One Perturbation is Enough: On Generating Universal Adversarial Perturbations against Vision-Language Pre-training Models

Hao Fang[1,*] Jiawei Kong[1,2,*] Wenbo Yu[1] Bin Chen[2,3†] Jiawei Li[4]
Hao Wu[5] Shu-Tao Xia[1,3] Ke Xu[1]
[1]Tsinghua University [2] Harbin Institute of Technology, Shenzhen [3]Peng Cheng Laboratory
[4]Huawei Technology [5]Shenzhen ShenNong Information Technology Co., Ltd.
ffhibnese@gmail.com, kongjiawei@stu.hit.edu.cn, wenbo.research@gmail.com, chenbin2021@hit.edu.cn,
li-jw15@tsinghua.org.cn, whpc79@163.com, xiast@sz.tsinghua.edu.cn, xuke@tsinghua.edu.cn;

Abstract

Vision-Language Pre-training (VLP) models have exhibited unprecedented capability in many applications by taking full advantage of the learned multimodal alignment. However, previous studies have shown they are vulnerable to maliciously crafted adversarial samples. Despite recent success, these attacks are generally instance-specific and require generating perturbations for each input sample. In this paper, we reveal that VLP models are also susceptible to the instance-agnostic universal adversarial perturbation (UAP). Specifically, we design a novel Contrastive-training Perturbation Generator with Cross-modal conditions (C-PGC). In light that the pivotal multimodal alignment in VLP models is achieved via contrastive learning, we devise to turn this powerful weapon against VLP models themselves. I.e., we employ a malicious version of contrastive learning to train the proposed generator using our carefully crafted positive and negative image-text pairs. Once training is complete, the generator is able to produce universal perturbations that can essentially destroy the established alignment relationship in VLP models. Besides, C-PGC fully utilizes the characteristics of Vision-and-Language (V+L) scenarios by incorporating both unimodal and cross-modal information as effective guidance. Extensive experiments show that C-PGC successfully forces adversarial samples to move away from their original area in the VLP model's feature space, thus fundamentally enhancing attack performance across various victim models and V+L tasks.

1. Introduction

Vision-Language Pre-training (VLP) models have recently demonstrated remarkable efficacy in a wide range of Vision-and-Language (V+L) tasks. By self-supervised pre-training on large-scale image-text pairs, VLP models efficiently align cross-modal features and capture rich information from the aligned multimodal embeddings, thereby providing expressive representations for various applications.

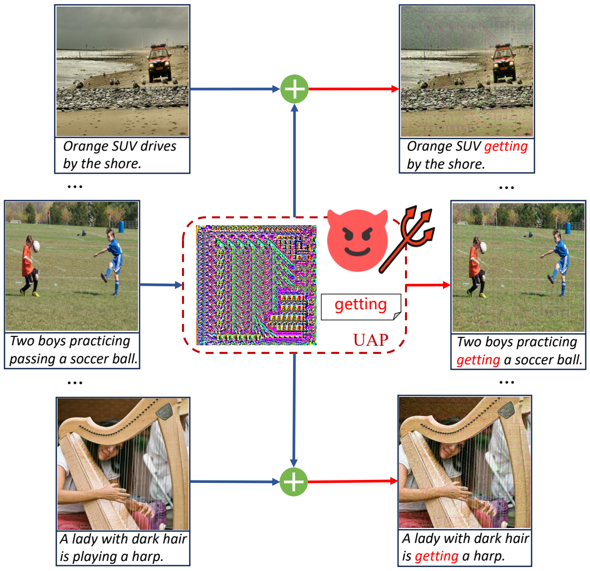
Orange SUV drives by the shore.
Orange SUV getting by the shore.
...
...
Two boys practicing passing a soccer ball.
getting
UAP
Two boys practicing getting a soccer ball.
...
...
A lady with dark hair is playing a harp.
A lady with dark hair is getting a harp.

Fig. 1: Illustration of universal adversarial attacks. With only a pair of image-text perturbations, the proposed method can effectively mislead different VLP models on diverse V+L tasks.

Adversarial attacks [7, 13], which aim to deceive models during inference time, have attracted extensive attention due to their significant threat to security-critical scenarios [11]. Recent studies have shown that VLP models are also vulnerable to adversarial samples. The pioneering work Co-Attack [48] proposes the first multimodal attack that simultaneously perturbs both image and text modalities and displays excellent performance. However, Co-Attack only considers relatively easier white-box attacks where victim

[*] Equal Contribution.
[†] Corresponding Author.

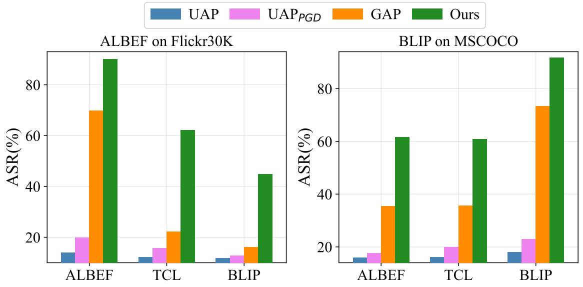

Fig. 2: Performance of existing UAP on text retrieval with ALBEF [23] and BLIP [24] as surrogate models. Note that UAP [31] is initially based on DeepFool [30], and the corresponding PGD-learned version UAP$_{PGD}$ is provided for a fair comparison.

models are completely accessible. To handle more practical black-box settings, subsequent studies propose various transferable adversarial samples generated on an available surrogate model to fool other inaccessible models. Specifically, SGA [29] significantly improves the adversarial transferability through the set-level cross-modal guidance obtained from data augmentations. Subsequently, TMM [40] proposes to jointly destroy the modality-consistency features within the clean image-text pairs and include more modality-discrepancy features in the perturbations to further enhance transferability. While existing methods have achieved great success, they are all instance-specific and need to generate a perturbation for each input pair, which results in substantial computational overhead. Meanwhile, universal adversarial attacks (illustrated in Fig. 1), as an efficient instance-agnostic approach that uses only one Universal Adversarial Perturbation (UAP) to conduct attacks, have not been fully investigated for VLP models. This naturally leads to a question, *is it possible to design a UAP that can deceive VLP models across different image-text pairs?*

