world scenarios, it is crucial to investigate the transferability of multimodal adversarial examples [70; 72]. A series of works focus on generating adversarial examples with high transferability for VLP models. For example, Lu et al. [42] propose to improve adversarial transferability for VLP models by introducing input diversity through data augmentation. Although previous works have achieved some effectiveness in boosting adversarial transferability in vision-language attacks, they mainly focus on maximizing the contrastive loss function in the image-text feature space to generate adversarial examples and increasing their

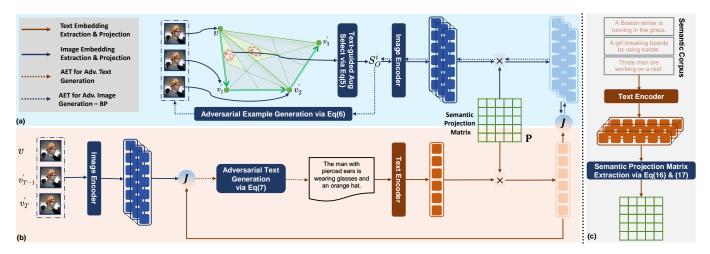


Fig. 2. The Pipeline of the Proposed SA-AET: (a) Pipeline for the Adversarial Evolution Triangle (AET) in Adversarial Image Generation. (b) Pipeline for the Adversarial Evolution Triangle (AET) in Adversarial Text Generation. (c) Pipeline for Extracting the Semantic Projection Matrix.

diversity along the optimization path to improve adversarial transferability. While these methods predominantly enhance diversity in online adversarial examples, they still have a lot of room for improvement in improving adversarial transferability.

Specifically, as shown in Figure 1, during each optimization iteration of the attack, previous works [42; 22] perform data augmentation on the generated adversarial examples (*i.e.*, online adversarial image) to improve transferability. This adversarial strategy increases the variety of adversarial examples throughout the optimization pathway, consequently yielding improvements in their transferability. Nonetheless, the strategy still risks overfitting the victim model due to the heavy reliance on examples from the adversarial trajectory, which results in reduced attack success rates when the adversarial examples are transferred to other VLP models.

To address these issues, as shown in Figure 2 (a) and (b), we first propose adopting the intersection evolution triangles along the adversarial trajectory to enhance the diversity of adversarial examples during optimization. Specifically, in each attack iteration, we propose constructing an adversarial evolution triangle that incorporates the original image, the adversarial image from the previous step, and the current adversarial image. Then, we propose to circumvent overfitting by strategic sampling within this evolution triangle, thereby avoiding excessive focus on adversarial example diversity only around adversarial images. We explore the impact of different adversarial evolution sub-triangle sampling on adversarial transferability and propose to sample from the adversarial evolution sub-triangle close to clean examples and previous adversarial examples. After obtaining the sampling samples, we generate the adversarial perturbations on the samples to stay away from the text. Subsequently, we apply these perturbations to current adversarial examples and select the one that diverges most significantly from the text. We also propose generating adversarial text that deviates from the previous adversarial evolution triangle along the optimization path rather than being distant from the last adversarial image example. We provide a theoretical analysis to illustrate the effectiveness of the proposed adversarial evolution triangle.

In addition, image feature embeddings usually contain much information irrelevant to text features [39; 71; 26]. In the original image-text contrast space, only a limited number of feature dimensions in the image features may be activated, while the remaining dimensions remain inactive and redundant. These redundant dimensions may dominate the similarity calculation, potentially distorting the feature matching of images and texts. This feature distortion makes the generated adversarial examples highly dependent on the victim model, reducing their transferability. Hence, generating adversarial examples within the native image-text feature contrast space increases the likelihood of overfitting the victim model, diminishing the examples' transferability. To overcome this limitation, we propose generating adversarial examples within the semantic image-text feature contrast space, which maps the original feature space into a semantic corpus subspace. Specifically, as shown in Figure 2 (c), we construct a semantic subspace through a series of independent text descriptions and project the image features of the original space into the semantic subspace. Then, we maximize the contrast loss between images and text in the semantic subspace to generate adversarial perturbations.

By assembling the proposed methods, we conduct our vision-language attack by exploiting semantically aligned adversarial evolution triangle, i.e., SA-AET. We conduct a series of experiments to evaluate the effectiveness of the proposed method on the two widely used multimodal datasets, consisting of Flickr30K [47] and MSCOCO [41]. The evaluation experiments are also conducted on three vision-and-language downstream tasks, which include image-text retrieval (ITR), visual grounding (VG), and image captioning (IC). The experimental results indicate that the proposed method can significantly improve the transferability of multimodal adversarial examples, surpassing the state-of-the-art adversarial attack methods. Furthermore, when adversarial examples generated from ITR by the proposed method are applied to other vision-andlanguage downstream tasks, attack performance is significantly enhanced. Our main contributions are in five aspects:

· We propose enhancing the diversity of adversarial ex-

amples during optimization by leveraging the intersection evolution triangle of adversarial trajectories, thereby improving the transferability of multimodal adversarial examples against VLP models. Furthermore, we provide a theoretical analysis to support the proposed adversarial evolution triangle.

- We investigate how sampling from different adversarial evolution sub-triangles affects adversarial transferability and propose sampling from the evolution sub-triangle that is close to clean examples and previous adversarial examples.
- We propose to generate the adversarial text by deviating from the final adversarial evolution triangle along the optimization trajectory rather than the final adversarial example, minimizing overfitting on the surrogate model to improve its transferability.
- To further enhance the transferability of adversarial examples, we propose generating them in the semantic imagetext feature contrast space by mapping the original feature space onto a subspace defined by a semantic corpus.
- Our extensive experiments across various network architectures and datasets demonstrate that the proposed method significantly enhances the transferability of multimodal adversarial examples and outperforms state-of-the-art multimodal transfer adversarial attack methods.

This paper is a journal extension of our conference paper [17] (called DRA). Compared to the preliminary conference version, we have made significant improvements and extensions in this version. The main differences are in four aspects: 1) In addition to sampling the intersection evolution triangle of adversarial trajectories proposed in the previous version, we explore the impact of different sampling strategies and propose to sample adversarial evolution triangles close to clean examples and previously generated adversarial examples in **Section III-B**. This can further improve the transferability of multimodal adversarial examples. We add a theoretical analysis to demonstrate the effectiveness of the proposed adversarial evolution triangle in **Section III-D**. 2) We propose to generate the adversarial examples in the semantic imagetext feature contrast space in **Section III-E**, which can reduce reliance on victim models, thereby improving transferability. 3) We conduct more experiments and analyses, which include comparisons with state-of-the-art methods, ablation studies, and performance analyses. We adopt some state-ofthe-art adversarial attack methods as the new comparison in **Section IV-C.** We add the ablation study versus the different proposed elements in **Section IV-F**. We analyze the effective performance of the proposed method in Section IV-G. 4) We have thoroughly revised the abstract, introduction, method, experiment, and conclusion sections to offer a more detailed overview of our motivation and approach. Furthermore, we have updated all figures and tables to enhance clarity and presentation.

II. RELATED WORK

In this section, we begin by discussing the existing research on vision-language pre-training models. Subsequently,

we explore the studies related to downstream vision and language tasks. Finally, we introduce the research concerning the transferability of multimodal adversarial examples for vision-language pre-training models.

3

A. Vision-Language Pre-training Models

Vision-language pre-training (VLP) models enhance various Vision-and-Language (V+L) tasks by using multimodal learning from extensive image-text pairs [33]. Initially, VLP models heavily relied on pre-trained object detectors to generate multimodal representations [8; 37; 68; 56; 52]. However, recent developments have seen a shift towards employing end-to-end image encoders such as the Vision Transformer (ViT) [13; 53; 64], which provide faster inference speeds. Consequently, some recent studies suggest replacing the computationally intensive object detectors with these more efficient image encoders [14; 33; 34; 55; 61]. The learning visionlanguage representations in VLP models can be divided into two categories: the fused architecture and the aligned architecture. Fused VLP models, such as ALBEF [34] and TCL [61], first adopt two separate unimodal encoders to extract features from text and images. These features are combined using a multimodal encoder to create a joint representation. On the other hand, aligned VLP models, like CLIP [48], aim to harmonize the feature spaces of different unimodal encoders, significantly enhancing performance in downstream tasks [1]. In this paper, we focus on assessing our proposed method using various popular fused and aligned VLP models.

B. Downstream Vision-and-Language Tasks

The downstream vision-and-language tasks can be divided into three categories: Image-text Retrieval (ITR), Visual Grounding (VG), and Image Captioning (IC). Image-text Retrieval (ITR): it involves retrieving relevant information, both textual and visual, in response to queries presented in a different modality [7; 10; 59; 69]. Typically, it includes two sub-tasks: image-to-text retrieval and text-to-image retrieval. Specifically, for aligned VLP models, the Text Retrieval (TR) and Image Retrieval (IR) tasks utilize ranking based on the similarity between text and image embeddings. For fused VLP models, similarity scores are computed for all image-text pairs to identify the Top-N candidates since the embedding spaces of unimodal encoders are not aligned. The multimodal encoder then processes these candidates, calculating matching scores to determine the final ranking. Visual Grounding (VG): Identifying and mapping out specific areas within a visual scene corresponding to entities or concepts described in natural language. For example, in VLP models, ALBEF enhances Grad-CAM [49] by leveraging the resulting attention map to prioritize the detected proposals. Image Captioning (IC): It involves creating a textual description that accurately reflects or conveys the content of a given visual input, usually through generating captions for images. To evaluate the performance of image captioning models, metrics such as BLEU [45], METEOR [6], ROUGE [40], CIDEr [54] and SPICE [3] are commonly used. These metrics measure the quality and relevance of the generated captions by comparing them to reference captions.

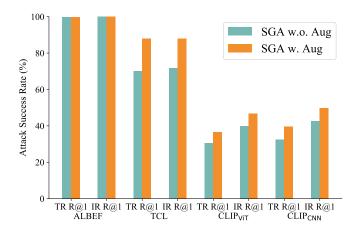


Fig. 3. Attack Success Rate (%) of SGA with and without Image Augmentation. The SGA w.o. Aug does not utilize image augmentation techniques. We employ ALBEF to generate multimodal adversarial examples.

C. Adversarial Transferability

Adversarial attack methods [28; 20; 19; 23; 18] can be divided into two categories: white-box attacks [29; 27] and black-box attacks [5; 46]. The white-box attacks indicate that the attacker has full access to the model (e.g., model parameters and architecture), while the black-box attacks cannot. An increasing number of researchers focus on black-box attacks, as they are more practical for real-world applications where model information is often limited. In the context of image classification tasks, data augmentation techniques like TIM [12], ADMIX [57], and PAM [67] are employed to enhance the transferability of adversarial examples. Recent works [42; 22; 11; 15] have begun to explore improving the transferability of multimodal adversarial examples for VLP models. One common method involves integrating unimodal adversarial attacks [44; 35] from each modality. For example, Sep-Attack [42] directly combines BERT-Attack [35] for text and PGD [44] for image attacks. Zhang et al. [66] develop a white-box attack, called Co-Attack, to attack popular VLP models by considering cross-modal interactions. Building on this foundation, Lu et al. [42] propose the SGA method to boost the transferability of multimodal adversarial examples by expanding a single image-text pair into diverse sets to increase adversarial example variety. However, SGA predominantly considers diversity near adversarial examples throughout the optimization, which may elevate the likelihood of overfitting to the targeted model and reduce the adversarial transferability. Consequently, this work aims to thoughtfully broaden the diversity of adversarial examples without excessively concentrating on diversity solely around adversarial images.

