Asymmetric Relational-Geometry Driven Universal Adversarial Perturbations for Vision-Language Models

Jiaxin Ye[1] **Weihai Li**[1*] **Ying Wang**[2] **Simeng Qin**[3] **Zhitao Zeng**[4] **Zikai Xu**[1]

[1]University of Science and Technology of China, China
[2]Beijing University of Technology, China
[3]Northeastern University, Shenyang, China
[4]National University of Singapore, Singapore

[*]Corresponding author: whli@ustc.edu.cn

Abstract

Although vision-language pre-trained (VLP) models have achieved remarkable success across multimodal tasks, they remain vulnerable to adversarial perturbations. Existing universal adversarial perturbation (UAP) methods in multimodal settings—whether generator-based or optimization-based—often suffer from limited cross-model transferability, especially in black-box scenarios. We attribute this limitation to the prevalent use of symmetric or distribution-level objectives that overlook the asymmetric roles of image and text modalities and the relational nature of vision-language representations. To address this issue, we propose ARG-Attack, an optimization-based framework that learns universal perturbations under an asymmetric relational-geometry driven objective. Our method integrates three complementary components: a cosine-based loss that induces directional semantic drift in visual features, a center shift loss that geometrically regularizes adversarial embeddings toward a shared semantic center, and a relational polarity loss that explicitly disrupts image–text matching relationships. Together, these objectives enable effective cross-modal interaction without relying on model-specific training losses or probabilistic distribution matching. In addition, we adopt an adaptive gradient update strategy inspired by Adam optimization to stabilize training and accelerate convergence. Extensive experiments across multiple vision-language models and tasks demonstrate that ARG-Attack achieves competitive white-box performance and significantly outperforms state-of-the-art methods in black-box transfer settings.

1 Introduction

Vision-language pre-trained (VLP) models have achieved remarkable progress in bridging vision and language(Kim et al., 2021; Huang et al., 2020; Zhu et al., 2022; Li et al., 2021; Radford et al., 2021), enabling strong performance in multimodal

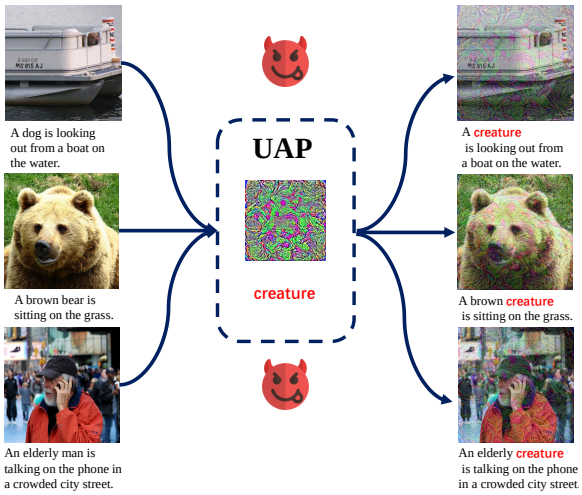

Figure 1: Illustration of universal adversarial attacks targeting vision-language models. By utilizing a single pair of image and text perturbations, the proposed attack can deceive various models across VL tasks.

tasks such as image-text retrieval, image captioning, and visual grounding. Representative models including CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and ALBEF (Li et al., 2021) learn joint visual-textual representations from large-scale image-text corpora and have become widely adopted in both academic and industrial applications. However, as their usage spreads to high-stakes settings, concerns regarding their robustness against adversarial inputs become increasingly critical(Eykholt et al., 2018).

Recent studies show that VLP models are vulnerable to multimodal adversarial examples, where imperceptible perturbations can significantly alter model predictions (Zhang et al., 2022a). Early efforts mainly focused on instance-specific attacks (Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Lu et al., 2023; Wang et al., 2024a), which require per-sample optimization and incur high computational cost. To improve scalability, universal adversarial perturbations (UAPs) have

Findings of the Association for Computational Linguistics: ACL 2026, pages 7997–8014
July 2-7, 2026

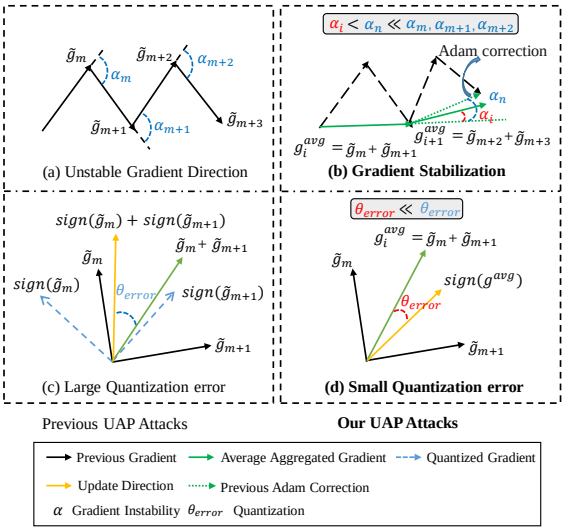

Figure 2: Comparison with prior UAPs. (a)(c): Traditional attacks suffer from unstable gradients and quantization errors. (b)(d): Our method reduces both via Adam-based correction.

been proposed to fool models using a single, input-agnostic perturbation (Moosavi-Dezfooli et al., 2017; Poursaeed et al., 2018; Liu et al., 2023), as illustrated in Figure 1.

Existing multimodal UAP methods generally fall into two categories (Gu et al., 2023): generator-based(Yang et al., 2022b; Mao et al., 2020; Phan et al., 2020) and optimization-based(Byun et al., 2022; Wang et al., 2021; Xiong et al., 2022; Qian et al., 2023; Chen et al., 2023; Zhang et al., 2022b; Wu et al., 2020; Naseer et al., 2021). Generator-based methods rely on training additional neural networks to synthesize perturbations, resulting in high training complexity, significant computational overhead, and limited controllability. Despite these limitations, a recent generator-based method, CPGC (Fang et al., 2024), has achieved state-of-the-art performance in universal multimodal attacks through contrastive training objectives.