Motivation. To this end, we make an intuitive attempt to transplant renowned approaches UAP [31] and GAP [35] to attack several VLP models by maximizing the distance between the embeddings of the adversarial image and its matched texts. Unfortunately, Fig. 2 demonstrates that these methods yield unsatisfactory attack success rates (ASR), especially for black-box attacks. Empirically, this failure stems from their narrow focus on the image modal, disregarding the other modality and the multimodal information that plays a pivotal role in VLP models. To overcome this challenge, we revisit VLP models' basic training paradigm and emphasize that regardless of the downstream V+L tasks, their achieved outstanding performance is heavily reliant on the well-established multimodal alignment, which draws the embeddings of matched image-text pairs closer while distancing those of non-matched pairs. In light of this consideration, we argue that the core of an effective universal adversarial attack is to obtain a UAP that can fundamentally destroy this learned alignment relationship to mislead VLP models into making incorrect decisions. Besides, Fig. 2 shows that the generator-based GAP consistently outperforms UAP methods, due to the excellent distributional modeling capability of generators. This suggests the superiority of the generative paradigm, which is also corroborated by numerous studies [14, 15].

Based on these insights, we propose a novel generative framework that learns a Contrastive-training Perturbation Generator with Cross-modal conditions (C-PGC) to launch universal attacks on VLP models. To essentially destroy the multimodal alignment, we devise to utilize VLP models' most powerful weapons to attack against themselves. *I.e.*, utilize the contrastive learning mechanism to train the generator based on maliciously constructed image-text pairs that completely violate the correct V+L matching relationship, to produce a perturbation that pushes the embeddings of matched pairs apart while pulling those of non-matched ones together. Moreover, most previous studies [29, 48, 49] simply exploit the crucial cross-modal information by maximizing the feature distance between samples of different modals to optimize perturbations, without deeper exploration for attack enhancement. In contrast, we fully harness V+L characteristics by refining the generator architecture to incorporate cross-modal knowledge through cross-attention mechanisms for better guidance. Besides, we also consider the intra-modal influence and introduce a unimodal distance loss to further improve attack effectiveness. We highlight that the proposed framework is seamlessly compatible with text perturbation generation, achieving a truly multimodal universal attack that benefits from the synergy between V+L modalities. Our contributions are as follows:

- We propose C-PGC, a novel perturbation generator conditioned on cross-modal knowledge, to produce both image and text UAPs for powerful attacks on VLP models.
- We design a malicious contrastive learning paradigm that incorporates both unimodal and multimodal guidance to train the generator to produce UAP that can essentially disrupt the multimodal alignment in VLP models.
- Extensive experiments on 6 various VLP models across diverse V+L tasks reveal that our method achieves outstanding attack performance in different scenarios.

2. Related Work

2.1. Vision-Language Pre-training Models

VLP models are pre-trained on massive image-text pairs to learn the semantic correlations across modalities and serve diverse multimodal user demands [8, 10]. We next illustrate the basis of VLP models from multiple perspectives.

Architectures. Based on the ways of multimodal fusion, the architectures of VLP models can be classified into two types: *single-stream* and *dual-stream* architectures. Single-stream architectures [9, 25] directly concatenate the text and image features, and calculate the attention in the same Transformer block for multimodal fusion. On the contrary, dual-stream architectures [24, 37] separately feed the text and image features to different Transformer blocks and leverage the cross-attention mechanism for multimodal fusion. Generally, single-stream architectures are more parameter-efficient than dual-stream architectures since they adopt the same set of parameters in a Transformer block for the text and image modalities.

Pre-training Objectives. The pre-training objectives for VLP models mainly include *masked feature completion*, *multimodal feature matching*, and *specific downstream objectives*. Masked feature completion [9] encourages VLP models to predict the deliberately masked tokens based on the remaining unmasked tokens. Multimodal feature matching [23] pre-trains VLP models to precisely predict whether the given image-text pairs are matched. Specific downstream objectives [2] directly utilize the training objectives of downstream tasks for pre-training VLP models.

Downstream Tasks. In this paper, we mainly consider the following multimodal downstream tasks: (1) Image-text retrieval (ITR) [41]: finding the most matched image for the given text and vice versa, including image-to-text retrieval (TR) and text-to-image retrieval (IR). (2) Image caption (IC) [4]: generating the most suitable descriptions for the given image. (3) Visual grounding (VG) [20]: locating specific regions in the image that correspond with the given textual descriptions. (4) Visual entailment (VE) [42]: analyzing the input image and text and predicting whether their relationship is entailment, neutral, or contradiction.

2.2. Adversarial Attacks

Among various attacks [12, 16, 17, 36, 46, 51], adversarial attacks stand out as a particularly powerful type.

Instance-specific Attacks on VLP Models. The adversarial robustness of VLP Models has become a research focus. Early works [22, 44] impose perturbations only on a single modal and lack cross-modal interactions when attacking multimodal models. To solve this, Co-Attack [48] conducts the first white-box attack on VLP models. Based on Co-Attack, [29] extends the attacks to more rigorous black-box settings and proposes SGA, which utilizes set-level alignment-preserving argumentations with carefully designed cross-modal guidance. However, [40] points out that SGA fails to fully exploit modality correlation and proposes TMM to better leverage cross-modal interactions via modality-consistency and modality-discrepancy features. Nonetheless, these methods are all instance-specific and need to craft perturbations for each input pair.