III. THE PROPOSED METHOD

A. Motivation

Adversarial attacks on VLP models typically cause a mismatch between adversarial images and their corresponding adversarial text while conforming to predefined limits on perturbations in both the image and text domains. Let $(\mathbf{x}_I, \mathbf{x}_T) \in \mathcal{D}$ denote an image-text pair extracted from a multimodal

dataset \mathcal{D} . $(\tilde{\mathbf{x}}_I, \tilde{\mathbf{x}}_T)$ represents the corresponding adversarial image-text pair in the image-text searching spaces $B[\mathbf{x}_I, \epsilon_I]$ and $B[\mathbf{x}_T, \epsilon_T]$, where ϵ_I represents the maximal perturbation bound for the image, and ϵ_T represents the maximal number of changeable words in the caption. $F_I(\cdot)$ and $F_T(\cdot)$ represent the image and text encoders of VLP models, respectively. The multimodal adversarial examples are generated by maximizing the loss function J of VLP models. The objective function can be defined as:

$$\begin{cases}
\max J\left(F_{I}\left(\tilde{\mathbf{x}}_{I}\right), F_{T}\left(\tilde{\mathbf{x}}_{T}\right)\right) \\
s.t.\tilde{\mathbf{x}}_{I} \in B\left[\mathbf{x}_{I}, \epsilon_{I}\right], \tilde{\mathbf{x}}_{T} \in B\left[\mathbf{x}_{T}, \epsilon_{T}\right].
\end{cases} (1)$$

We adopt $J(\tilde{\mathbf{x}}_I, \mathbf{x}_T)$ to represent $J(F_I(\tilde{\mathbf{x}}_I), F_T(\mathbf{x}_T))$. Previous works for improving the transferability of multimodal adversarial examples, specifically SGA [42], employ augmented image-text pairs to enhance the diversity of adversarial examples throughout the optimization process. During the generation of adversarial images, let $\tilde{\mathbf{x}}_I^i$ denote the adversarial image generated at the i-th step. In the next step (i+1), the SGA procedure begins by performing data augmentation (resize images to different scales) on $\tilde{\mathbf{x}}_I^i$, producing n augmented examples $\{\tilde{\mathbf{x}}_I^{i1}, \tilde{\mathbf{x}}_I^{i2}, \ldots, \tilde{\mathbf{x}}_I^{in}\}$. The whole process can be formulated as follows:

$$\tilde{\mathbf{x}}_{I}^{i+1} = \prod_{\mathbf{x}_{I}, \epsilon_{I}} \left(\tilde{\mathbf{x}}_{I}^{i} + \alpha \cdot \operatorname{sign} \left(\frac{\nabla_{\mathbf{x}_{I}} \sum_{j=1}^{n} J(\tilde{\mathbf{x}}_{I}^{ij}, \mathbf{x}_{T})}{\left\| \nabla_{\mathbf{x}_{I}} \sum_{j=1}^{n} J(\tilde{\mathbf{x}}_{I}^{ij}, \mathbf{x}_{T}) \right\|} \right) \right),$$
(2)

where α represents the step size and $\prod_{\mathbf{x}_I, \epsilon_I}$ is to constrain adversarial examples \mathbf{x}_I to stay within an ϵ_I -radius of the clean image under the L_{∞} norm.

To better understand the role of image augmentation during the optimization phase in SGA, ALBEF is used as a surrogate model to generate multimodal adversarial examples. The generated adversarial examples are then applied to attack target VLP models, such as TCL and CLIP, to evaluate the transferability of the multimodal adversarial examples. The experimental results are shown in Figure 3. It can be observed that SGA can significantly boost the transferability of multimodal adversarial examples for VLP models, with improvements ranging from 6.14% to 17.81%. However, it is important to note that the attack success rates of the target models are still considerably lower than the source model. This discrepancy is mainly due to SGA's focus on promoting diversity around the adversarial example $\tilde{\mathbf{x}}_I^i$ during optimization while insufficiently accounting for the diversity of adversarial examples relative to the clean image. This can lead to overfitting on the victim model, thereby reducing adversarial transferability.

To improve adversarial transferability, we propose to sample from adversarial evolution triangles composed of clean examples, historical adversarial examples, and current adversarial examples to boost the diversity of adversarial examples. Besides, we investigate the impact of various sampling strategies within adversarial evolution triangles on adversarial transferability and propose sampling from evolution triangles near clean examples and previously generated adversarial examples. Additionally, as for the text adversarial perturbations, our goal is to generate the adversarial perturbation by deviating

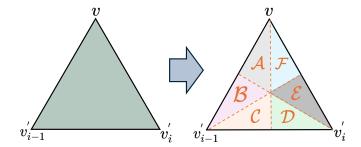


Fig. 4. Adversarial evolution sub-triangle partitioning. v represents the clean sample, v_{i-1}' represents the last adversarial example, and v_i' represents the current adversarial example. We conduct a more detailed investigation of this triangle by partitioning it into six sub-triangles based on the distance relationships among the clean example, the last adversarial example, and the current adversarial example.

from the adversarial image and adversarial evolution triangle, thereby reducing overfitting the source model and boosting the adversarial transferability. Moreover, we propose generating adversarial examples within the semantic image-text feature contrast space. This approach projects the original feature space into a semantically aligned subspace, reducing dependency on source models. The semantically coherent subspace enhances adversarial transferability across VLP models.

B. Adversarial Evolution Triangle for Adv Image Generation

Adversarial evolution triangle (AET): During the i-th iteration of the optimization, we can obtain the benign example \mathbf{x}_I , the adversarial example from the previous step $\tilde{\mathbf{x}}_I^{i-1}$ (previous adversarial example), and the current adversarial example $\tilde{\mathbf{x}}_I^i$. Then, these sample points together form a triangular region, referred to as the adversarial evolution triangle $\Delta \mathbf{x}_I \tilde{\mathbf{x}}_I^{i-1} \tilde{\mathbf{x}}_I^i$. Next, we can sample multiple instances from the adversarial evolution triangle $\Delta \mathbf{x}_I \tilde{\mathbf{x}}_I^{i-1} \tilde{\mathbf{x}}_I^i$, and denote the collection of samples as $\mathcal{S}^i = \left\{s_1^i, s_2^i, \dots, s_m^i\right\}$. Each sample can be calculated as follows:

$$s_k = \lambda \cdot \mathbf{x}_I + \beta \cdot \tilde{\mathbf{x}}_I^{i-1} + \gamma \cdot \tilde{\mathbf{x}}_I^i$$
, where $\lambda + \beta + \gamma = 1.0$, (3)

where λ , β , and γ represent the hyper-parameters, which decide how to do the sampling.

Text-guided augmentation selection: For our sampling set S^i , an adversarial perturbation direction can be calculated for each image. The k-th sample is denoted as s^i_k . Its perturbation direction can be obtained through the following:

$$\epsilon_{k}^{i} = \alpha \cdot \operatorname{sign}\left(\frac{\nabla_{s_{k}^{i}} J\left(s_{k}^{i}, \mathbf{x}_{T}\right)}{\left\|\nabla_{s_{k}^{i}} J\left(s_{k}^{i}, \mathbf{x}_{T}\right)\right\|}\right). \tag{4}$$

At this stage, we obtain a diverse set of perturbations, denoted as $\epsilon^i = \left\{\epsilon_1^i, \epsilon_2^i, \ldots, \epsilon_m^i\right\}$. To fully leverage modality interactions, we propose a text-guided augmentation selection approach to identify the optimal sample. Specifically, we integrate each element from the perturbation set ϵ^i into the adversarial image $\tilde{\mathbf{x}}_I^i$ individually. The objective is to determine which perturbation maximizes the distance between $\tilde{\mathbf{x}}_I^i$ and \mathbf{x}_T . This process can be formalized as follows:

$$o = \underset{\epsilon_k^i \in \epsilon^i}{\operatorname{arg\,max}} J\left(\tilde{\mathbf{x}}_I^i + \epsilon_k^i, \mathbf{x}_T\right). \tag{5}$$

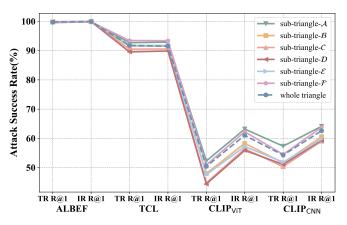


Fig. 5. Attack Success Rate (%) of different adversarial evolution subtriangles, which is used to boost the diversity of adversarial examples for improving adversarial transferability.

At this juncture, s_o^i represents the optimal sample.

Adversarial example generation: Through sampling and text-guided filtering, we have obtained the optimal sample s_o^i in the proposed adversarial evolution triangle $\triangle \mathbf{x}_I \tilde{\mathbf{x}}_I^{i-1} \tilde{\mathbf{x}}_I^i$. In addition, we follow SGA[42] as our baseline and integrate the image augmentation techniques (image resizing) during its optimization process. The chosen optimal sample s_o^i is augmented and expanded into the set $S_O^i = \left\{s_o^{i1}, s_o^{i2}, \ldots, s_o^{in}\right\}$. Subsequently, we use the extended set S_O^i to generate adversarial perturbation and update $\tilde{\mathbf{x}}_I^i$ to $\tilde{\mathbf{x}}_I^{i+1}$. This process can be formalized as follows:

$$\tilde{\mathbf{x}}_{I}^{i+1} = \prod_{\mathbf{x}_{I}, \epsilon_{I}} \left(\tilde{\mathbf{x}}_{I}^{i} + \alpha \cdot \operatorname{sign} \left(\frac{\nabla_{s_{o}^{i}} \sum_{j=1}^{n} J\left(s_{o}^{ij}, \mathbf{x}_{T}\right)}{\left\| \nabla_{s_{o}^{i}} \sum_{j=1}^{n} J\left(s_{o}^{ij}, \mathbf{x}_{T}\right) \right\|} \right) \right).$$
(6)

Adversarial evolution sub-triangle: To explore the impact of different adversarial evolution sub-triangle sampling on adversarial transferability, as shown in Figure 4, we divide the entire adversarial evolution triangle into six evolution sub-triangles based on their proximity to three key points: the clean example, the previous adversarial example, and the current adversarial example. Each evolution sub-triangle is characterized by which of these points is the nearest and which is the second nearest. The evolution sub-triangles are defined as follows:

- **Sub-triangle-***A***:** Nearest to the clean example; second nearest to the previous adversarial example.
- **Sub-triangle-***B***:** Nearest to the previous adversarial example; second nearest to the clean example.
- **Sub-triangle-***C*: Nearest to the previous adversarial example; second nearest to the current adversarial example.
- **Sub-triangle-***D*: Nearest to the current adversarial example; second nearest to the previous adversarial example.
- **Sub-triangle-**\$\mathcal{E}\$: Nearest to the current adversarial example; second nearest to the clean example.
- **Sub-triangle-** \mathcal{F} : Nearest to the clean example; second nearest to the current adversarial example.