In contrast, optimization-based methods directly update perturbations through gradient-based optimization, offering better efficiency and flexibility. However, we observe that most existing approaches adopt *symmetric objectives* for image and text modalities, implicitly assuming similar semantic roles and geometric behaviors, which limits attack transferability.

From an optimization perspective, gradient-based UAP generation also faces two key challenges (Moosavi-Dezfooli et al., 2017; Liu et al., 2023): *gradient instability*, caused by conflicting gradients across samples (Wang et al., 2024b; Wang and He, 2021; Xiong et al., 2022), and *quantization error*, introduced by repeated sign-based gradient discretization (Cheng et al., 2021; Zhang et al., 2022c) (Figure 2). These issues hinder convergence and reduce perturbation generalization.

Motivated by these observations, we propose an optimization-based framework for multimodal UAP generation that explicitly accounts for the asymmetric roles of image and text modalities. Our method jointly perturbs visual and textual inputs under strict imperceptibility constraints, without relying on auxiliary generators. By targeting both global embedding geometry and cross-modal relational structure, the proposed framework achieves strong transferability across diverse vision-language architectures and tasks.

To enable stable and efficient optimization, we adopt an Adam-inspired adaptive gradient aggregation strategy that mitigates gradient instability and reduces quantization error. This strategy accelerates convergence while serving as a supporting optimization component, without altering the core asymmetric relational-geometry formulation. As a result, our method achieves comparable or superior performance to state-of-the-art approaches such as CPGC, while reducing training cost by up to **30**×.

To promote semantic plausibility and cross-model transferability, we introduce an asymmetric relational-geometry driven attack objective. Specifically, cosine-based losses regulate semantic drift, while a relational polarity loss (RPL) explicitly disrupts similarity ordering between matched and mismatched image–text pairs. Unlike distribution-matching objectives, our formulation directly targets relational structures in the joint embedding space, yielding superior black-box transferability.

Our contributions are summarized as follows:

- We propose an **asymmetric multimodal universal adversarial perturbation formulation** that explicitly distinguishes the geometric and relational roles of image and text modalities in vision-language models, enabling more effective and transferable attacks.
- We introduce a **relational-geometry driven attack objective**, centered around a relational polarity loss (RPL), which disrupts cross-modal similarity ordering while preserving semantic coherence, leading to significantly improved black-box transferability across architectures.

- We develop an efficient **optimization-based framework** for multimodal UAP generation, which achieves competitive white-box performance and substantially outperforms state-of-the-art methods such as CPGC in cross-model and cross-task evaluations.

2 Related Work

2.1 Vision-Language Pre-training Models

Vision-Language Pre-training (VLP) underpins a wide range of multimodal tasks by learning aligned image–text representations from large-scale paired data. Existing VLP models generally follow two architectural paradigms (Li et al., 2021): aligned models (e.g., CLIP (Radford et al., 2021)), which encode images and texts separately and align them in a shared embedding space via contrastive learning, and fused models (e.g., ALBEF (Li et al., 2021), BLIP (Li et al., 2022), X-VLM (Zeng et al., 2021)), which integrate visual and textual features through cross-modal attention. While achieving state-of-the-art performance on downstream tasks such as image–text retrieval, captioning, and visual grounding, their robustness under adversarial settings remains underexplored.

2.2 Adversarial Attacks on VLPs

Early adversarial research primarily focused on unimodal settings, with classical methods such as FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017), and UAP (Moosavi-Dezfooli et al., 2017) laying the foundation for input perturbation. Initial multimodal attacks extended these techniques by independently perturbing each modality (e.g., PGD for images and token-level substitutions for text), but lacked cross-modal coordination.

Co-Attack (Zhang et al., 2022a) first enabled joint multimodal adversarial optimization via cross-modal gradient fusion, including perturbations on image–text attention maps, substantially improving white-box attack performance on VLP models. However, its objective remains a unified cross-modal similarity manipulation, without explicitly modeling the distinct functional roles of image and text representations in the joint embedding space.

Subsequent methods further explored transferability and optimization. SGA (Lu et al., 2023) proposed a set-level attack that enhances black-box transferability by preserving global distribution alignment and local sample relationships, while CMI-Attack (Fu et al., 2024) introduced dynamic weighting to facilitate cross-modal gradient collaboration for improved instance-level optimization. Despite their effectiveness, these approaches mainly emphasize optimization or distributional alignment, relying on per-sample objectives shared across modalities, which limits scalability and makes transferability sensitive to model-specific embedding geometries.

While instance-specific attacks such as SGA and CMI-Attack demonstrate strong transferability across vision–language models, their reliance on per-sample optimization severely limits scalability in large-scale or real-time settings. To address this issue, universal adversarial perturbations (UAPs) (Moosavi-Dezfooli et al., 2017; Khrulkov and Oseledets, 2018; Mopuri et al., 2018; Co et al., 2021; Zhang et al., 2021) have emerged as a scalable alternative. By learning a single, input-agnostic perturbation that generalizes across samples, UAPs substantially reduce computational cost. However, achieving strong cross-model transferability in multimodal settings remains challenging without explicitly modeling cross-modal relational structures.

Building on this direction, recent studies have explored UAPs for vision–language pre-training models. AdvCLIP (Zhou et al., 2023) proposed downstream-agnostic adversarial patches for CLIP by attacking its contrastive objective, but its reliance on visible perturbations and model-specific design limits general applicability. ETU (Zhang et al., 2024) improves black-box transferability via local–global feature mixing, yet still adopts a unified objective across modalities. C-PGC (Fang et al., 2024) generates joint image–text UAPs through contrastive learning, but treats both modalities symmetrically and relies on an auxiliary generator, without explicitly modeling the asymmetric geometric anchoring of visual representations and the relational matching behavior of text. Consequently, existing multimodal UAP methods remain constrained in transferability, efficiency, and semantic stability.