Universal Adversarial Examples. Universal adversarial attacks [31, 32, 50] aim to deceive the victim model by exerting a uniform adversarial perturbation to all samples. These attacks save the redundant procedures of redesigning perturbations for each input sample and are hence more efficient than instance-specific methods. Generally, universal adversarial attacks can be categorized into optimization-based [28, 31, 39] and generation-based [3, 15, 18, 50] methods. Benefiting from the powerful modeling abilities of generative models, generation-based methods are more versatile and can produce more natural samples than optimization-based ones. A concurrent work, ETU [49], also investigates UAP on VLP models and proposes a data augmentation named ScMix. However, *ETU adopts a non-generative approach that narrowly focuses on image UAP, failing to constitute a truly multimodal attack for VLP models.* Moreover, ETU demonstrates insufficient attack effects, especially for black-box transferability. In contrast, we propose a generative multimodal attack framework based on a malicious variant of contrastive learning, which yields UAP with strong attack effects and high transferability.

3. Universal Multimodal Attacks

In this section, we first present the problem statement of universal adversarial attacks on VLP models. Next, we introduce the overview of our framework. Finally, we illustrate the detailed design of the proposed C-PGC.

3.1. Problem Statement

We define an input image-text pair as (v, t) and denote $\boldsymbol{e}_v$ and $\boldsymbol{e}_t$ as the image and text embeddings encoded by the image encoder $f_I(\cdot)$ and text encoder $f_T(\cdot)$ of the targeted VLP model $f(\cdot)$. Let $\mathcal{D}_s$ be an available dataset consisting of image-text pairs collected by a malicious adversary. The attack objective is to utilize $\mathcal{D}_s$ to train a generator $G_w(\cdot)$ for producing a powerful pair of universal image-text perturbations (δ_v, δ_t) that can affect the vast majority of test dataset $\mathcal{D}_t$ to fool models into making incorrect decisions. Formally, the attack goal can be formulated as:

$$\mathcal{T}(f(v + \delta_v, t \oplus \delta_t)) \neq y, \text{ s.t. } \|\delta_v\|_\infty \leq \epsilon_v, \|\delta_t\|_0 \leq \epsilon_t, \tag{1}$$

where $\mathcal{T}(\cdot)$ denotes the operation that uses the output V+L features to obtain the final predictions, $\oplus$ indicates the text perturbation strategy [29, 48] that replaces certain important tokens of the original sentence with crafted adversarial words, and y is the correct prediction of the considered V+L task. To ensure the perturbation's imperceptibility, we constrain the pixel-level image perturbation with l_∞ norm of a given budget ϵ_v. Following previous studies [29, 40, 48], the textual perturbation is token-level and the stealthiness is accordingly constrained by the number of modified words ϵ_t. To ensure the stealthiness of text perturbation, we apply a rigorous restriction that permits only a single word to be substituted ($\epsilon_t = 1$). On the premise of imperceptibility, the attacker attempts to generalize the crafted UAP to a wider range of test data and victim models.

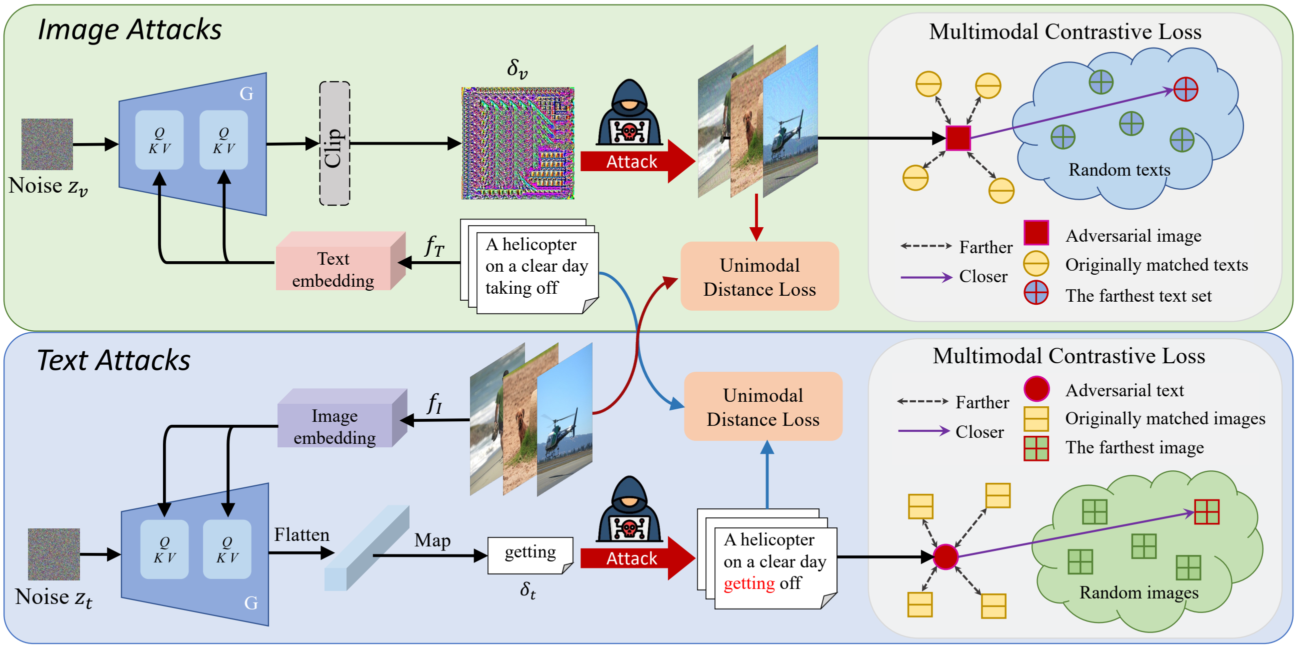

Fig. 3: An overview of our proposed universal adversarial attack. Benefiting from the well-designed unimodal distance loss $\mathcal{L}_{Dis}$ and multimodal contrastive loss $\mathcal{L}_{CL}$, the generator $G_w(\cdot)$, conditioned with cross-modal embeddings, learns rich knowledge from features of different modalities and thus produces δ_v and δ_t of superior generalization ability across diverse models and downstream tasks.