We partition the adversarial evolution triangle into six distinct evolution sub-triangles by adjusting the hyperparameters β , γ , and λ . Specifically, these evolution sub-triangles are defined by all possible strict orderings of the three parameters: sub-triangle- \mathcal{A} is defined by $\gamma < \beta < \lambda$, sub-triangle- \mathcal{B} by $\gamma < \lambda < \beta$, sub-triangle- \mathcal{C} by $\lambda < \gamma < \beta$, sub-triangle- \mathcal{D} by $\lambda < \beta < \gamma$, sub-triangle- \mathcal{E} by $\beta < \lambda < \gamma$, and sub-triangle- \mathcal{F} by $\beta < \gamma < \lambda$. To study the impact of the adversarial evolution sub-triangles on adversarial transferability, we adopt ALBEF as the source model to generate multimodal adversarial examples and evaluate the adversarial transferability on other VLP models. The experiment results are shown in Figure 5. Analyses are as follows. Different adversarial evolution subtriangles can achieve different performances of adversarial transferability. Specifically, the transferability performance of sub-triangle- \mathcal{C} is the lowest among all, while sub-triangle- \mathcal{A} demonstrates higher transferability performance compared to both the whole triangle and the other sub-triangles. The results indicate that focusing sampling efforts on the sub-triangle near clean and previous adversarial examples significantly enhances adversarial transferability.

C. Adversarial Evolution Triangle for Adv Text Generation

As for the adversarial text generation, previous works generate adversarial texts in an iterative process by only deviating from the final generated adversarial image. Specifically, given the total iterations T and the final adversarial image $\tilde{\mathbf{x}}_I^T$, the adversarial text $\tilde{\mathbf{x}}_T$ is generated by deviating from the features of the final adversarial image. Since the final adversarial example completely depends on the source model, the generated adversarial text is only far away from the final adversarial example, resulting in the final multimodal adversarial example overfitting the source model. To address this issue, we propose to deviate from the last adversarial evolution triangle for generating the adversarial text. The evolution triangle consists of the clean image \mathbf{x}_I , the previous adversarial example $\tilde{\mathbf{x}}_I^{T-1}$, and the current adversarial example $\tilde{\mathbf{x}}_I^T$. The adversarial text can be calculated as follows:

$$\tilde{\mathbf{x}}_{T} = \underset{\tilde{\mathbf{x}}_{T} \in B[\mathbf{x}_{T}, \epsilon_{t}]}{\arg \max} \left(\kappa \cdot J\left(\mathbf{x}_{I}, \tilde{\mathbf{x}}_{T}\right) + \mu \cdot J\left(\tilde{\mathbf{x}}_{I}^{T-1}, \tilde{\mathbf{x}}_{T}\right) + \nu \cdot J\left(\tilde{\mathbf{x}}_{I}^{T}, \tilde{\mathbf{x}}_{T}\right) \right),$$

$$(7)$$

where κ , μ , and ν represent the hyper-parameters used to generate the adversarial text. They are constrained such that $\kappa + \mu + \nu = 1.0$.

D. Theoretical Analysis

Here we provide theoretical analysis to show that the proposed method can improve the transferability of adversarial examples compared with the SGA [42]. Wang *et al.* [58] introduce the Shapley value [50] to analyze the interactions inside adversarial perturbations. They discover that the adversarial transferability and the interaction inside adversarial perturbation have a negative correlation, *i.e.*, the adversarial examples with smaller interactions have better black-box transferability. These observations are also confirmed by the follow-up work like [60; 63; 65]. We give the Shapley interaction indices of the proposed method and the SGA [42] by the following theorem.

Theorem 1. The adversarial perturbations $\{\delta_t\}$ generated by the proposed method are given as

$$\boldsymbol{\delta}_t = \sum_{i=1}^t \boldsymbol{g}_i, \quad t = 2, 3, \dots,$$
 (8)

where

$$\boldsymbol{g}_{t} = \begin{cases} g(\boldsymbol{x} + \beta \cdot \boldsymbol{\delta}_{t-2} + \gamma \cdot \boldsymbol{\delta}_{t-1}), & \text{if } t \geq 2, \\ g(\boldsymbol{x}), & \text{if } t = 1, \end{cases}$$
(9)

 $g(\cdot)$ is the gradient of loss function $L(\cdot)$, β , $\gamma \in [0,1]$ are given constants, and $\delta_0 = (0,\ldots,0)$. Meanwhile, the adversarial perturbations generated by the SGA [42] $\{\zeta_t\}^{-1}$ are

$$\zeta_t = \sum_{i=1}^t h_i, \ h_i = g(x + \zeta_{i-1}), \ t = 1, \ 2, \dots$$
 (10)

Then the interaction inside adversarial perturbation $\boldsymbol{\delta}_t$ *will be*

$$\mathbb{E}_{i,j}[\boldsymbol{I}_{i,j}(\boldsymbol{\delta}_t)] = (\beta + \gamma)B \cdot t^3 + (A - 2\beta B) \cdot t^2 + (2A - (\beta + \gamma)B) \cdot t + A + 2\beta B,$$
(11)

where

$$A = \mathbb{E}_{i,j}[g(i) \cdot g(j) \cdot \boldsymbol{H}_{i,j}],$$

$$B = \mathbb{E}_{i,j}[g(i) \cdot \boldsymbol{H}_{i,j} \cdot g^{\top} \boldsymbol{H}_{*j}],$$
(12)

g = [g(1), ..., g(n)] is the gradient of L(x), $H_{i,j}$ is the (i, j) element of the Hessian matrix H for L(x), and H_{*j} is the j-th column of H. Moreover, the interaction inside ζ_t

$$\mathbb{E}_{i,j}[\boldsymbol{I}_{i,j}(\boldsymbol{\zeta}_t)] = B \cdot t^3 + (A - B) \cdot t^2. \tag{13}$$

Compared (11) with (13), it is clear that the interaction inside adversarial perturbation generated by the proposed method is lower than SGA [42] as $\beta + \gamma < 1$ and $t \geq 1$. Consequently, we know the proposed method will have better black-box transferability against the models with unknown parameters. We provide the proof details in the supplementary material.

E. Contrast Space Optimization for Semantic-aligned AET

We generate multimodal adversarial examples by optimizing the maximization of the loss function J of VLP models. Especially given the visual embeddings $F_I(\mathbf{x}_I) \in \mathbb{R}^d$, and the text embeddings $F_T(\mathbf{x}_T) \in \mathbb{R}^d$, where d represents the size of the feature embeddings, the loss function, abbreviated as J, in the original feature space can be calculated as:

$$J = \frac{F_I(\mathbf{x}_I) \cdot F_T(\mathbf{x}_T)}{d},\tag{14}$$

where \cdot represents the dot product of vectors, used to calculate the similarity between each pair of image and text features.

Previous works [71; 26] have indicated that image feature embeddings often carry substantial information unrelated to text features. In the original image-text contrast space, only a small subset of feature dimensions within the image embeddings is typically active, while the rest remain inactive and

 $^{1}{\rm SGA}$ [42] can be treated as a special case of the proposed method as $\beta=0$ and $\gamma=1.$

TABLE I

ATTACK SUCCESS RATE (%) OF THE PROPOSED METHOD WITH DIFFERENT PROPORTIONS OF TEXT DATA IN THE DATASET, WHICH ARE USED TO GENERATE THE PROJECTION MATRIX. WE USE ALBEF TO GENERATE MULTIMODAL ADVERSARIAL EXAMPLES ON THE FLICKR30K TO EVALUATE ADVERSARIAL TRANSFERABILITY.

Proportion	ALBEF		TCL		$ ext{CLIP}_{ ext{ViT}}$		$\mathrm{CLIP}_{\mathrm{CNN}}$	
roporuon	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
DRA [17]	99.90	99.93	91.57	91.17	46.26	56.80	49.55	59.01
Baseline	99.90	99.93	91.57	91.17	46.26	56.80	49.55	59.01
10%	100.0	100.0	94.73	94.69	47.73	58.96	53.38	61.51
20%	100.0	100.0	94.52	94.29	47.12	57.96	53.26	60.65
30%	100.0	100.0	95.15	94.26	47.98	58.02	53.51	60.79
40%	100.0	100.0	94.84	94.10	49.08	58.47	53.51	61.51
50%	100.0	100.0	94.52	94.19	48.34	57.51	51.47	61.20
60%	100.0	100.0	95.15	94.05	47.24	58.25	52.11	60.82
70%	100.0	100.0	94.73	94.55	47.85	57.64	52.49	60.62
80%	100.0	100.0	95.15	94.45	47.85	58.27	50.45	60.48
90%	100.0	100.0	94.42	94.26	47.24	57.64	52.11	60.48
100%	100.0	100.0	94.94	94.17	48.71	58.18	52.62	59.86

redundant. These redundant dimensions can impact similarity calculations, potentially distorting the feature alignment between images and text. The feature distortion may cause the generated adversarial examples to rely highly on the specific victim model, reducing their transferability. Hence, we propose a semantic image-text feature contrast space to generate multi-mode adversarial examples. It maps the original feature space into a semantic corpus subspace by using a semantic projection matrix. As shown in Figure 2 (c), assuming we have a d dimensional semantic representation space \mathbb{R}^d , we have \mathcal{N} semantic texts $U = [c_1, c_2, \ldots, c_{\mathcal{N}}]$. The text feature embeddings of those texts can be calculated as follows:

$$\mathcal{T} = [t_1, t_2, \dots, t_{\mathcal{N}}] \in \mathcal{R}^{\mathcal{N} \times d}, \tag{15}$$

where $t_i = F_T(c_i)$ for $i \in \{1, 2, ..., \mathcal{N}\}$. Then, we perform singular value decomposition (svd) [31] on these text features. It can be calculated as follows:

$$\mathcal{U}, \mathcal{S}, \mathcal{V} = \operatorname{svd}(\mathcal{T}).$$
 (16)