In contrast to prior generator-based and symmetric optimization-based UAP methods, we propose a lightweight optimization-based framework that explicitly introduces an *objective-level asymmetric formulation*. By distinguishing the geometric and relational roles of image and text modalities in the joint embedding space, our method achieves improved efficiency and substantially enhanced black-box transferability, as detailed in Section 3.

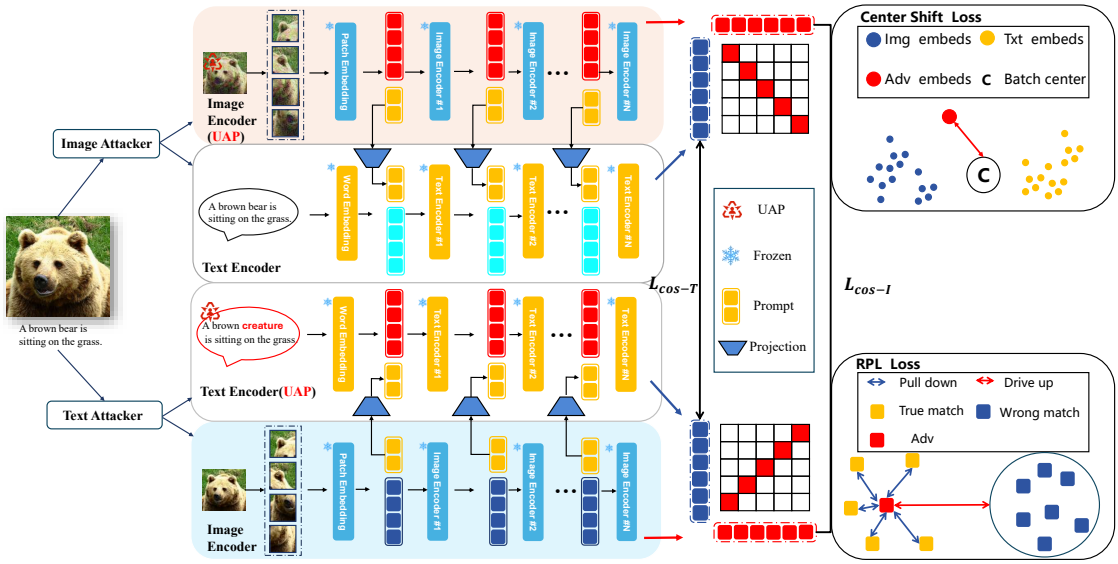

Figure 3: Overview of our multimodal universal adversarial perturbation framework. Two shared perturbations (image Δ_i and text Δ_t) are jointly optimized via gradient-based updates with feedback from surrogate model losses.

3 Methodology

3.1 Problem Definition

Our goal is to learn a pair of universal perturbations (Δ_i, Δ_t) that can fool the model across most natural inputs from a given distribution, while allowing modality-specific perturbation behaviors.

Formally, given an input distribution $\mathcal{D}$, the objective is to solve:

$$(\Delta_i^*, \Delta_t^*) = \arg\max_{\Delta_i,\, \Delta_t} \mathbb{E}_{(i,t)\sim\mathcal{D}} \mathcal{L}\big(f(i+\Delta_i,\ t\oplus\Delta_t),\ y\big). \tag{1}$$

$$\text{s.t.} \quad \|\Delta_i\|_\infty \le \epsilon_i, \qquad \|\Delta_t\|_0 \le \epsilon_t. \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes a composite attack objective that may incorporate modality-specific geometric and relational loss components. The operator $\oplus$ represents a text perturbation mechanism (Zhang et al., 2022a) that replaces one or more tokens, inducing non-continuous perturbations in the textual embedding space.

To ensure the imperceptibility of the generated perturbations, we apply strict budget constraints to both modalities: the image perturbation is bounded by an ℓ_∞ norm with a pixel-level budget of $\epsilon_i = 12/255$, while the text perturbation is restricted to a single token replacement (i.e., $\epsilon_t = 1$).

We further consider a black-box transfer setting where the attacker has no access to the target model's internal architecture or parameters, observing only its output predictions. In this case, adversarial perturbations are generated using a set of surrogate models $\{f_k\}_{k=1}^K$ and transferred to the target model f_{target}.To evaluate transferability, we use the Attack Success Rate (ASR) defined as:

$$\text{ASR} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(i,t)\in\mathcal{D}_{\text{test}}} \mathbb{I}[\Delta(i,t)]\,, \tag{3}$$

$$\Delta(i,t) \triangleq f_{\text{target}}(i+\Delta_i,\ t\oplus\Delta_t) \neq f_{\text{target}}(i,t). \tag{4}$$

In essence, the core challenge lies in crafting perturbations that are semantically aligned, architecture-agnostic, and resilient to distributional and geometric variations across diverse VLMs.

3.2 The Proposed Attack Framework Overview

Figure 3 illustrates the overall pipeline of our multimodal UAP framework. We aim to jointly optimize two universal perturbations: an image-space perturbation Δ_i and a text-space perturbation Δ_t, such that perturbed image–text pairs $(i+\Delta_i,\ t\oplus\Delta_t)$ can mislead VLMs, while preserving modality-specific perturbation behaviors.

We initialize Δ_i and Δ_t and inject them into batches of image–text pairs sampled from the training set. For the image branch, Δ_i is applied to each image followed by data augmentations, including geometric transformations (e.g., padding, cropping,

resizing) and photometric distortions (e.g., brightness, contrast, saturation, and hue). These augmentations improve transformation robustness and enhance cross-model transferability. For the text branch, Δ_t operates at the token level via universal token replacement, introducing discrete and non-continuous perturbations.

The augmented images and modified texts are processed by a surrogate vision-language model with separate image and text encoders. Instead of adopting a symmetric objective across modalities, our framework employs a modality-aware optimization strategy that explicitly accounts for the distinct geometric and relational roles of image and text embeddings in the joint representation space.