3.2. Overview of the Proposed Framework

The overview of C-PGC is depicted in Fig. 3. We adopt a multimodal perturbation strategy and generate perturbations for both image and text modalities. Given the high similarity between the workflows of image and text, we then take image attacks as an example for illustration.

Firstly, a fixed noise z_v is randomly initialized and subsequently fed into the conditional generator. For each image v and its descriptions $\mathbf{t}$, the generator $G_w(\cdot)$ translates the input noise z_v into the adversarial perturbation δ_v that is of the same size as v. During generation, the network G_w additionally benefits from cross-modal information by integrating the embedding of text descriptions corresponding to the current input image v, *i.e.*, $\delta_v = G_w(z_v; f_T(\mathbf{t}))$. Next, the generated adversarial noise δ_v is injected into the clean image to obtain the adversarial image via $v_{adv} = v + \delta_v$. To better guide the training process, we design two effective unimodal and multimodal losses as our optimization objectives. Unimodal loss is straightforward and aims to push the adversarial images away from the clean images in the latent embedding space, while multimodal loss is based on contrastive learning using our manually constructed positive and negative samples to strongly destroy the image-text matching relationship achieved by feature alignment. Once we finish training C-PGC using the proposed loss function, the input fixed noise is transformed into a UAP with great generalization and transferability.

3.3. Detailed Design of C-PGC

Next, we provide a detailed introduction to each of the proposed designs. Note that we primarily discuss the image attack as an example, given that the design of the text attack is completely symmetrical. The pseudocode of the training procedure is provided in Appendix A.

Perturbation Generator Conditioned on Cross-modal knowledge. Previous generative universal attacks [3, 15] have shown excellent efficacy in fooling the discriminative models. Nevertheless, since existing generative attacks are limited to a single modality, directly utilizing the off-the-shelf generators might fail to leverage the multimodal interactions in these special V+L scenarios. To address this limitation, we additionally introduce cross-modal embeddings as auxiliary information to further facilitate the process of perturbation generation. Specifically, we modify the architecture of existing decoder-based generators by adding several cross-attention modules that have been proven effective in tasks with multiple input modalities. The obtained textual embeddings $\boldsymbol{e}_t$ encoded by $f_T(\cdot)$ are then incorporated into our generator through:

$$Q = \boldsymbol{h}_t W_q, K = \boldsymbol{e}_t W_k, V = \boldsymbol{e}_t W_v,$$
$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \tag{2}$$

where $\boldsymbol{h}_t \in \mathbb{R}^{B \times d_\alpha}$ is the flattened intermediate features within $G_w(\cdot)$, and $W_q \in \mathbb{R}^{d_\alpha \times d}$, $W_k \in \mathbb{R}^{512 \times d}$, $W_v \in$

$\mathbb{R}^{512\times d}$ are the optimized parameters of attention modules.

Multimodal Contrastive Loss. The preceding analysis regarding the failures of existing UAP attacks encourages us to design a loss function that can guide the generated UAP to break the learned multimodal feature alignment. Motivated by the fact that contrastive learning underpins the cross-modal alignment, we advocate leveraging this mechanism to attack VLP models themselves by contrastively training our C-PGC to essentially disrupt the benign alignment relationship. Concretely, we adopt the widely recognized InfoNCE [19] as our basic contrastive loss.

To establish the contrastive paradigm, we first define the adversarial image v_{adv} as the anchor sample. Besides, it is also necessary to construct an appropriate set of positive and negative samples. Based on the fundamental objective of our attack, it is natural to leverage the originally matched text description set $\mathbf{t} = \{t_1, t_2, \ldots, t_M\}$ as negative samples $\mathbf{t}_{neg}$ to amplify the discrepancy between matched image-text pairs in the feature space of VLP models. Additionally, we need to select a set of positive samples to further pull the adversarial image v_{adv} away from its corresponding text descriptions $\mathbf{t}$. To this end, we propose a *farthest selection strategy*, which associates the anchor image v_{adv} with target texts $\mathbf{t}_{pos}$ whose embeddings differ significantly from that of the original clean image v, to reach a more strong disruption of the multimodal alignment. Specifically, we randomly sample a batch of text sets from $\mathcal{D}_s$ and select the text set with the largest feature distances from the current image v as positive samples, *i.e.*, $\mathbf{t}_{pos} = \{t'_1, t'_2 \ldots, t'_K\}$. Moreover, we utilize data augmentations that resize the clean v into diverse scales and apply random Gaussian noise to acquire a more diverse image set $\mathbf{v} = \{v_1, v_2 \ldots, v_N\}$ for set-level guidance [29]. With these well-constructed positive and negative samples, the multimodal contrastive loss $\mathcal{L}_{CL}$ can be formulated as:

$$\mathcal{L}_{CL} = \log \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} s(v_i+\delta_v, t_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} s(v_i+\delta_v, t_j) + \sum_{i=1}^{N}\sum_{j=1}^{K} s(v_i+\delta_v, t'_j)}, \tag{3}$$

where δ_v is the universal image perturbation; $s(v,t) = \exp(\text{sim}(f_I(v), f_T(t))/\tau)$, where τ denotes the temperature parameter and $\text{sim}(\cdot,\cdot)$ represents the cosine similarity.