We can obtain $\mathcal{U} = [e_1, e_2, \dots, e_{\mathcal{N}}]$ as an orthonormal basis for the span of \mathcal{T} . The orthonormal basis \mathcal{U} can be utilized to compute a semantic projection matrix $\mathbf{P} \in \mathcal{R}^{d \times d}$, which can be defined as follows:

$$\mathbf{P} = \mathcal{U} \odot \mathcal{U}^{\top},\tag{17}$$

where \odot represents the matrix multiplication. Then, we can project the image and text features from the original space into the semantic subspace with the projection matrix **P**. They can be calculated as follows:

$$\widetilde{F_I(\mathbf{x}_I)} = F_I(\mathbf{x}_I) \odot \mathbf{P},
\widetilde{F_T(\mathbf{x}_T)} = F_T(\mathbf{x}_T) \odot \mathbf{P},$$
(18)

where $\widetilde{F_I(\mathbf{x}_I)}$ and $\widetilde{F_T(\mathbf{x}_T)}$ respectively represent the image feature embeddings and text feature embeddings after semantic projection. The loss function J, which is used to generate multimodal adversarial examples, can be rewritten as:

$$J = \frac{\left(F_I\left(\mathbf{x}_I\right) \odot \mathbf{P}\right) \cdot \left(F_T\left(\mathbf{x}_T\right) \odot \mathbf{P}\right)}{d},\tag{19}$$

The core of the semantic projection matrix is how to build an effective semantic corpus. In this paper, we construct the semantic corpus by creating a subset of the text-testing dataset, which is generated through random sampling of the complete testing dataset. To better understand the role of the semantic corpus, we analyze the impact of different sampling proportions $\mathcal{P} = \mathcal{N}/\mathcal{M}$, where \mathcal{N} represents the number of samples used to generate the matrix and $\mathcal M$ represents the number in the entire text dataset. We adopt ALBEF as the source model to generate the multimodal adversarial examples. TCL, CLIP_{CNN}, and CLIP_{ViT} are used as the target models to evaluate the adversarial transferability. We adopt varying proportions of text data from the Flickr30K dataset to generate the projection matrix, i.e., 20%, 40%, 60%, 80%, and 100%. The experiment results are shown in Table I. Analyses are summarized as follows. First, compared to the baseline DRA [17], which calculates the loss value in the original contrast space, the proposed loss function incorporating varying proportions of text data demonstrates improved adversarial transferability. It demonstrates the effectiveness of the proposed semantic contrast space in improving adversarial transferability. Second, the projection matrices generated from text data with varying proportions yield different levels of improvement in adversarial transferability, with 40% achieving the best improvement. Considering the computational expenses, we set the proportion to 40% for the experiments. In the appendix, we provide a detailed algorithm for the proposed SA-AET.

IV. EXPERIMENTS

To evaluate the effectiveness of the proposed method, we conduct extensive experiments across different datasets and VLP models. First, we introduce the detailed experimental setup, including the VLP models, image-text pair benchmark datasets, the settings of adversarial attack, and the evaluation metrics in **Section IV-A**. Subsequently, we introduce the selection of the optimal hyper-parameters in **Section IV-B**. Then, we present the comparative experimental results with state-of-the-art methods on the cross-model adversarial transferability within the image-text retrieval task in **Section IV-C**. After

TABLE II OPTIMAL HYPER-PARAMETERS SELECTION. ATTACK SUCCESS RATE (%) ON DIFFERENT SETTINGS, TOP FOR DIFFERENT VALUES OF κ, μ, ν IN Equation 7 and Bottom for different numbers of samples m.

Setting	ALI	BEF	TO	CL	CLIF	${ m CLIP_{ViT}}$		$\mathrm{CLIP}_{\mathrm{CNN}}$	
	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	
$\kappa, \mu, \nu = [0.2, 0.0, 0.8]$	99.9	99.98	96.42	96.05	51.53	59.47	52.23	61.41	
$[\kappa,\mu, u] = [0.2,0.2,0.6]$	99.9	99.98	96.21	96.10	52.27	59.60	52.75	61.44	
$[\kappa, \mu, \nu] = [0.2, 0.4, 0.4]$	99.9	99.98	96.00	96.10	52.52	59.60	52.87	61.13	
$[\kappa,\mu, u] = [0.2,0.6,0.2]$	99.9	99.98	96.00	96.07	52.76	59.73	53.38	61.23	
$[\kappa,\mu, u] = [0.2,0.8,0.0]$	99.9	100.0	95.89	96.05	53.13	59.95	53.64	61.17	
$[\kappa,\mu, u] = [0.4,0.0,0.6]$	99.9	99.98	96.84	96.02	53.50	62.34	55.43	63.88	
$[\kappa,\mu, u] = [0.4,0.2,0.4]$	99.9	100.0	96.63	96.02	53.62	62.21	55.30	63.84	
$[\kappa,\mu, u] = [0.4,0.4,0.2]$	99.9	100.0	96.94	96.02	54.48	62.34	55.30	63.98	
$[\kappa, \mu, \nu] = [0.4, 0.6, 0.0]$	99.9	99.98	96.84	96.05	54.60	62.21	55.30	63.95	
$[\kappa, \mu, \nu] = [0.6, 0.0, 0.4]$	99.9	99.98	96.42	96.05	55.09	63.56	57.22	65.49	
$[\kappa,\mu, u] = [0.6,0.2,0.2]$	99.9	99.98	96.42	96.02	55.71	63.69	57.22	65.63	
$[\kappa, \mu, \nu] = [0.6, 0.4, 0.0]$	99.9	99.98	96.52	96.05	55.45	63.72	57.09	65.56	
$[\kappa,\mu, u] = [0.8,0.0,0.2]$	99.9	100.0	96.63	95.93	54.85	64.18	56.45	65.63	
$[\kappa,\mu, u] = [0.8,0.2,0.0]$	99.9	99.98	96.73	95.83	54.97	64.47	56.32	65.49	
m=3	99.9	100.0	96.00	95.93	54.56	63.50	56.49	65.00	
m=4	99.9	100.0	96.10	96.00	54.83	63.65	56.64	64.83	
m = 5	99.9	99.98	96.42	96.02	55.58	63.89	57.22	65.59	
m = 6	99.9	100.0	95.89	96.21	54.23	63.56	57.24	65.19	
m = 7	99.9	100.0	96.33	96.17	55.58	63.95	56.98	65.15	

that, we present the cross-task transferability of multimodal adversarial examples from the image-text retrieval task to other tasks, along with adversarial transferability testing of current state-of-the-art LLMs in **Section IV-D** and **Section IV-E**, respectively. Additionally, we present the ablation study of the proposed method in **Section IV-F**. Finally, we introduce a knowledge transfer-based metric to offer an alternative perspective on the observed improvements in transferability in **Section IV-G**.

A. Experimental Settings

Benchmark datasets. We adopt two widely used imagetext pair datasets as the benchmark dataset to conduct experiments on the image-text retrieval task, i.e., Flickr30K [47] and MSCOCO [41]. The Flickr30K dataset contains 31,783 images, each annotated with five captions. Similarly, the MSCOCO dataset includes 123,287 images, each accompanied by approximately five captions. For the visual grounding task, we adopt the RefCOCO+ [62] dataset to conduct experiments. RefCOCO+ is a dataset with 141,564 referring expressions linked to 50,000 objects across 19,992 images from the MSCOCO collection. It is widely used to assess grounding models, focusing on how well they can locate objects described by natural language. For the Image Captioning task, which is another key challenge in visionand-language research, we utilize the MSCOCO dataset. This dataset is widely used because it provides rich images paired with diverse captions.

VLP models. We employ two standard architectures of VLP models, *i.e.*, fused and aligned VLP models, to carry out our experiments. In the case of aligned VLP models, CLIP [48] is used for the aligned VLP model. Specifically, we adopt different image encoders for CLIP, including ResNet-101 [24] and ViT-B/16 [13], *i.e.*, CLIP_{CNN} and CLIP_{ViT}, respectively.

In the case of fused VLP models, we adopt ALBEF [34] and TCL [61] to conduct our experiments. ALBEF incorporates a 12-layer ViT-B/16 as its image encoder, alongside two 6layer transformers dedicated to text and multimodal encoding. Although TCL adopts the same architectural framework, it distinguishes itself by employing different pre-training objectives. **Adversarial attack settings.** Following previous works [42; 17; 36], for the text adversarial attack, we set the text perturbation bound $\epsilon_T = 1$, with the word list W containing 10 words. For the image adversarial attack, we set the image perturbation bound $\epsilon_I = 8/255$ under L_{∞} norm. Moreover, the number of iteration steps T is set to 10, with each step size $\alpha = 2/255$. We follow the image augmentation techniques in SGA during its optimization process, enlarging the image dataset by resizing the original images to five different scales: {0.50, 0.75, 1.00, 1.25, 1.50}, utilizing bicubic interpolation. We compare the proposed SA-AET with the following baselines: (1) PGD [44], (2) Bert-Attack [35], (3) Sep-Attack [66], (4) Co-Attack [66], (5) SGA [42], and (6) DRA [17].

Evaluation metrics. We adopt the Attack Success Rate (ASR) at the top-1 rank (R@1) as the metric to evaluate the adversarial transferability. It represents the percentage of successful attacks among all generated adversarial examples. A higher ASR indicates a more effective attack with higher transferability.

B. Optimal Hyper-Parameters Selection

The number of samples m taken from the sub-triangle- \mathcal{A} of the adversarial trajectory and scaling factors in Equation 7 are the hyper-parameters. We conduct a series of experiments to determine the optimal value of hyper-parameters and analyze their impact on the overall performance of the proposed method. Specifically, we employ ALBEF to generate

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE FLICKR30K DATASET FOR THE IMAGE-TEXT RETRIEVAL TASK. THE SOURCE COLUMN SHOWS VLP MODELS WE USE TO GENERATE MULTIMODAL ADVERSARIAL EXAMPLES. FOR BOTH IMAGE RETRIEVAL AND TEXT RETRIEVAL, WE PROVIDE AN R@1 attack success rate (%).