Specifically, we design a hybrid loss consisting of three complementary components. A cosine consistency loss regularizes geometric alignment between clean and adversarial embeddings to prevent excessive semantic drift. A center shift loss encourages adversarial embeddings to collapse toward modality-specific feature centers, degrading instance-level discrimination. Finally, a relational polarity loss explicitly disrupts cross-modal correspondence by suppressing similarities of matched image–text pairs while amplifying mismatched relations at the batch level.

The total loss is backpropagated to iteratively update the shared perturbations Δ_i and Δ_t using accumulated gradients across batches. This lightweight, optimization-based framework yields a single pair of universal perturbations that are geometry-aware, relation-disruptive, and highly transferable across diverse VLMs and multimodal tasks.

3.3 The Proposed Hybrid Loss Function Design

We introduce a hybrid loss function for optimizing multimodal universal adversarial perturbations, with the key principle of explicitly decoupling *geometric degradation* and *relational disruption* in the joint vision-language embedding space. Although image and text attackers operate on different input domains, they share a similar optimization pipeline. Without loss of generality, we describe the loss formulation using the image attacker, while the text branch follows an analogous design. The overall objective consists of three complementary components: a cosine-based geometric consistency loss, a center shift loss that degrades instance-level structures, and a relational polarity loss that disrupts cross-modal correspondence. The image attacker pseudocode is provided in Appendix A.

Cosine Consistency Loss. We adopt a cosine-based consistency loss to induce semantic deviation between adversarial and clean visual embeddings. Despite its name, this loss is optimized in an *anti-consistency* manner by minimizing cosine similarity, thereby encouraging controlled semantic drift that enhances attack transferability.

Formally, let $v_{\text{adv}}^{(i)}$ and v_i denote the adversarial and clean visual embeddings of the i-th sample in a mini-batch of size N, extracted from the surrogate image encoder. The cosine consistency loss is defined as:

$$\mathcal{L}\text{cos} = -\frac{1}{N}\sum i = 1^N \frac{1}{\tau}\frac{\langle v_{\text{adv}}^{(i)}, v_i\rangle}{|v_{\text{adv}}^{(i)}| \cdot |v_i|} \quad (5)$$

Here, τ is a temperature hyperparameter controlling similarity sharpness. By explicitly pushing adversarial features away from their clean counterparts, this loss undermines feature-level consistency and promotes transferable semantic drift across vision-language models.

Center Shift Loss. To regulate the global geometric structure of adversarial embeddings, we introduce a **Center Shift Loss**. Unlike probabilistic divergence objectives, this loss operates purely in the embedding space and softly constrains adversarial features toward a batch-level semantic center, aligning naturally with the geometric properties of vision–language representations. In contrast to the cosine consistency loss, which induces pairwise directional deviation, the center shift loss shapes global embedding geometry at the batch level.

Let $\mathbf{v}_{\text{adv}}^{(i)} \in \mathbb{R}^D$ denote the adversarial embedding of the i-th sample, and let $\{\mathbf{v}_j\}_{j=1}^N$ represent the corresponding clean embeddings from either modality within a mini-batch. The batch center is computed as

$$\mathbf{c} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{v}_j, \quad (6)$$

followed by ℓ_2 normalization. The center shift loss is then defined as

$$\mathcal{L}_{\text{center}} = \frac{1}{N}\sum_{i=1}^{N}\cos\left(\frac{\mathbf{v}_{\text{adv}}^{(i)}}{\|\mathbf{v}_{\text{adv}}^{(i)}\|}, \frac{\mathbf{c}}{\|\mathbf{c}\|}\right). \quad (7)$$

By maximizing this similarity, adversarial embeddings are constrained toward a shared geometric region, degrading instance-level discrimination while preserving global semantic coherence. This geometric regularization prevents drift into

modality-specific or out-of-manifold regions, enhancing stability and cross-model transferability.

Relational Polarity Loss (RPL). While geometric regularization weakens instance-level separability, effective multimodal attacks must further disrupt the relational structure between images and texts. To this end, we introduce a **Relational Polarity Loss (RPL)**, which explicitly reverses the similarity polarity between matched and mismatched image–text pairs in the joint embedding space.

Let $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{B_v}$ and $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^{B_t}$ denote the ℓ_2-normalized adversarial image embeddings and clean text embeddings, respectively. The cross-modal similarity matrix is computed as

$$\mathbf{S} = \mathbf{V}\mathbf{T}^\top \in \mathbb{R}^{B_v \times B_t}. \tag{8}$$

Given pairing annotations, we define a binary mask $\mathbf{P} \in \{0,1\}^{B_v \times B_t}$, where $\mathbf{P}_{ij} = 1$ denotes a matched pair. The RPL is then formulated as

$$\mathcal{L}_{\text{RPL}} = \frac{1}{B_v} \sum_{i=1}^{B_v} \Big(\frac{\sum_j \mathbf{S}_{ij}(1 - \mathbf{P}_{ij})}{\sum_j (1 - \mathbf{P}_{ij})} - \frac{\sum_j \mathbf{S}_{ij}\mathbf{P}_{ij}}{\sum_j \mathbf{P}_{ij}} \Big). \tag{9}$$

By suppressing matched-pair similarities and amplifying mismatched ones, RPL inverts the relational polarity encoded in vision–language representations. Importantly, this loss is defined purely on similarity statistics without relying on model-specific heads, making it architecture-agnostic and well-suited for black-box transfer attacks.

Total Loss. The final optimization objective is a weighted combination of the three complementary losses:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{cos}} + \beta \cdot \mathcal{L}_{\text{center}} + \gamma \cdot \mathcal{L}_{\text{RPL}}, \tag{10}$$

where $\mathcal{L}_{\text{cos}}$ induces directional semantic drift, $\mathcal{L}_{\text{center}}$ regularizes global embedding geometry, and $\mathcal{L}_{\text{RPL}}$ disrupts cross-modal relational consistency. The weights α, β, and γ control each component and are determined empirically. Together, these losses form a hybrid objective that integrates geometric control and relational manipulation, yielding transferable universal perturbations in black-box multimodal attack settings.