Unimodal Distance Loss. Apart from the multimodal guidance, we also consider the unimodal influence by directly pushing adversarial images away from their initial visual semantic area to improve the attack. Similarly, the input image v also undergoes resizing and noise perturbation to generate the augmented image set $\mathbf{v} = \{v_1, v_2 \ldots, v_N\}$ for set-level guidance. Then, we craft the adversarial image through $v_{adv} = v + \delta_v$ and process v_{adv} with the same augmentation operation to obtain the adversarial image set $\mathbf{v}_{adv} = \{v_1^{adv}, v_2^{adv} \ldots, v_N^{adv}\}$. Finally, we minimize the negative Euclidean distance between the embeddings of adversarial images and clean images to optimize the UAP generator. Formally, the loss $\mathcal{L}_{Dis}$ is formulated as:

$$\mathcal{L}_{Dis} = -\sum_{i=1}^{N}\sum_{j=1}^{N}\|f_I(v_i^{adv}) - f_I(v_j)\|_2. \tag{4}$$

Taking advantage of the unimodal guidance, $\mathcal{L}_{Dis}$ ensures an effective optimization direction during the generator training and further enhances the attack effectiveness.

Training Objective. With the above two well-designed loss terms $\mathcal{L}_{Dis}$ and $\mathcal{L}_{CL}$, the overall optimization objective of our conditional generator for image attacks can be formulated as:

$$\begin{aligned} &\min_{w} \mathbb{E}_{(v,\mathbf{t})\sim\mathcal{D}_s, \mathbf{t}_{pos}\sim\mathcal{D}_s}(\mathcal{L}_{CL} + \lambda\mathcal{L}_{Dis}), \\ &\qquad \text{s.t. } \|G_w(z_v; f_T(\mathbf{t}))\|_\infty \le \epsilon_v, \end{aligned} \tag{5}$$

where λ is the pre-defined hyperparameter to balance the contributions of $\mathcal{L}_{CL}$ and $\mathcal{L}_{Dis}$. By training the network with the proposed loss function over the entire data distribution of the multimodal training dataset D_s, the generator $G_w(\cdot)$ is optimized to generate UAPs that push the features of mismatched image-text pairs together while pulling the embeddings of the matched ones apart. This finally enables the generation of UAPs with strong generalization capabilities and high adversarial transferability.

Text Modality Attacks. In textual attacks, the generator architecture and training loss are completely symmetrical with those of image attacks. Correspondingly, embeddings of the matched image v are used as the cross-modal conditions for the generator. Given an adversarial text t_{adv} as the anchor sample, we use the set $\mathbf{v} = \{v_1, v_2 \ldots, v_N\}$ scaled from the originally matched image v as negative samples while the $\mathbf{v}' = \{v'_1, v'_2 \ldots, v'_N\}$ augmented from the farthest image v' within the randomly sampled image set as positive samples to formulate the $\mathcal{L}_{CL}$ loss. $\mathcal{L}_{Dis}$ is consequently calculated as the negative Euclidean distance between the embeddings of t_{adv} and the clean input t. Accordingly, the conditional generator is utilized to output the adversarial textual embeddings, which are subsequently mapped back to the vocabulary space to obtain a universally applicable word-level perturbation.

A notable distinction between image and text attacks is the way to inject adversarial perturbations. We align with previous studies [29, 40, 48] and apply the token-wise substitute strategy that replaces certain important words in the original sentence with crafted adversarial words. Prior to the word replacement, a meticulous process is undertaken to identify the most optimal position within the sentence to insert the perturbation. Our strategy intends to replace the words that are more likely to have a greater influence during decision-making. Concretely, for each word w_i within

Table 1. ASR (%) of different methods for image-text retrieval tasks on Flickr30k dataset. TR indicates text retrieval based on the input image, while IR is image retrieval using the input text. The results on the MSCOCO dataset are in Appendix C due to space limits.

Source	Method	ALBEF		TCL		X-VLM		CLIP$_{\text{ViT}}$		CLIP$_{\text{CNN}}$		BLIP	
		TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
ALBEF	GAP	69.78	81.59	22.15	29.97	6.61	18.37	23.40	37.54	29.92	44.29	16.09	28.12
	ETU	78.01	84.56	29.92	35.91	14.33	22.03	23.77	39.20	33.55	47.69	22.61	32.28
	Ours	**90.13**	**88.82**	**62.11**	**64.48**	**20.53**	**39.38**	**43.10**	**65.93**	**54.40**	**72.51**	**44.79**	**56.36**
TCL	GAP	33.50	40.61	82.41	80.67	6.61	17.79	21.55	38.56	30.57	45.48	21.45	31.82
	ETU	28.26	35.03	90.48	87.57	9.65	20.56	25.00	39.68	36.14	49.33	18.93	29.19
	Ours	**50.26**	**56.29**	**94.93**	**90.64**	**14.94**	**33.96**	**46.92**	**66.41**	**52.98**	**70.66**	**35.75**	**52.52**
X-VLM	GAP	16.14	24.43	17.08	26.20	90.24	85.98	24.51	41.15	42.62	53.08	16.19	25.74
	ETU	12.33	21.93	13.98	24.04	93.19	90.85	23.89	39.62	35.62	51.19	12.09	23.59
	Ours	**24.46**	**47.77**	**29.19**	**50.15**	**93.29**	**91.90**	**43.47**	**66.03**	**59.20**	**72.79**	**32.39**	**52.24**
CLIP$_{\text{ViT}}$	GAP	11.72	23.34	15.32	26.39	8.54	20.48	85.73	90.45	48.83	60.78	14.83	26.46
	ETU	14.80	25.23	21.22	30.87	10.87	24.96	84.14	90.45	57.51	65.51	16.40	27.22
	Ours	**23.23**	**38.67**	**25.05**	**41.79**	**15.85**	**35.59**	**88.92**	**93.05**	**66.06**	**75.42**	**26.71**	**45.70**
CLIP$_{\text{CNN}}$	GAP	13.57	25.21	19.05	28.87	11.59	23.13	27.46	43.16	73.18	81.60	15.25	27.94
	ETU	8.94	20.59	13.25	24.41	8.94	20.82	21.92	40.51	91.71	92.40	11.15	23.82
	Ours	**19.01**	**41.86**	**22.98**	**47.02**	**19.61**	**43.26**	**40.89**	**65.77**	**96.50**	**94.22**	**24.19**	**48.17**
BLIP	GAP	12.23	23.94	14.49	25.44	6.91	17.81	20.32	37.00	26.81	43.59	47.21	73.33
	ETU	19.32	27.91	19.98	29.15	11.99	20.91	24.38	39.84	31.61	46.22	59.52	77.82
	Ours	**32.17**	**44.40**	**33.44**	**44.51**	**18.60**	**35.53**	**43.35**	**60.26**	**48.96**	**66.95**	**71.82**	**82.82**