	<u> </u>	ALI	BEF	TO	CL	CLI	$P_{ m ViT}$	${f CLIP}_{{ m CNN}}$	
Source	Attack	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
	PGD [44]	93.74	94.43	24.03	27.90	10.67	15.82	14.05	19.11
	BERT-Attack [35]	11.57	27.46	12.64	28.07	29.33	43.17	32.69	46.11
	Sep-Attack [66]	95.72	96.14	39.30	51.79	34.11	45.72	35.76	47.92
ALBEF	Co-Attack [66]	97.08	98.36	39.52	51.24	29.82	38.92	31.29	41.99
	SGA [42]	99.90	99.98	87.88	88.05	36.69	46.78	39.59	49.78
	DRA [17]	99.90	99.93	91.57	91.17	46.26	56.80	49.55	59.01
	SA-AET (ours)	99.90	99.98	96.42	96.02	55.58	63.89	57.22	65.59
	PGD [44]	35.77	41.67	99.37	99.33	10.18	16.3	14.81	21.10
	BERT-Attack [35]	11.89	26.82	14.54	29.17	29.69	44.49	33.46	46.07
	Sep-Attack [66]	52.45	61.44	99.58	99.45	37.06	45.81	37.42	49.91
TCL	Co-Attack [66]	49.84	60.36	91.68	95.48	32.64	42.69	32.06	47.82
	SGA [42]	92.49	92.77	100.0	100.0	36.81	46.97	41.89	51.53
	DRA [17]	95.20	95.58	100.0	99.98	47.24	57.28	52.23	62.23
	SA-AET (ours)	98.85	98.50	100.0	100.0	56.20	63.47	59.77	67.86
	PGD [44]	3.13	6.48	4.43	8.83	69.33	84.79	13.03	17.43
	BERT-Attack [35]	9.59	22.64	11.80	25.07	28.34	39.08	30.40	37.43
	Sep-Attack [66]	7.61	20.58	10.12	20.74	76.93	87.44	29.89	38.32
$\text{CLIP}_{\mathrm{ViT}}$	Co-Attack [66]	8.55	20.18	10.01	21.29	78.53	87.50	29.50	38.49
	SGA [42]	22.42	34.59	25.08	36.45	100.0	100.0	53.26	61.10
	DRA [17]	27.84	42.84	27.82	44.60	100.0	100.0	64.88	69.50
	SA-AET (ours)	36.60	50.44	39.20	51.10	100.0	100.0	71.01	74.10
	PGD [44]	2.29	6.15	4.53	8.88	5.40	12.08	89.78	93.04
	BERT-Attack [35]	8.86	23.27	12.33	25.48	27.12	37.44	30.40	40.10
	Sep-Attack [66]	9.38	22.99	11.28	25.45	26.13	39.24	93.61	95.30
$CLIP_{\mathrm{CNN}}$	Co-Attack [66]	10.53	23.62	12.54	26.05	27.24	40.62	95.91	96.50
	SGA [42]	15.64	28.60	18.02	33.07	39.02	51.45	99.87	99.90
	DRA [17]	19.50	34.59	21.60	37.88	48.47	59.12	99.87	99.90
	SA-AET (ours)	23.98	38.28	27.29	41.81	54.11	64.21	100.0	99.97

multimodal adversarial examples on the Flickr30K dataset and evaluate their transferability to three other VLP models.

Number of samples m taken from the sub-triangle-A of the adversarial trajectory: The results are shown at the bottom of Table II. It can be observed that transfer effects become apparent for both the Image Retrieval and Text Retrieval tasks when the sample size reaches 5. Beyond this point, as the sample size increases, transferability shows fluctuations without any significant enhancement. Considering the balance between transfer effects and computational efficiency, a sample size of 5 is the most effective hyperparameter.

Scaling factors κ, μ, ν in adversarial text generation: In Equation 7, we introduce three hyper-parameters that represent the weights of the clean image, the previous adversarial example, and the current adversarial example, respectively. We stipulate that the proportion of adversarial images cannot be zero, i.e., $\mu + \nu \neq 0$. Under the above conditions, we assign κ, μ, ν to 14 different sets of values. Using ALBEF as the surrogate model, we generate adversarial examples and evaluate their transferability to the remaining three models. The detailed results are shown at the top of Table II. Adversarial transferability from ALBEF to TCL shows minimal variation across settings (around 1%), while transferability to CLIP models varies significantly, with a performance gap between 4.18% and 5.0%. Thus, our hyper-parameter selection focuses on maximizing transferability to CLIP models. Increasing κ , which introduces a higher proportion of clean images, initially boosts transferability, but an excessive κ reduces the adversarial ratio, decreasing overall effectiveness. Based on Table II, we select $[\kappa, \mu, \nu] = [0.6, 0.2, 0.2]$ for optimal transferability.

C. Cross-model Adversarial Transferability

We adopt four widely used VLP models, *i.e.*, ALBEF, TCL, CLIP $_{\rm ViT}$, and CLIP $_{\rm CNN}$, to conduct comparative experiments on the image-text retrieval task. Specifically, we adopt one of four VLP models to generate multimodal adversarial examples and then use the other three VLP models to evaluate the transferability of adversarial examples. We compare the proposed SA-AET with a series of advanced adversarial attack methods for VLP models on the Flickr30K and MSCOCO datasets.

Performance on the Flickr30K dataset. Comparative experimental results on the Flickr30K dataset are shown in Table III. It can be observed that our SA-AET, along with SGA and DRA, consistently outperforms other adversarial attack methods under all white box attack scenarios. Regardless of whether using TR or IR, ASR consistently surpasses 99.8%. Compared with previous works, the proposed SA-AET achieves the best performance of adversarial transferability across different VLP models under all adversarial attack scenarios. For example, when ALBEF is used as the source model to generate multimodal adversarial examples, the advanced SGA method achieves an ASR of 87.88% on TCL, 36.69% on

TABLE IV

Comparison with state-of-the-art methods on the MSCOCO dataset for the image-text retrieval task. The source column shows VLP models we use to generate multimodal adversarial examples. For both Image Retrieval and Text Retrieval, we provide an R@1 attack success rate (%).

_		ALI	BEF	TO	CL	CLI	$P_{ m ViT}$	CLII	CNN
Source	Attack	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
I	PGD [44]	94.35	93.26	34.15	36.86	21.71	27.06	23.83	30.96
	BERT-Attack [35]	24.39	36.13	24.34	33.39	44.94	52.28	47.73	54.75
	Sep-Attack [66]	97.01	96.59	61.06	66.13	57.19	65.82	58.81	68.61
ALBEF	Co-Attack [66]	65.22	72.41	94.95	97.87	55.28	62.33	56.68	66.45
	SGA [42]	99.95	99.94	87.46	88.17	63.72	69.71	63.91	70.78
	DRA [17]	99.90	99.93	88.81	90.06	69.25	75.31	68.53	75.09
	SA-AET (ours)	100.0	99.99	97.28	96.88	76.57	80.24	76.17	80.64
	PGD [44]	40.81	44.09	98.54	98.20	21.79	26.92	24.97	32.17
	BERT-Attack [35]	35.32	45.92	38.54	48.48	51.09	58.80	52.23	61.26
TCL	Sep-Attack [66]	66.38	72.19	99.15	98.98	59.94	65.95	60.77	69.37
	Co-Attack [66]	49.84	60.36	91.68	95.48	32.64	42.69	32.06	47.82
	SGA [42]	92.70	92.99	100.0	100.0	59.79	65.31	60.52	67.34
	DRA [17]	94.72	95.89	100.0	100.0	70.51	74.95	70.29	76.99
	SA-AET (ours)	97.78	98.08	100.0	99.99	76.12	79.74	75.89	80.92
	PGD [44]	10.26	13.69	12.72	15.81	82.91	90.51	21.62	28.78
	BERT-Attack [35]	20.34	29.74	21.08	29.61	45.06	51.68	44.54	53.72
CLIP _{ViT}	Sep-Attack [66]	25.91	36.84	28.20	38.47	88.36	97.09	47.57	57.79
	Co-Attack [66]	26.35	36.69	88.78	96.72	28.23	38.42	47.36	58.45
	SGA [42]	43.75	51.08	44.05	51.02	100.0	100.0	70.66	75.58
	DRA [17]	52.69	61.50	51.88	61.06	100.0	100.0	80.18	84.11
	SA-AET (ours)	57.64	66.88	57.30	65.16	100.0	100.0	83.98	86.72
	PGD [44]	8.38	12.73	11.90	15.68	13.66	20.62	92.68	94.71
	BERT-Attack [35]	23.38	34.64	24.58	29.61	51.28	57.49	54.43	62.17
	Sep-Attack [66]	29.13	40.64	31.40	42.99	52.23	59.73	96.16	97.54
$CLIP_{CNN}$	Co-Attack [66]	29.49	41.50	31.83	43.44	53.15	60.15	97.79	98.54
	SGA [42]	36.94	46.79	38.81	48.90	62.19	67.73	99.92	99.97
	DRA [17]	41.40	52.25	43.62	54.15	70.43	74.14	99.80	99.92
	SA-AET (ours)	43.62	55.19	47.01	57.39	73.67	76.90	100.0	99.92

TABLE V

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.

		$ITR \to VG$		$\mathbf{ITR} \to \mathbf{IC}$					
Attack	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 [42]	50.56	57.42	40.66	28.0	24.6	51.2	91.4	17.7	
DRA [17]	49.70	56.32	40.54	27.2	24.2	50.7	88.3	17.2	
SA-AET (ours)	47.44	53.27	38.58	21.0	20.5	45.2	65.7	13.6	

CLIP $_{\rm ViT}$, and 39.59% on CLIP $_{\rm CNN}$ for text retrieval tasks. For image retrieval tasks, it attains an ASR of 88.05% on TCL, 46.78% on CLIP $_{\rm ViT}$, and 49.78% on CLIP $_{\rm CNN}$. In contrast, the proposed SA-AET method outperforms SGA with an ASR of 96.42% on TCL, 55.58% on CLIP $_{\rm ViT}$, and 57.22% on CLIP $_{\rm CNN}$ for text retrieval tasks. For image retrieval tasks, SA-AET achieves an ASR of 96.02% on TCL, 63.89% on CLIP $_{\rm ViT}$, and 65.59% on CLIP $_{\rm CNN}$. Moreover, when applying ALBEF to target CLIP $_{\rm CNN}$, the proposed SA-AET improves the ASR by 7.67% for TR R@1 and 6.58% for IR R@1, compared to DRA. The results demonstrate the effectiveness of our proposed method in improving the transferability of multimodal adversarial examples on the Flickr30K dataset.

Performance on the MSCOCO dataset. Comparative experimental results on the MSCOCO are shown in Table IV. It can

also be observed that compared with previous works, the proposed method achieves the best performance in improving the adversarial transferability for VLP models. Specifically, when using CLIP $_{\rm ViT}$ to generate adversarial examples, the advanced SGA method demonstrates an ASR of 43.75% on ALBEF, 44.05% on TCL, and 70.66% on CLIP $_{\rm CNN}$ for text retrieval tasks. For image retrieval tasks, it achieves an ASR of 51.08% on ALBEF, 51.02% on TCL, and 75.58% on CLIP $_{\rm CNN}$. Conversely, the proposed SA-AET method surpasses SGA, achieving an ASR of 57.64% on ALBEF, 57.3% on TCL, and 83.98% on CLIP $_{\rm CNN}$ for text retrieval tasks. In image retrieval tasks, SA-AET records ASRs of 66.88% on ALBEF, 65.16% on TCL, and 86.72% on CLIP $_{\rm CNN}$. Additionally, when applying CLIP $_{\rm ViT}$ to target TCL, the proposed SA-AET improves the ASR by 5.42% for TR R@1 and 4.1% for IR

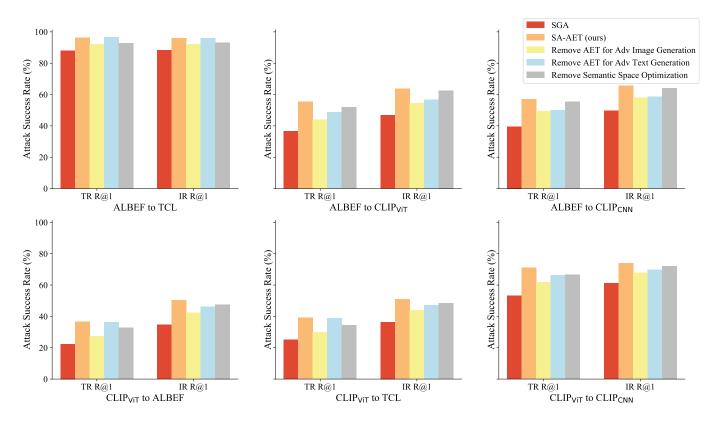


Fig. 6. Ablation Study. 'Remove AET for Adv Image Generation' refers to eliminating the proposed adversarial evolution triangle for adversarial image generation. 'Remove AET for Adv Image Generation' refers to eliminating the proposed adversarial evolution triangle for adversarial text generation. 'Remove Semantic Space Optimization' refers to eliminating the proposed semantic contrast space.