4 Experiment

4.1 Experimental Setup

We conduct comprehensive experiments across multiple vision-language tasks, datasets, and pre-trained models to evaluate the effectiveness of our universal multimodal attack.

Tasks and Datasets. We evaluate three core tasks: image-text retrieval (ITR), image captioning (IC), and visual grounding (VG). ITR is conducted on both Flickr30K (Plummer et al., 2015) and MSCOCO (Lin et al., 2014), IC on MSCOCO, and VG on RefCOCO+ (Yu et al., 2016). Full dataset details are provided in Appendix B.

Models. We consider six widely-used vision-language models: ALBEF (Li et al., 2021), BLIP (Li et al., 2022), TCL (Yang et al., 2022a), X-VLM (Zeng et al., 2021), and two CLIP variants (ViT-B/16(Radford et al., 2021) and RN101). All models are used in ITR; IC attacks are generated using ALBEF/TCL/BLIP and tested on BLIP; VG experiments target ALBEF.

Evaluation Metric. For ITR, we compute Attack Success Rate (ASR) via Recall@1 degradation. For IC and VG, we use task-specific metrics (e.g., CIDEr, SPICE, grounding accuracy) reported in the corresponding result sections.

Baselines. We compare with GAP (Poursaeed et al., 2018), adapted to the multimodal setting, and CPGC (Fang et al., 2024), a recent SOTA method. DO-UAP (Yang et al., 2024) is additionally included in our time comparison for completeness, but is not used as a main baseline in ITR experiments due to its limited cross-model transferability.

Implementation. Visual perturbations are bounded by $\epsilon_i = 12/255$; text perturbations allow one token change ($\epsilon_t = 1$). UAPs are trained for 2 epochs with our adaptive gradient strategy, using data augmentations (resize, pad, color jitter). All experiments are conducted on a single NVIDIA A100 GPU.

4.2 Transferability of Universal Adversarial Perturbations

Attack Performance on Flickr30K. Table 1 reports the Attack Success Rate (ASR) under white-box and black-box settings. In the white-box scenario, our method achieves competitive performance compared to the state-of-the-art CPGC (Fang et al., 2024), e.g., 83.56% TR and 85.41% IR vs. 90.13%/88.82% using ALBEF as the source model. In all cases, it clearly outperforms the baseline GAP (Poursaeed et al., 2018).

In black-box settings, which better reflect real-world adversarial scenarios, our method substantially improves transferability. For example, when transferring from ALBEF to X-VLM, our method

Table 1: **Attack Success Rate (ASR %) for image-text retrieval tasks on Flickr30k.** TR indicates text retrieval based on the input image, while IR denotes image retrieval using input text.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Flickr30k	ALBEF	GAP	69.78	81.59	22.15	29.97	6.61	18.37	23.4	37.54	29.92	44.29	16.09	28.12
		CPGC	**90.13**	**88.82**	62.11	64.48	20.53	39.38	43.1	65.93	54.4	72.51	44.79	56.36
		ours	83.56	85.41	**62.84**	**69.25**	**27.44**	**50.86**	42.98	**68.33**	**54.4**	**75.77**	**55.63**	**66.63**
	TCL	GAP	33.5	40.61	82.41	80.67	6.61	17.79	21.55	38.56	30.57	45.48	21.45	31.82
		CPGC	50.26	56.29	**94.93**	**90.64**	14.94	33.96	46.92	66.41	52.98	70.66	35.75	52.52
		ours	**53.44**	**66.6**	88.92	86.67	**28.35**	**51.51**	**52.71**	**73.05**	**59.84**	**77.77**	**53.52**	**69.03**
	X-VLM	GAP	16.14	24.43	17.08	26.2	90.24	85.98	24.51	41.15	42.62	53.08	16.19	25.74
		CPGC	24.46	47.77	29.19	50.15	93.29	91.9	43.47	66.03	59.2	72.79	32.39	52.24
		ours	**29.29**	**49.22**	**36.54**	**52.73**	**96.54**	**91.97**	**44.58**	**66.54**	57.9	**72.83**	**35.86**	**53.63**
	CLIPvit	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
		CPGC	23.23	38.67	25.05	41.79	15.85	35.59	88.92	**93.05**	66.06	75.42	26.71	45.7
		ours	**24.25**	**40.96**	**30.75**	**44.48**	**20.53**	**38.98**	**89.53**	92.34	**70.21**	**79.34**	**28.81**	**48.29**
	CLIPcnn	GAP	13.57	25.21	19.05	28.87	11.59	23.13	27.46	43.16	73.18	81.6	15.25	27.94
		CPGC	15.31	38.93	19.77	**43.72**	17.17	41.65	39.9	64.82	81.74	88.9	22.19	46.11
		ours	**26.52**	**41.23**	**25.78**	42.83	**25.41**	**42.31**	**50.25**	**69.92**	**90.67**	**92.58**	**29.23**	**51.11**
	BLIP	GAP	12.23	23.94	14.49	25.44	6.91	17.81	20.32	37	26.81	43.59	47.21	73.33
		CPGC	32.17	44.4	33.44	44.51	**18.6**	**35.53**	43.35	60.26	48.96	66.95	**71.82**	**82.82**
		ours	**34.53**	**47.95**	**38.41**	**49.34**	15.65	32.52	**46.43**	**62.68**	**54.53**	**70.1**	62.15	73.24

achieves 27.44% TR and 50.86% IR, outperforming CPGC (20.53%/39.38%). Similar gains appear in other configurations, e.g., a 10-point IR improvement on BLIP. These results collectively highlight our method's robustness and superior generalization across diverse target models, without requiring internal access to their architectures.