a given sentence, we compute the distance between the embeddings of the original sentence and the w_i-masked version to determine its contribution. By convention, the imperceptibility of text UAP is controlled by the number of modified words ϵ_t [6]. As aforementioned, we set $\epsilon_t = 1$ for high stealthiness, *i.e.*, choose the single word exerting the highest feature distance as the target for replacement.

4. Experiments

We conduct comprehensive experiments on diverse V+L tasks across multiple VLP models. Please see more experimental results in the Appendix. Our code is available at: https://github.com/ffhibnese/CPGC_VLP_Universal_Attacks.

4.1. Experimental Setup

Downstream tasks and datasets. We evaluate C-PGC on four downstream V+L tasks, including image-text retrieval (ITR), image captioning (IC), visual grounding (VG), and visual entailment (VE). Following [29, 48], we employ the Flickr30K [34] and MSCOCO [27] datasets for ITR tasks, . The MSCOCO is also adopted for evaluating the IC task. For VG and VE tasks, we evaluate on SNLI-VE [42] and RefCOCO+ [45] datasets, respectively.

Models. We conduct experiments on a wide range of VLP models, including ALBEF [23], TCL [43], X-VLM [47], CLIP$_{\text{ViT}}$ [37], CLIP$_{\text{CNN}}$ [37], and BLIP [24]. Note that for different V+L tasks, we correspondingly select different VLP models for evaluation based on their capability [40]. For instance, among the six considered VLP models, only ALBEF, TCL, and X-VLM can handle VG tasks, while only ALBEF and TCL can deal with VE tasks.

Baselines. We transplant the representative GAP [35] to V+L scenarios by appropriately editing its original loss function [29]. We also consider a concurrent UAP study ETU [49], which adopts a non-generative method that narrowly focuses on image perturbation, despite the multimodal nature of V+L scenarios. Note that it implements several versions, and we report their best results.

Implementation details. Following the SGA [29], we adopt the Karpathy split [21] to preprocess the dataset and build the test set for evaluation. The test set is disjoint from the generator's training data for rigorous assessment. To ensure perturbation invisibility, we follow [40] and limit the perturbation budgets ϵ_v to $12/255$ and ϵ_t to 1. For augmentation, we resize the original images into five scales $\{0.5, 0.75, 1, 1.25, 1.5\}$, and apply Gaussian noise $\mathcal{N}(0, 0.5^2)$. See Appendix B for more details.

4.2. Universal Attack Effectiveness

To align with previous studies [29, 48], we first consider the typical V+L task image-text retrieval and calculate the ASR as the proportion of successful adversarial samples within the originally correctly predicted pairs. We present the performance based on R@1 retrieval results in Table 1. Appendix C supplements the results of R@5 and R@10.

Table 2. Performance of C-PGC on the visual grounding task. The first row displays the source models, and the Baseline indicates the clean performance of the target model on clean data.

Target	Baseline			ALBEF			TCL			X-VLM		
	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB
ALBEF	58.4	65.9	46.2	**37.1**	**39.8**	**32.0**	42.2	46.9	35.2	37.6	40.2	33.0
TCL	59.6	66.8	48.1	43.6	47.8	36.9	**39.0**	**41.4**	**33.6**	39.5	41.7	34.1
X-VLM	70.8	67.8	61.8	51.8	54.7	47.7	52.7	55.9	47.8	**33.1**	**34.7**	**28.8**

Table 3. Performance of C-PGC on image captioning task. The Baseline represents the clean performance. The target is BLIP.

Source	B@4	METEOR	ROUBE_L	CIDEr	SPICE
Baseline	39.7	31.0	60.0	133.3	23.8
ALBEF	30.1	23.7	51.2	92.5	17.5
TCL	29.5	23.5	51.0	88.9	17.3
BLIP	**21.2**	**19.1**	**45.5**	**62.5**	**13.7**

White-box attack performance. By observing the white-box ASR in the gray-shaded area, we demonstrate that the proposed algorithm stably achieves excellent ASR on all the evaluated VLP models, validating the outstanding capability of the produced UAP. With only a single pair of perturbations, we reach a noteworthy average white-box ASR of nearly 90% on two large datasets in terms of both TR and IR tasks. Compared with GAP [35] and ETU [49], the proposed method consistently enhances the fooling rates in the white-box scenario, confirming the great validity of our suggested multimodal contrastive-learning mechanism. Essentially, the exceptional performance stems from the efficacy of our generated UAP in destroying the alignment between the image and text modalities, thereby misleading the VLP model during inference.

Black-box attack performance. We also conduct thorough experiments regarding the adversarial transferability of the generated UAP by transferring from surrogate models to other inaccessible models. As demonstrated in Table 1, the proposed C-PGC displays great attack effects in the more realistic black-box scenarios, *e.g.*, 72.51% from ALBEF to CLIP$_{\text{CNN}}$ for IR tasks. We highlight that the advantage of C-PGC over the concurrent study ETU is greatly amplified in the more challenging black-box scenarios, which achieves a significant average improvement of 17.76% on the Flickr30K dataset. These experimental results indicate that our generative contrastive learning framework does not overly rely on the encoded feature space tailored to the surrogate model. Conversely, it is well capable of transferring to breaking the multimodal alignment of other unseen target models, thus attaining superior adversarial transferability.