Fig. 7. Visualization on Image Captioning Task. We use the ALBEF model, pre-trained on Image Text Retrieval (ITR) task, to generate adversarial images on the MSCOCO dataset and use the BLIP [33] model for Image Captioning on both clean images and adversarial images, respectively.

R@1, compared to DRA. The results highlight the effectiveness of our proposed method in enhancing the transferability of multimodal adversarial examples on the MSCOCO dataset.

D. Cross-task Adversarial Transferability

We not only evaluate the transferability of multimodal adversarial examples generated by our proposed method across

TABLE VI
TRANSFERABILITY PERFORMANCE ON THE ADVANCED COMMERCIAL
MULTIMODAL LARGE LANGUAGE MODELS.

Attack	GPT-4 R@1	Claude-3 R@1	Qwen-VL R@1
No Attack	0.0	0.0	0.0
SGA [42]	6.0	18.0	6.0
DRA [17]	16.0	22.0	15.0
SA-AET (ours)	18.0	24.0	16.0

different VLP models but also conduct experiments to evaluate its effectiveness in transferring across diverse V+L tasks. Specifically, we generate adversarial examples for the Image-Text Retrieval (ITR) task and evaluate them on Visual Grounding (VG) and Image Captioning (IC) tasks. As evident from Table V and visual results in Figures 7 and 8, the adversarial examples generated for ITR demonstrate transferability, successfully impacting both VG and IC tasks. Compared to SGA and DRA, adversarial examples generated by SA-AET lead to a substantial decline in metrics for both VG and IC tasks, with a notable drop of 25.7 and 22.6 points in the CIDEr score on the IC task, respectively. This demonstrates the superior effectiveness of SA-AET in transfer attacks, emphasizing its crosstask transferability. Moreover, our transferability consistently surpasses that of all other methods.

E. Adversarial Transferability in MLLMs

Multimodal Large Language Models (MLLMs) have recently gained significant attention for their wide range of appli-

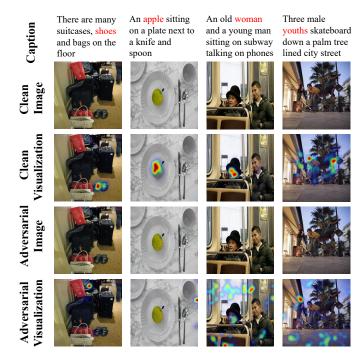


Fig. 8. Visualization on Visual Grounding Task. We use the ALBEF model, pre-trained on the ITR task, to generate adversarial images on the RefCOCO+ dataset and use the same model, pre-trained on Visual Grouding (VG) task, to localize the regions corresponding to red words on both clean images and adversarial images, respectively.

cations, demonstrating remarkable versatility and effectiveness in handling complex tasks. To further investigate the adversarial robustness of MLLMs, we conduct a series of extension experiments. Using ALBEF as a surrogate model, we generate adversarial examples under constrained perturbation settings, with a magnitude of 16/255 and a single-step perturbation of 0.5/255, iterating over 100 steps. These adversarial examples are then evaluated on advanced MLLMs (e.g., GPT-4 [2], Claude-3 [4]) by prompting them with the query "Describe this image". As shown in Figure 9, it illustrates that adversarial examples generated by our method are capable of deceiving state-of-the-art MLLMs, leading them to produce incorrect responses. In addition to these qualitative results on MLLMs, we further extend our experiments by randomly selecting 100 images from the Flick30K dataset. We generate adversarial images using SGA, DRA, and our method SA-AET, with ALBEF as the surrogate model, under the same perturbation setting (16/255). These adversarial images are then fed into MLLMs, where the query is prompted with "Please describe this image, limited to 50 words". We rank the generated descriptions against 100 irrelevant texts based on CLIP similarity. If a description does not rank first, we consider it a successful attack. The results are shown in Table VI. It can be observed that our SA-AET achieves the best transferability performance among advanced commercial MLLMs.

F. Ablation Study

Our proposed method builds upon the SGA [42] method and introduces three significant improvements: 1) the adversarial evolution triangle for adversarial image generation, 2) the

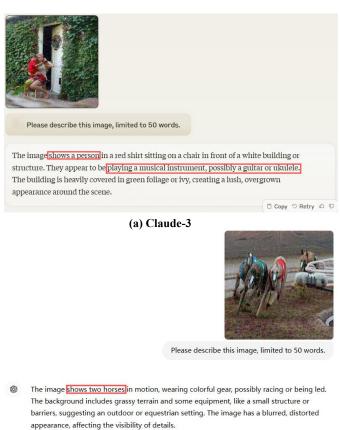


Fig. 9. Adversarial Transferability on Multimodal Large Language Models (MLLMs). We input adversarial images along with the query "Please describe this image, limited to 50 words." into MLLMs. The upper section represents the results from Claude-3, while the lower section shows the results from GPT-4. Incorrect descriptions are highlighted in red boxes.

(b) GPT-4

0 0 0 0 0 CV

adversarial evolution triangle for adversarial text generation, and 3) the contrast space optimization for semantic-aligned adversarial evolution triangle. To evaluate the contribution of each element, we conduct ablation studies on the image-text retrieval (ITR) task and evaluate the transferability of multimodal adversarial examples from two different architectures: the fused and aligned architectures, corresponding to ALBEF and CLIP_{ViT}, respectively. We then test how well these adversarial examples generalize to other VLP models. The results of the ablation studies can be seen in Figure 6. In the ablation studies, we compare five methods. The first is SGA [42], which serves as our baseline, followed by SA-AET, representing our complete method. The other three methods are derived from SA-AET by sequentially removing each of the three key contributions we introduce. 'Remove AET for Adv Image Generation' refers to eliminating the proposed adversarial evolution triangle for adversarial image generation. 'Remove AET for Adv Text Generation' refers to eliminating the proposed adversarial evolution triangle for adversarial text generation. 'Remove Semantic Space Optimization' refers to eliminating the proposed semantic contrast space. It is evident from the six sets of transfer attack experiments on ALBEF and $CLIP_{ViT}$ that removing any of the improvements from

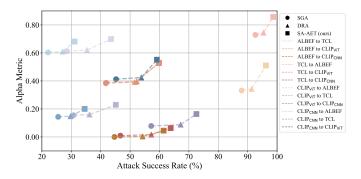


Fig. 10. Attack Success Rate vs. Alpha Metric in Transfer Attacks on VLP Models. The horizontal axis represents the attack success rate of transfer attacks on the target model, while the vertical axis represents the alpha metric.

SA-AET results in weaker transfer attacks compared to the full SA-AET method. It demonstrates the effectiveness of each element of our SA-AET.

G. Performance Analysis

In addition to examining the transfer attack success rates of multimodal adversarial examples generated by various VLP models, we introduce a knowledge transfer-based metric proposed by Liang *et al.* [38] to offer an alternative perspective on the observed improvements in transferability and to provide insights into the underlying reasons for these enhancements, from the baseline method, SGA [42], to the advanced approaches, DRA [17] and SA-AET. The metric, known as the alpha metric, represents how often the adversarial attack transfers, which can be calculated as follows:

$$\frac{J\left(F_{I}\left(\mathbf{x}_{I}+\epsilon^{sur}\right),F_{T}\left(\tilde{\mathbf{x}}_{T}^{sur}\right)\right)}{J\left(F_{I}\left(\mathbf{x}_{I}+\epsilon^{tar}\right),F_{T}\left(\tilde{\mathbf{x}}_{T}^{tar}\right)\right)}.$$
(20)

In the above formula, F_I and F_T represent the image and text encoders of the target VLP model for transfer attacks, respectively, while $J(\cdot)$ denotes the loss function of the VLP model. ϵ^{sur} and $\tilde{\mathbf{x}}_T^{sur}$ represent the adversarial image perturbation and adversarial text generated on the surrogate model, respectively, while ϵ^{tar} and $\tilde{\mathbf{x}}_T^{tar}$ denote the adversarial image perturbation and adversarial text generated on the target model. In this context, the alpha metric can be viewed as the difference in loss between the multimodal adversarial examples generated on the surrogate model and its own multimodal adversarial examples. To measure the alpha metric, we select the Flickr30k dataset and utilize 12 sets of transfer attacks from four chosen VLP models, calculating the metric according to Formula (20). We present the final results in Figure 10. Notably, the x-axis represents the attack success rates of the three methods: SGA, DRA, and SA-AET. These success rates are obtained by averaging the TR R@1 and IR R@1 metrics. The alpha metric is calculated as the average across all Flickr30k data.

Based on the results presented in Figure 10, we find that SA-AET consistently outperforms both SGA and DRA, demonstrating significantly higher attack success rates and alpha metrics across all 12 sets of transfer attacks. Furthermore, the DRA method exhibits favorable performance, with both

the attack success rates and alpha metrics for DRA and SA-AET showing substantial improvements over SGA. This finding suggests that our contribution to increasing adversarial transferability in both DRA and SA-AET is effective. While the difference in alpha metrics between DRA and SGA is relatively modest, there is a significant gap between SA-AET and the other two methods. The transfer attack success rates of DRA compared to SGA have already reached a high level, making further improvements challenging. However, by introducing semantic-aligned adversarial sub-triangles, we achieve a notable increase in the alpha metric. This indicates that we enhance how often the adversarial attack transfers, thereby further improving adversarial transferability.

V. CONCLUSION

In this paper, we focus on generating highly transferable adversarial examples (AEs) for vision-language pre-training (VLP) models. We find that previous methods focus only on augmenting image-text pairs around current adversarial examples, which offers limited improvements in transferability. To overcome this limitation, we propose leveraging the intersection regions along the adversarial trajectory during optimization, specifically by sampling from adversarial evolution triangles that are composed of clean, historical, and current adversarial examples. It significantly enhances the diversity of AEs, leading to improved transferability. Additionally, we tackle the issue of feature-matching distortion caused by redundant image features in VLP models. Adversarial examples generated in the original image-text feature contrast space are highly dependent on the victim model. We propose a semantic image-text feature contrast space to mitigate this, projecting the original features into a semantically aligned subspace. This projection reduces feature redundancy and enhances the transferability of adversarial examples. Our theoretical analysis validates the effectiveness of the proposed adversarial evolution triangles, and extensive experiments across various datasets and models demonstrate that our method outperforms state-of-the-art adversarial attack methods for VLP models.