Attack Performance on MSCOCO. We observe consistent trends on the MSCOCO dataset (see Appendix C), where our method continues to outperform prior methods in both white-box and black-box scenarios, further demonstrating its robustness and strong generalization.

Results of R@5 and R@10. To further evaluate the effectiveness of our proposed ARG method under relaxed retrieval metrics, we report attack success rates based on R@5 and R@10 in Appendix F. As the results show, ARG basically outperforms both CPGC and GAP across all settings, demonstrating superior cross-model attack capability and stronger robustness across different retrieval thresholds.

Comparison of Time Consumption. To assess the practical efficiency of our method, we compare the training time required to other universal perturbations on the Flickr30K dataset using ALBEF as the surrogate model. As shown in Table 2, our method converges in just 4.15 hours—significantly faster than CPGC (Fang et al., 2024) (153.52 hours), and more efficient than GAP (Poursaeed et al., 2018) (5.92 hours) and DO-UAP (Yang et al., 2024) (6.84 hours). Despite this substantial speedup, our method achieves superior black-box transferability and maintains white-box performance comparable to CPGC. These results highlight both the computational efficiency and strong generalization of our method, making it well-suited for real-world large-scale adversarial applications.

4.3 Evaluation on More Downstream Tasks

To further assess our attack on generative multimodal tasks, we evaluate it on the image captioning (IC) task, which aims to generate natural language descriptions aligned with visual content. ALBEF, TCL, and BLIP are used as source models, with BLIP as the target to evaluate transferability. Experiments are conducted on the MSCOCO dataset, and performance is measured using standard captioning metrics: BLEU@4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016).

Table 2: Comparison of training time across different UAP generation methods on Flickr30K using ALBEF as the surrogate model.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIP_{ViT}		CLIP_{CNN}		BLIP		Time (hours)
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	
Flickr30k	ALBEF	GAP	69.78	81.59	22.15	29.97	6.61	18.37	23.4	37.54	29.92	44.29	16.09	28.12	5.92
		CPGC-40	90.13	88.82	62.11	64.48	20.53	39.38	43.1	65.93	54.4	72.51	44.79	56.36	153.52
		DO-UAP	**94.24**	**93.06**	56.31	63.95	20.12	41.67	40.64	61.88	44.56	65.93	45.11	57.63	6.84
		Ours	83.56	85.41	**62.84**	**69.25**	**27.44**	**50.86**	42.98	**68.33**	**54.4**	**75.77**	**55.63**	**66.63**	**4.15**

Table 3: Attack results of image captioning on MSCOCO. Baseline refers to performance on clean images.

Source	B@4	METEOR	ROUGE_L	CIDEr	SPICE
Baseline	39.7	31.0	60.0	133.3	23.8
ALBEF	26.6	23.4	50.0	87.4	16.4
TCL	26.3	23.5	49.6	87.8	16.5
BLIP	**24.5**	**22.6**	**48.4**	**82.0**	**15.6**

As shown in Table 3, the BLIP model achieves 39.7 BLEU@4 and 133.3 CIDEr on clean images. After applying our universal adversarial perturbations, all metrics drop significantly. For instance, when perturbations generated from TCL are transferred to BLIP, BLEU@4 and CIDEr decrease to 26.3 and 87.8, a 33.4% and 34.1% relative drop, respectively. Furthermore, using BLIP as the source model results in even stronger attacks, yielding the lowest scores across all metrics (e.g., 24.5 BLEU@4, 15.6 SPICE). These results demonstrate the powerful disruption capacity and cross-model transferability of our method in the generative captioning task. For qualitative examples, please refer to the IC visualization provided in Appendix 5.

Additional Results. We provide evaluations on two additional downstream tasks: visual grounding and cross-dataset transfer. Please refer to Appendix D.1, Appendix D.2.

4.4 Ablation Study

We conduct ablation studies on the Flickr30K dataset using ALBEF as the surrogate model under a challenging cross-model transfer setting. Specifically, adversarial perturbations are optimized on ALBEF and directly evaluated on CLIPcnn, whose architecture and training paradigm differ substantially from ALBEF, enabling a faithful assessment of transferability.To evaluate the contribution of each component, we remove one loss term at a time from the overall objective, including the cosine consistency loss ($\mathcal{L}\text{cos}$), the center shift loss ($\mathcal{L}_{\text{CS}}$), and the relational polarity loss ($\mathcal{L}_{\text{RPL}}$). Text Retrieval (TR) and Image Retrieval (IR) performance on CLIPcnn are reported in Appendix 7.

As shown in the table, removing any single component consistently degrades transfer performance, indicating that all loss terms contribute complementary effects. In particular, the center shift loss is crucial for maintaining geometric stability across architectures, while the relational polarity loss plays a key role in disrupting cross-modal associations.

Overall, these results demonstrate that the proposed loss design generalizes well to architecturally distinct vision–language models, confirming its robustness in black-box transfer scenarios.Additional ablations on hyperparameter sensitivity and perturbation budgets are provided in Appendix E.

5 Conclusion

This paper investigates the challenge of generating universal adversarial perturbations (UAPs) for vision-language models (VLMs) and proposes ARG-Attack, an efficient gradient-based framework that jointly perturbs image and text inputs. By integrating an Adam-inspired adaptive gradient optimization strategy with an asymmetric relational-geometry driven loss design, our method significantly improves cross-model transferability while maintaining competitive white-box performance. Unlike prior approaches that rely on symmetric or distribution-level objectives, ARG-Attack explicitly models directional semantic drift, geometric regularization, and cross-modal relational disruption to construct more transferable universal perturbations. Compared to existing methods such as CPGC, ARG-Attack achieves substantially faster convergence while demonstrating superior generalization across multiple VLP models and representative downstream tasks. We hope this work sheds new light on the geometric and relational vulnerabilities of vision-language models and provides a solid foundation for future research on transferable multimodal adversarial attacks and robust vision–language representation learning.