4.3. Evaluation on More Downstream Tasks

We then provide results on more downstream V+L tasks. Specifically, we consider Image Captioning (IC), Visual Grounding (VG), and Visual Entailment (VE). The results of VE are shown in Appendix C due to space limit.

Visual grounding. This is another common V+L task, which aims to locate the correct position in an image based on a given textual description. We conduct experiments on RefCOCO+ using ALBEF, TCL, and X-VLM as source and target models. Table 2 indicates that C-PGC brings a notable negative impact on the localization accuracy in both white-box and black-box settings, again verifying that the produced UAP strongly breaks the cross-modal alignment.

Image captioning. The objective of IC is to generate text descriptions relevant to the semantic content based on the given image. We use ALBEF, TCL, and BLIP as source models and attack the commonly used captioning model BLIP. Similar to SGA [29], several typical evaluation metrics of IC are calculated to measure the quality of generated captions, including BLEU [33], METEOR [5], ROUGE [26], CIDEr [38], and SPICE [1]. The results in Table 3 demonstrate that our algorithm again displays prominent attack effectiveness, *e.g.*, the crated UAP induces notable drops of 10.2% and 9% in the B@4 and ROUGE_L respectively when transferred from TCL to BLIP.

4.4. Ablation Study

This part employs ALBEF [23] as the surrogate model and provides ablation studies on Flickr30K. We begin our analysis on the contribution of each proposed technique. Next, we examine the sensitivity of certain hyperparameters.

The effect of $\mathcal{L}_{CL}$ and $\mathcal{L}_{Dis}$. To investigate the impact of the proposed loss terms, we introduce two variants C-PGC$_{CL}$ and C-PGC$_{Dis}$ that remove $\mathcal{L}_{CL}$ and $\mathcal{L}_{Dis}$ from the overall training loss respectively. As shown in Table 4, the removal of $\mathcal{L}_{CL}$ leads to significant degradation, particularly for black-box transferable attacks. *E.g.*, a 27.12% ASR drop occurs in TR tasks when transferring from ALBEF to TCL. This validates the considerable contribution of $\mathcal{L}_{CL}$ to guarantee a successful attack. Regarding the influence of $\mathcal{L}_{Dis}$, we demonstrate that the unimodal guidance further enhances the attack on the basis of $\mathcal{L}_{CL}$. Especially in white-box scenarios, $\mathcal{L}_{Dis}$ brings a 10.59% increase in the ASR of TR tasks on ALBEF. The proposed two loss terms complement each other and jointly underpin the generalizability and transferability of the produced UAP.

Table 4. ASR (%) of C-PGC and its variants averaged across six target models on retrieval tasks.

Method	ALBEF		TCL		X-VLM		CLIP$_{\text{ViT}}$		CLIP$_{\text{CNN}}$		BLIP	
	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
C-PGC	**90.13**	**88.82**	**62.11**	**64.48**	**20.53**	**39.38**	**43.10**	**65.93**	**54.40**	**72.51**	**44.79**	**56.36**
C-PGC$_{CL}$	76.46	77.58	34.99	47.55	14.33	33.61	42.98	62.81	46.11	65.58	27.13	46.44
C-PGC$_{Dis}$	79.54	82.46	56.52	62.21	20.24	38.26	39.78	65.10	52.20	71.01	42.43	55.52
C-PGC$_{Rand}$	61.87	65.17	43.69	52.54	19.51	35.47	40.33	65.77	54.15	70.62	39.43	52.59
C-PGC$_{CA}$	85.18	83.07	45.76	53.73	15.24	34.02	39.29	60.61	47.15	40.64	32.39	48.29

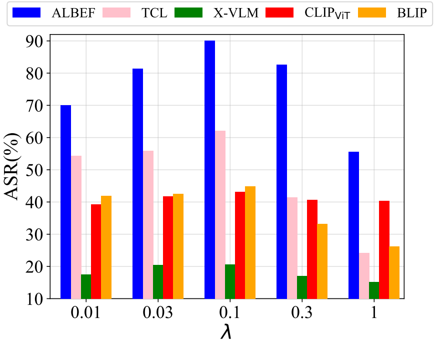

Fig. 4: ASR of five target models on TR tasks under various values of λ.

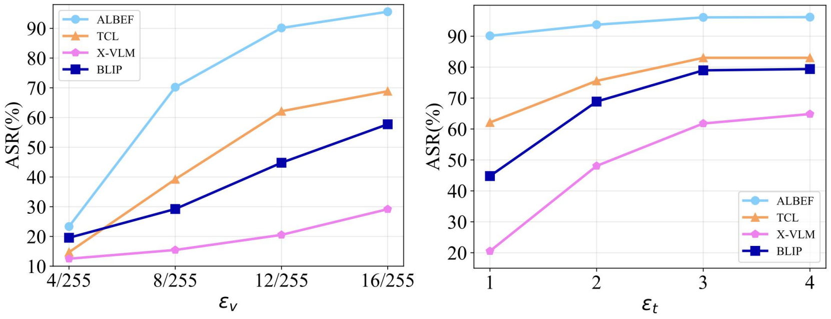

Fig. 5: ASR of five target models on the TR task under different values of perturbation budgets for ϵ_v and ϵ_t respectively.

The effect of positive sample selection. To validate the *farthest selection strategy* for positive sample construction, we design another variant C-PGC$_{Rand}$ that adopts randomly sampled data points as positive samples. Table 4 reveals the necessity of the proposed technique as it brings an average improvement of 25.96% in white-box ASR and 4.95% in black-box ASR. We also find that if positive samples are not adequately defined, adding the $\mathcal{L}_{CL}$ would even harm the white-box performance (see C-PGC$_{CL}$ and C-PGC$_{Rand}$), highlighting the significance of our selection strategy.