REFERENCES

- [1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1348–1357, 2023.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- [3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen 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*, pages 382–398. Springer, 2016.
- [4] Anthropic. Claude 3 haiku: our fastest model yet. 2024.

- [5] Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. Improving query efficiency of black-box adversarial attack. In *Computer Vision–ECCV* 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pages 101–116. Springer, 2020.
- [6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [7] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663, 2020.
- [8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In European conference on computer vision, pages 104– 120. Springer, 2020.
- [9] Hao Cheng, Erjia Xiao, Jiahang Cao, Le Yang, Kaidi Xu, Jindong Gu, and Renjing Xu. Typography leads semantic diversifying: Amplifying adversarial transferability across multimodal large language models. arXiv preprint arXiv:2405.20090, 2024.
- [10] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiong-wei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2022.
- [11] Nhat Chung, Sensen Gao, Tuan-Anh Vu, Jie Zhang, Aishan Liu, Yun Lin, Jin Song Dong, and Qing Guo. Towards transferable attacks against vision-llms in autonomous driving with typography. arXiv preprint arXiv:2405.14169, 2024.
- [12] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. arXiv preprint arXiv:2010.11929, 2010.
- [14] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.
- [15] Jiyuan Fu, Zhaoyu Chen, Kaixun Jiang, Haijing Guo,

- Jiafeng Wang, Shuyong Gao, and Wenqiang Zhang. Improving adversarial transferability of visual-language pre-training models through collaborative multimodal interaction. *arXiv* preprint arXiv:2403.10883, 2024.
- [16] Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. *arXiv preprint* arXiv:2405.09981, 2024.
- [17] Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. *arXiv preprint arXiv:2403.12445*, 2024.
- [18] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. arXiv preprint arXiv:2310.17626, 2023.
- [19] Qing Guo, Ziyi Cheng, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yang Liu, and Jianjun Zhao. Learning to adversarially blur visual object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10839–10848, 2021.
- [20] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, and Yang Liu. Watch out! motion is blurring the vision of your deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 975–985, 2020.
- [21] Dongchen Han, Xiaojun Jia, Yang Bai, Jindong Gu, Yang Liu, and Xiaochun Cao. Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. *arXiv preprint arXiv:2312.04403*, 2023.
- [22] Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. Sa-attack: Improving adversarial transferability of vision-language pretraining models via self-augmentation. *arXiv preprint arXiv:2312.04913*, 2023.
- [23] Bangyan He, Jian Liu, Yiming Li, Siyuan Liang, Jingzhi Li, Xiaojun Jia, and Xiaochun Cao. Generating transferable 3d adversarial point cloud via random perturbation factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 764–772, 2023.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [25] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17980– 17989, June 2022.
- [26] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pretraining with source free domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications

- of Computer Vision, pages 2994-3003, 2024.
- [27] Yihao Huang, Liangru Sun, Qing Guo, Felix Juefei-Xu, Jiayi Zhu, Jincao Feng, Yang Liu, and Geguang Pu. Ala: Naturalness-aware adversarial lightness attack. In Proceedings of the 31st ACM International Conference on Multimedia, pages 2418–2426, 2023.
- [28] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han. Adv-watermark: A novel watermark perturbation for adversarial examples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1579–1587, 2020.
- [29] Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao Sr. Improving fast adversarial training with prior-guided knowledge. *arXiv* preprint arXiv:2304.00202, 2023.
- [30] Zaid Khan and Yun Fu. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042, 2021.
- [31] Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980.
- [32] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. In Proceedings of the 29th ACM International Conference on Multimedia, pages 2567–2576, 2021.
- [33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [34] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34:9694–9705, 2021.
- [35] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert, 2020.
- [36] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability via intermediate-level perturbation decay. *Advances in Neural Information Processing Systems*, 36, 2024.
- [37] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, pages 121–137. Springer, 2020.
- [38] Kaizhao Liang, Jacky Y Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability. In *International Conference on Machine Learning*, pages 6577–6587. PMLR, 2021.

- [39] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems, 35:17612–17625, 2022.
- [40] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence 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, pages 740–755. Springer, 2014.
- [42] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of visionlanguage pre-training models, 2023.
- [43] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv* preprint arXiv:2403.09766, 2024.
- [44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.
- [45] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. Advances in Neural Information Processing Systems, 34:23296–23308, 2021.
- [46] Jeonghwan Park, Paul Miller, and Niall McLaughlin. Hard-label based small query black-box adversarial attack. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3986–3995, 2024.
- [47] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015.
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [50] Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, Contributions to the Theory of Games II, pages 307–317. Princeton University Press, 1953.

- [51] Lei Shi, Kai Shuang, Shijie Geng, Peng Gao, Zuohui Fu, Gerard de Melo, Yunpeng Chen, and Sen Su. Dense contrastive visual-linguistic pretraining. In *Proceedings* of the 29th ACM International Conference on Multimedia, pages 5203–5212, 2021.
- [52] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* preprint arXiv:1908.07490, 2019.
- [53] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.
- [54] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on* computer vision and pattern recognition, pages 4566– 4575, 2015.
- [55] Teng Wang, Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie, and Ping Luo. Accelerating visionlanguage pretraining with free language modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23161–23170, 2023.
- [56] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learn-ing*, pages 22680–22690. PMLR, 2022.
- [57] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International* Conference on Computer Vision, pages 16158–16167, 2021.
- [58] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A unified approach to interpreting and boosting adversarial transferability. In *International Conference on Learning Representations*, 2021.
- [59] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Crossmodal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5764–5773, 2019.
- [60] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, Yu-Gang Jiang, and Larry S. Davis. Towards transferable adversarial attacks on image and video transformers. *IEEE Transactions on Image Processing*, 32:6346–6358, 2023.
- [61] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.
- [62] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara 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, pages 69–85. Springer, 2016.
- [63] Wenqian Yu, Jindong Gu, Zhijiang Li, and Philip Torr. Reliable evaluation of adversarial transferability. arXiv preprint arXiv:2306.08565, 2023.
- [64] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.
- [65] Zheng Yuan, Jie Zhang, Zhaoyan Jiang, Liangliang Li, and Shiguang Shan. Adaptive perturbation for adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5663–5676, 2024.
- [66] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.
- [67] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael 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*, pages 8173–8182, 2023.
- [68] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in visionlanguage models. In *Proceedings of the IEEE/CVF* conference on computer vision and pattern recognition, pages 5579–5588, 2021.
- [69] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3536–3545, 2020.
- [70] Yechao Zhang, Shengshan Hu, Leo Yu Zhang, Junyu Shi, Minghui Li, Xiaogeng Liu, Wei Wan, and Hai Jin. Why does little robustness help? a further step towards understanding adversarial transferability. In 2024 IEEE Symposium on Security and Privacy (SP), pages 3365–3384. IEEE, 2024.
- [71] Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. Clip-pae: projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023
- [72] Rongyi Zhu, Zeliang Zhang, Susan Liang, Zhuo Liu, and Chenliang Xu. Learning to transform dynamically for better adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24273–24283, 2024.

APPENDIX

DETAILED ALGORITHM

Algorithm 1 Attack formulation.

Input: Image encoder F_I , Text encoder F_T , Image-caption pair $(\mathbf{x}_I, \mathbf{x}_T)$, iteration steps T, number of samples m, image maximal perturbation bound ϵ_I , text maximal number of changeable words ϵ_T , single-step perturbation size α , image augmentation scale set $\mathcal{I} = \{I_1, I_2, ..., I_n\}$

Output: Adversarial Image $\tilde{\mathbf{x}}_I^T$, Adversarial Text $\tilde{\mathbf{x}}_T$

- 1: /*Follow SGA and get $\tilde{\mathbf{x}}_I^0$ and $\tilde{\mathbf{x}}_I^1$ */
- 2: $\tilde{\mathbf{x}}_I^0 = \mathbf{x}_I + \epsilon_I \cdot N(0, 1)$

2:
$$\mathbf{x}_{I} = \mathbf{x}_{I} + \epsilon_{I} \cdot \mathcal{N}(0, 1)$$

3: Construct the image augmentation set $\left\{\tilde{\mathbf{x}}_{I}^{01}, \tilde{\mathbf{x}}_{I}^{02}, \dots, \tilde{\mathbf{x}}_{I}^{0n}\right\}$ for $\tilde{\mathbf{x}}_{I}^{0}$ using the scales in \mathcal{I} .
4: $\tilde{\mathbf{x}}_{I}^{1} = \prod_{\mathbf{x}_{I}, \epsilon_{I}} \left(\tilde{\mathbf{x}}_{I}^{0} + \alpha \cdot \operatorname{sign}\left(\frac{\nabla_{\mathbf{x}_{I}} \sum_{j=1}^{n} J(\tilde{\mathbf{x}}_{I}^{0j}, \mathbf{x}_{T})}{\left\|\nabla_{\mathbf{x}_{I}} \sum_{j=1}^{n} J(\tilde{\mathbf{x}}_{I}^{0j}, \mathbf{x}_{T})\right\|}\right)\right)$.
5: /* Adversarial Image Generation */

- 6: **for** i = 1 to T 1 **do**
- Sample n instances from adversarial evolution sub-triangle- \mathcal{A} and build set $\mathcal{S}^i = \{s_1^i, s_2^i, \dots, s_m^i\}$.
- $$\begin{split} s_k &= \lambda \cdot \mathbf{x}_I + \beta \cdot \tilde{\mathbf{x}}_I^{i-1} + \gamma \cdot \tilde{\mathbf{x}}_I^i, \text{ where } \lambda + \beta + \gamma = 1.0 \text{ and } \lambda > \beta > \gamma. \\ \text{Build perturbation set } \epsilon^i &= \left\{ \epsilon_1^i, \epsilon_2^i, \dots, \epsilon_m^i \right\}. \end{split}$$

10:
$$\epsilon_k^i = \alpha \cdot \text{sign}\left(\frac{\nabla_{s_k^i} J(s_k^i, \mathbf{x}_T)}{\left\|\nabla_{s_k^i} J(s_k^i, \mathbf{x}_T)\right\|}\right).$$

- Text-guided augmentation selection: obtain the optimal sample s_o^i . 11:
- $o = \arg \max J\left(\tilde{\mathbf{x}}_I^i + \epsilon_k^i, \mathbf{x}_T\right).$ 12:
- 13:

13: Construct the image augmentation set
$$\mathcal{S}_O^i = \left\{s_o^{i1}, s_o^{i2}, \dots, s_o^{in}\right\}$$
 for s_o^i using the scales in \mathcal{I} .

14: $\tilde{\mathbf{x}}_I^{i+1} = \prod_{\mathbf{x}_I, \epsilon_I} \left(\tilde{\mathbf{x}}_I^i + \alpha \cdot \operatorname{sign}\left(\frac{\nabla_{s_o^i} \sum_{j=1}^n J(s_o^{ij}, \mathbf{x}_T)}{\left\|\nabla_{s_o^i} \sum_{j=1}^n J(s_o^{ij}, \mathbf{x}_T)\right\|}\right)\right)$.