6 Limitations

While our method demonstrates superior black-box transferability and substantially reduces training cost compared to existing state-of-the-art methods, it still exhibits certain limitations. Specifically, in white-box scenarios, our performance—though competitive—does not universally surpass prior generator-based methods such as CPGC across all models and metrics. This suggests room for further refinement, particularly in enhancing architecture-specific fitting and fully leveraging model-internal gradients. Future work may explore hybrid optimization-generation schemes or adaptive task-aware loss weighting to further bridge this gap.

7 Impact Statement

This work aims to deepen our understanding of the vulnerabilities in widely deployed vision-language models (VLMs) by introducing a computationally efficient and transferable universal adversarial attack framework. On the positive side, our findings can serve as valuable diagnostic tools for the machine learning community and practitioners, helping to identify blind spots in VLMs, strengthen defenses, and guide the development of more robust multimodal systems in safety-critical applications such as autonomous vehicles, medical imaging, or AI-assisted decision-making.

However, as with any research on adversarial attacks, there is potential for misuse. The proposed method could be repurposed to deceive multimodal AI systems in real-world deployments, potentially leading to misinformation, security breaches, or harmful manipulation of AI outputs. To mitigate such risks, our work is intended solely for academic purposes and can support the creation of standardized adversarial benchmarks, model auditing protocols, and robust training schemes. We encourage responsible use by sharing results in controlled research settings and highlighting defense strategies alongside attack capabilities.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. 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.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. 2022. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253.

Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. 2023. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4489–4498.

Yaya Cheng, Jingkuan Song, Xiaosu Zhu, Qilong Zhang, Lianli Gao, and Heng Tao Shen. 2021. Fast gradient non-sign methods. *arXiv preprint arXiv:2110.12734*.

Kenneth T Co, Luis Muñoz-González, Leslie Kanthan, Ben Glocker, and Emil C Lupu. 2021. Universal adversarial robustness of texture and shape-biased models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 799–803. IEEE.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Hao Wu, Shutao Xia, and Ke Xu. 2024. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. *arXiv preprint arXiv:2406.05491*.

Jiyuan Fu, Zhaoyu Chen, Kaixun Jiang, Haijing Guo, Jiafeng Wang, Shuyong Gao, and Wenqiang Zhang. 2024. Improving adversarial transferability of vision-language pre-training models through collaborative multimodal interaction. *arXiv preprint arXiv:2403.10883*.

Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. 2024. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, pages 442–460. Springer.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, and 1 others. 2023. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Valentin Khrulkov and Ivan Oseledets. 2018. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8562–8570.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. 2023. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4435–4444.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. 2023. 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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Xiaofeng Mao, Yuefeng Chen, Yuhong Li, Yuan He, and Hui Xue. 2020. Gap++: Learning to generate target-conditioned adversarial examples. *arXiv preprint arXiv:2006.05097*.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. 2018. Nag: Network for adversary generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 742–751.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. 2021. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Huy Phan, Yi Xie, Siyu Liao, Jie Chen, and Bo Yuan. 2020. Cag: A real-time low-cost enhanced-robustness high-transferability content-aware adversarial attack generator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5412–5419.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4422–4431.

Yaguan Qian, Shuke He, Chenyu Zhao, Jiaqiang Sha, Wei Wang, and Bin Wang. 2023. Lea2: A lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4510–4521.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. 2024a. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1722–1740. IEEE.

Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Pinxue Guo, Haijing Guo, and Wenqiang Zhang. 2024b. Boosting the transferability of adversarial attacks with global momentum initialization. *Expert Systems with Applications*, 255:124757.

Xiaosen Wang and Kun He. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1924–1933.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. 2021. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16158–16167.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. 2020. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1161–1170.

Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14983–14992.

Fan Yang, Yihao Huang, Kailong Wang, Ling Shi, Geguang Pu, Yang Liu, and Haoyu Wang. 2024. Efficient and effective universal adversarial attack against vision-language pre-training models. *arXiv preprint arXiv:2410.11639*.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022a. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2022b. Boosting transferability of targeted adversarial examples via hierarchical generative networks. In *European conference on computer vision*, pages 725–742. Springer.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. 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.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. 2021. Data-free universal adversarial perturbation and black-box attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7868–7877.

Jiaming Zhang, Qi Yi, and Jitao Sang. 2022a. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013.

Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. 2022b. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14993–15002.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. 2024. 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.

Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. 2022c. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pages 26693–26712. PMLR.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. 2023. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815.

Appendix

A Algorithm Description: Image Attacker

As illustrated in Algorithm 1, we adopt an Adam-style adaptive gradient optimization strategy to generate universal adversarial perturbations (UAPs) for vision–language models. The algorithm operates on a multimodal training set D_s, and iteratively updates a shared image-space perturbation Δ_i (and its text counterpart when applicable) to ensure transferability across diverse samples and architectures.

At each optimization step, a minibatch of image–text pairs is sampled from D_s. For the image branch, stochastic visual augmentations (e.g., color jittering, random cropping, and padding) are applied to simulate realistic distributional shifts and enhance robustness. The universal image perturbation Δ_i is then added to the augmented images to form adversarial inputs v_{adv}.

The adversarial and clean images are forwarded through the frozen image encoder to obtain their corresponding embeddings. The overall attack objective is formulated as a weighted sum of three complementary losses: (1) a *cosine consistency loss* $\mathcal{L}_{\text{cos}}$, which promotes directional semantic drift between adversarial and clean visual embeddings; (2) a *center shift loss* $\mathcal{L}_{\text{center}}$, which geometrically regularizes adversarial embeddings toward the batch-level semantic center induced by clean image and text features, preventing modality-specific drift and improving optimization stability; and (3) a *relational polarity loss* $\mathcal{L}_{\text{rpl}}$, which explicitly disrupts image–text correspondence by suppressing similarities of true image–text pairs while amplifying mismatched associations, thereby degrading cross-modal relational structure.