The effect of cross-modal conditions. As aforementioned, cross-attention (CA) modules are introduced into the generator to exploit cross-modal information. We then design C-PGC$_{CA}$ that cancels these CA layers to explore the influence. As expected, it causes a notable 9.78% average decrease across six target models, verifying its vital role in boosting attacks. Another finding is that C-PGC$_{CA}$ induces a more pronounced drop in black-box attacks than white-box ones, indicating that cross-modal conditions exert a greater contribution to adversarial transferability.

Different regulatory factor λ. The value of λ is a critical factor as it adjusts the scales of the two loss terms $\mathcal{L}_{CL}$ and $\mathcal{L}_{Dis}$. We evaluate the attack performance under various values of λ to confirm the optimal value. Fig. 4 indicates that $\lambda = 0.1$ achieves superior performance.

Different perturbation budgets ϵ_v and ϵ_t. As shown in Fig. 5, we analyze varying perturbation budgets for ϵ_v and ϵ_t. Generally, the ASR increases with the larger perturbation magnitudes. Note that when $\epsilon_v = 4/255$, C-PGC's performance is severely compromised since the budget $4/255$ is too small to allow the UAP to carry enough information required to generalize to diverse data samples.

It also indicates that the improvement slows down as ϵ_v increases from $12/255$ to $16/255$. Thus, we select the moderate value of $12/255$ to reach a balance between attack utility and imperceptibility. For text perturbation, ϵ_t exhibits a more profound influence on the black-box attacks. In our experiments, we strictly set $\epsilon_t = 1$ for invisibility. Attackers can adjust the value of ϵ_t in accordance with their demands to trade off the attack efficacy and perturbation stealthiness.

5. Conclusion

In this paper, we investigate the challenging task of universal adversarial attacks on VLP models. We begin by revealing the deficiency of existing attacks and empirically explaining the underlying reasons. Based on the analysis, we propose to break the crucial multimodal alignment in VLP models by designing a contrastive-learning generative UAP framework that leverages both unimodal and multimodal information. Extensive experiments validate the efficacy of C-PGC on diverse VLP models and V+L tasks. We highlight that this work makes a significant step in exploring UAP in multimodal attacks and deepens our understanding of the mechanism underlying VLP models.

Acknowledgments

This work is supported in part by the National Science Foundation for Distinguished Young Scholars of China under No. 62425201, the National Natural Science Foundation of China under grant 62171248, 62301189, and Shenzhen Science and Technology Program under Grant KJZD20240903103702004, JCYJ20220818101012025, GXWD20220811172936001.

References

[1] 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. 7

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 3

[3] Gautham Anil, Vishnu Vinod, and Apurva Narayan. Generating universal adversarial perturbations for quantum classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10891–10899, 2024. 3, 4

[4] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018. 3

[5] 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

[6] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. Universal adversarial attacks on text classifiers. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7345–7349. IEEE, 2019. 6

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 1

[8] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 2

[9] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120. Springer, 2020. 2, 3

[10] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 2

[11] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 1

[12] Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4976, 2023. 3

[13] Hao Fang, Jiawei Kong, Bin Chen, Tao Dai, Hao Wu, and Shu-Tao Xia. Clip-guided generative networks for transferable targeted adversarial attacks. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 1

[14] Weiwei Feng, Nanqing Xu, Tianzhu Zhang, and Yongdong Zhang. Dynamic generative targeted attacks with pattern injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16404–16414, 2023. 2

[15] Haoran Gao, Hua Zhang, Jiahui Wang, Xin Zhang, Huawei Wang, Wenmin Li, and Tengfei Tu. Nuat-gan: Generating black-box natural universal adversarial triggers for text classifiers using generative adversarial networks. *IEEE Transactions on Information Forensics and Security*, 2024. 2, 3, 4

[16] Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*. 3

[17] Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4005–4014, 2023. 3

[18] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49. IEEE, 2018. 3

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 5

[20] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 684–696, 2019. 3

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 6

[22] Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learn-

ing with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 2, 3, 6, 7

[24] 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. 2, 3, 6

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7

[27] 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. 6

[28] Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4435–4444, 2023. 3

[29] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. 2, 3, 5, 6, 7

[30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 2

[31] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. 2, 3

[32] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018. 3

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7

[34] 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. 6

[35] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 2, 6, 7

[36] Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *European Conference on Computer Vision*, pages 109–126. Springer, 2024. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 7

[39] Donghua Wang, Wen Yao, Tingsong Jiang, and Xiaoqian Chen. Improving transferability of universal adversarial perturbation with feature disruption. *IEEE Transactions on Image Processing*, 2023. 3

[40] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 102–102. IEEE Computer Society, 2024. 2, 3, 5, 6

[41] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016. 3

[42] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 3, 6

[43] 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. 6

[44] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3340–3349, 2021. 3

[45] 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. 6

[46] Wenbo Yu, Hao Fang, Bin Chen, Xiaohang Sui, Chuan Chen, Hao Wu, Shu-Tao Xia, and Ke Xu. Gi-nas: Boosting gradient inversion attacks through adaptive neural architecture search. *arXiv preprint arXiv:2405.20725*, 2024. 3

[47] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International conference on machine learning*, pages 25994–26009. PMLR, 2022. 6

[48] 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. 1, 2, 3, 5, 6

[49] Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 862–871, 2024. 2, 3, 6, 7

[50] Ziqi Zhou, Shengshan Hu, Ruizhi Zhao, Qian Wang, Leo Yu Zhang, Junhui Hou, and Hai Jin. Downstream-agnostic adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4345–4355, 2023. 3

[51] Tianqu Zhuang, Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, and Shu-Tao Xia. Stealthy shield defense: A conditional mutual information-based approach against black-box model inversion attacks. In *The Thirteenth International Conference on Learning Representations*. 3