- 16: /* Adversarial Text Generation */
- 17: $\tilde{\mathbf{x}}_T = \underset{\tilde{\mathbf{x}}_T \in B[\mathbf{x}_T, \epsilon_t]}{\arg \max} \left(\kappa \cdot J(\mathbf{x}_I, \tilde{\mathbf{x}}_T) + \mu \cdot J(\tilde{\mathbf{x}}_I^{T-1}, \tilde{\mathbf{x}}_T) + \nu \cdot J(\tilde{\mathbf{x}}_I^T), \tilde{\mathbf{x}}_T \right) \right).$

PROOF OF THEOREM 1

We first consider the following proposition to prove Theorem 1.

Proposition 1 (Update Rules). The adversarial perturbation generated by the proposed method at t-step $(t \ge 2)$ is

$$g_t = a_t \cdot g + b_t \cdot Hg,$$

$$\delta_t = \sum_{i=1}^t g_i = c_t \cdot g + d_t \cdot Hg,$$
(21)

where q is the gradient and H is the Hessian matrix of the loss function L concerning x, and

$$a_{t} = 1,$$

$$b_{t} = \beta \cdot (t-2) + \gamma \cdot (t-1),$$

$$c_{t} = t,$$

$$d_{t} = \sum_{i=2}^{t} \beta(i-2) + \gamma(i-1) = \frac{(t-1)(t-2)}{2}\beta + \frac{t(t-1)}{2}\gamma.$$
(22)

Then the update rule of SAG will be

$$g'_{t} = e_{t} \cdot g + f_{t} \cdot Hg,$$

$$\zeta_{t} = \sum_{i=1}^{t} g'_{i} = h_{t} \cdot g + l_{t} \cdot Hg,$$
(23)

where

$$e_{t} = 1,$$
 $f_{t} = t - 1,$
 $h_{t} = t,$
 $l_{t} = \frac{1}{2} \cdot t(t - 1).$
(24)

Proof. We use mathematical induction to complete the proof.

1) When t = 2, by (8) and (9), we have

$$a_{2} = 1,$$
 $b_{2} = \gamma,$
 $c_{2} = 2,$
 $d_{2} = \gamma.$
(25)

2) Assuming that the update rule holds when $m = t \ge 2$, and when m = t + 1 we have

$$g_{t+1} = g(\mathbf{x} + \beta \cdot \boldsymbol{\delta}_{t-1} + \gamma \cdot \boldsymbol{\delta}_{t})$$

$$= \sum_{n=0}^{\infty} \frac{g^{(n)}(\mathbf{x})}{n!} \left(\mathbf{x} + \beta \cdot \sum_{i=1}^{t-1} \mathbf{g}_{i} + \gamma \cdot \sum_{i=1}^{t} \mathbf{g}_{i} - \mathbf{x} \right)^{n}$$

$$\approx \mathbf{g} + \beta \cdot \sum_{i=1}^{t-1} \mathbf{H} \mathbf{g}_{i} + \gamma \cdot \sum_{i=1}^{t} \mathbf{H} \mathbf{g}_{i}$$

$$= \mathbf{g} + (\beta + \gamma) \sum_{i=1}^{t-1} \mathbf{H} \mathbf{g}_{i} + \gamma \mathbf{H} \mathbf{g}_{t}$$

$$= \mathbf{g} + (\beta + \gamma) \sum_{i=1}^{t-1} \mathbf{H} (a_{i} \cdot \mathbf{g} + b_{i} \cdot \mathbf{H} \mathbf{g}) + \gamma \mathbf{H} (a_{t} \cdot \mathbf{g} + b_{t} \cdot \mathbf{H} \mathbf{g})$$

$$\approx \mathbf{g} + (\beta + \gamma) \sum_{i=1}^{t-1} a_{i} \cdot \mathbf{H} \mathbf{g} + \gamma \cdot a_{t} \mathbf{H} \mathbf{g}$$

$$= \mathbf{g} + \left((\beta + \gamma) \sum_{i=1}^{t-1} a_{i} \cdot \mathbf{H} \mathbf{g} + \gamma \cdot a_{t} \mathbf{H} \mathbf{g} \right)$$

$$= \mathbf{g} + \left((\beta + \gamma) \sum_{i=1}^{t-1} a_{i} + \gamma \cdot a_{t} \right) \mathbf{H} \mathbf{g},$$

$$(26)$$

where the first approximation holds as we ignore the high-order terms of $g^{(n)}$ $(n \ge 2)$ and the second approximation holds as we ignore the term related to HH. Then we have

$$a_{t+1} = 1,$$

 $b_{t+1} = (\beta + \gamma) \sum_{i=1}^{t-1} a_i + \gamma \cdot a_t = \beta \cdot (t-1) + \gamma \cdot t.$ (27)

Furthermore,

$$\delta_{t+1} = \delta_t + g_{t+1}$$

$$= c_t \cdot g + d_t \cdot Hg + a_{t+1} \cdot g + b_{t+1} \cdot Hg$$

$$= (c_t + a_{t+1}) \cdot g + (d_t + b_{t+1}) \cdot Hg$$

$$= c_{t+1} \cdot g + d_{t+1} \cdot Hg$$
(28)

Then we have

$$c_{t+1} - c_t = a_{t+1} = 1,$$

$$c_t - c_{t-1} = a_t = 1,$$

$$\cdots$$

$$c_3 - c_2 = a_3 = 1,$$

$$c_{t+1} = t + 1.$$
(29)

Meanwhile, it holds

$$d_{t+1} - d_t = b_{t+1},$$

$$d_t - d_{t-1} = b_t,$$

$$...$$

$$d_3 - d_2 = b_3,$$
(30)

and

$$d_{t+1} = \sum_{i=2}^{t+1} \beta(i-2) + \gamma(i-1)$$

$$= \frac{(t-1)(t-2)}{2} \beta + \frac{t(t-1)}{2} \gamma.$$
(31)

Theorem 1. The adversarial perturbations $\{\delta_t\}$ generated by the proposed method are given as

$$\boldsymbol{\delta}_t = \sum_{i=1}^t \boldsymbol{g}_i, \quad t = 2, 3, \dots,$$
 (8)

where

$$\mathbf{g}_{t} = \begin{cases} g(\mathbf{x} + \beta \cdot \boldsymbol{\delta}_{t-2} + \gamma \cdot \boldsymbol{\delta}_{t-1}), & \text{if } t \geq 2, \\ g(\mathbf{x}), & \text{if } t = 1, \end{cases}$$
(9)

 $g(\cdot)$ is the gradient of loss function $L(\cdot)$, β , $\gamma \in [0,1]$ are given constants, and $\delta_0 = (0,\ldots,0)$. Meanwhile, the adversarial perturbations generated by the SGA [42] $\{\zeta_t\}^2$ are

$$\zeta_t = \sum_{i=1}^t \mathbf{h}_i, \ \mathbf{h}_i = g(\mathbf{x} + \zeta_{i-1}), \ t = 1, 2, \dots$$
 (10)

Then the interaction inside adversarial perturbation δ_t will be

$$\mathbb{E}_{i,j}[\boldsymbol{I}_{i,j}(\boldsymbol{\delta}_t)] = (\beta + \gamma)B \cdot t^3 + (A - 2\beta B) \cdot t^2 + (2A - (\beta + \gamma)B) \cdot t + A + 2\beta B,$$
(11)

where

$$A = \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j}],$$

$$B = \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j}],$$
(12)

 $g = [g(1), \dots, g(n)]$ is the gradient of L(x), $H_{i,j}$ is the (i,j) element of the Hessian matrix H for L(x), and H_{*j} is the j-th column of H. Moreover, the interaction inside ζ_t

$$\mathbb{E}_{i,j}[I_{i,j}(\zeta_t)] = B \cdot t^3 + (A - B) \cdot t^2. \tag{13}$$

Proof. By [58], the Shapley interactions between adversarial perturbation i and j is defined as

$$I_{i,j}(\boldsymbol{\delta}_t) = \boldsymbol{\delta}_t(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{\delta}_t(j) + \mathcal{R}(\boldsymbol{\delta}_t(i), \boldsymbol{\delta}_t(j)), \tag{32}$$

where $\delta_t(i)$ represents the *i*-th element of δ_t , $\mathcal{R}(\cdot, \cdot)$ is the high order terms with respect to $\delta_t(i)$ and $\delta_t(j)$, and $\boldsymbol{H}_{i,j}$ is the element of the Hessian matrix in of *i*-th row and *j*-th column as

$$\boldsymbol{H}_{i,j} = \frac{\partial \boldsymbol{L}(\boldsymbol{x})}{\partial x_i \partial x_j}, \ \boldsymbol{x} = (x_1, \dots, x_n)^{\top}.$$
 (33)

Then we have

$$I_{i,j}(\boldsymbol{\delta}_{t}) \approx (c_{t} \cdot \boldsymbol{g}(i) + d_{t} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*i}) \boldsymbol{H}_{i,j} (c_{t} \cdot \boldsymbol{g}(j) + d_{t} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j})$$

$$= c_{t}^{2} \cdot \boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j} + c_{t} \cdot d_{t} \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*i} + c_{t} \cdot d_{t} \cdot \boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j} + \mathcal{O}(\boldsymbol{H}^{2})$$

$$\approx c_{t}^{2} \cdot \boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j} + c_{t} \cdot d_{t} \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*i} + c_{t} \cdot d_{t} \cdot \boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j}.$$

$$(34)$$

Furthermore

$$\mathbb{E}_{i,j}[\boldsymbol{I}_{i,j}(\boldsymbol{\delta}_{t})] = \mathbb{E}_{i,j}[c_{t}^{2} \cdot \boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j} + c_{t} \cdot d_{t} \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*i} + c_{t} \cdot d_{t} \cdot \boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j}]
= c_{t}^{2} \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j}] + 2c_{t} \cdot d_{t} \cdot \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j}]
= t^{2} A + [t(t-1)(t-2)\beta + t^{2}(t-1)\gamma]B
= (\beta + \gamma)B \cdot t^{3} + (A - 3\beta B - \gamma B) \cdot t^{2} + 2\beta B \cdot t$$
(35)

where

$$A = \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j}],$$

$$B = \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j}].$$
(36)

²SGA [42] can be treated as a special case of the proposed method as $\beta = 0$ and $\gamma = 1$.

Meanwhile, considering the updated rule of SAG, we have

$$\mathbb{E}_{i,j}[\boldsymbol{I}_{i,j}^{(t)}(\boldsymbol{\zeta}_t)] = t^2 \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{g}(j) \cdot \boldsymbol{H}_{i,j}] + t^2(t-1) \cdot \mathbb{E}_{i,j}[\boldsymbol{g}(i) \cdot \boldsymbol{H}_{i,j} \cdot \boldsymbol{g}^{\top} \boldsymbol{H}_{*j}]$$

$$= t^2 \cdot A + t^2 \cdot (t-1) \cdot B$$
(37)