Gradients of the aggregated loss are backpropagated to the universal perturbation and updated using Adam with first- and second-order moment estimates. After bias correction, the perturbation is updated with an adaptive step size and projected onto the ℓ_∞-ball to satisfy the perturbation budget constraint $\|\Delta_i\|_\infty \leq \epsilon$. This iterative process continues for a fixed number of steps until convergence, yielding a single universal perturbation that generalizes across samples, tasks, and model architectures.

Algorithm 1 Asymmetric Relational-Geometry Driven Image UAP Generation

Require: Surrogate VLP model f with image encoder f_I and text encoder f_T; Training set $D_s = \{(v, t)\}$; perturbation budget ϵ; loss weights α, β, γ; step size η; max iterations N

Ensure: Universal image perturbation Δ_i

Initialize universal image perturbation $\Delta_i \leftarrow \mathbf{0}$
Initialize optimizer states (e.g., momentum buffers)
for $n = 1$ to N **do**
  Sample a minibatch $(v, t) \sim D_s$
  Apply data augmentation to images v
  Generate adversarial images:
  $v_{\text{adv}} \leftarrow v + \Delta_i$
  Extract embeddings:
  $\mathbf{z}_{\text{adv}} \leftarrow f_I(v_{\text{adv}}), \quad \mathbf{z}_v \leftarrow f_I(v), \quad \mathbf{z}_t \leftarrow f_T(t)$
  Compute cosine loss:
  $\mathcal{L}_{\text{cos}} \leftarrow -\cos(\mathbf{z}_{\text{adv}}, \mathbf{z}_v)$
  Compute center shift loss:
  $\mathcal{L}_{\text{center}} \leftarrow \cos(\mathbf{z}_{\text{adv}}, \text{Center}(\mathbf{z}_v, \mathbf{z}_t))$
  Compute relational polarity loss (RPL):
  $\mathcal{L}_{\text{rpl}} \leftarrow \mathbb{E}[\text{sim}_{\text{neg}} - \text{sim}_{\text{pos}}]$
  Aggregate loss:
  $\mathcal{L} \leftarrow \alpha\mathcal{L}_{\text{cos}} + \beta\mathcal{L}_{\text{center}} + \gamma\mathcal{L}_{\text{rpl}}$
  Compute gradient $g \leftarrow \nabla_{\Delta_i}\mathcal{L}$
  Update Δ_i using adaptive gradient update
  Project Δ_i onto ℓ_∞-ball: $\|\Delta_i\|_\infty \leq \epsilon$
end for
return Δ_i

B Dataset Details

- **Flickr30K** (Plummer et al., 2015): This dataset includes 31,783 images, each annotated with five diverse human-written captions. It is commonly used for benchmarking ITR performance.
- **MSCOCO** (Lin et al., 2014): A large-scale dataset containing 123,287 images with five associated captions per image. We adopt this dataset for both ITR and IC evaluations.
- **RefCOCO+** (Yu et al., 2016): A referring expression dataset derived from MSCOCO, consisting of 141,564 region-level language annotations. It is widely used for evaluating VG tasks that require localizing referred objects in images.

Table 4: **Attack Success Rate (ASR %) for image-text retrieval tasks on MSCOCO.** TR indicates text retrieval based on the input image, while IR denotes image retrieval using input text.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
MSCOCO	ALBEF	GAP	82.65	84.35	53.6	45.46	15.09	15.64	25.18	29.94	28.06	35.28	37.44	33.61
		CPGC	**96.18**	**95.09**	**82.49**	**76.24**	39.97	48.58	59.71	67.05	61.27	70.8	59.18	63.89
		ours	93.14	93.48	80.82	74.24	**48.68**	**53.79**	**70.09**	**73.93**	**73.57**	**78.55**	**68.45**	**68.57**
	TCL	GAP	55.92	48.22	95.16	92.29	17.34	17.01	28.73	31.19	32.27	39.81	43.59	39.64
		CPGC	76.62	71.17	96.72	93.88	42.99	48.4	**70.32**	**79.08**	**74.1**	**82.97**	62.35	66.97
		ours	**79.71**	**75.95**	**96.85**	**94.25**	**48.24**	**55.56**	69.9	73.23	71.08	77.54	**76.5**	**75.74**
	X-VLM	GAP	26.35	23.72	27.8	22.91	95.1	88.84	32.39	38.16	52	55.4	24.67	22.65
		CPGC	51.46	65.71	52.8	64.99	**98.89**	**95.79**	67.42	75.45	75.49	**82.58**	55.74	66.7
		ours	**58.13**	**69.22**	**64.21**	**69.71**	97.64	94.29	**67.68**	**76.68**	**76.72**	82.16	**61.95**	**69.92**
	CLIPvit	GAP	35.96	31.91	37.33	32.56	33.42	29.25	97.71	96.04	74.63	74.67	33.47	31.99
		CPGC	46.92	**53.89**	46.03	**50.87**	41.49	48.6	**98.74**	98.01	81.58	86.5	47.35	57.55
		ours	**48.72**	51.2	**46.85**	50.04	**43.41**	**48.75**	98.7	**98.14**	**86.15**	**89.56**	**52.97**	**62.04**
	CLIPcnn	GAP	28.67	27.51	29.84	27.69	26.4	24.81	39.64	40.53	90.34	91.56	24.99	26.18
		CPGC	33.38	46.68	40.61	**50.76**	35.34	46.95	63.83	70.15	94.89	94.42	37.38	53.06
		ours	**56.25**	**53.52**	**49.37**	49.64	**49.72**	**52.17**	**74.93**	**80.34**	**96.69**	**97.16**	**56.73**	**62.15**
	BLIP	GAP	35.55	38.75	35.62	33.79	22.7	21.25	32.05	35.8	40.93	45.58	73.46	72.37
		CPGC	61.95	60.92	60.95	59.57	**51.81**	**52.53**	62.23	72.51	69.61	78.44	**91.67**	**90.42**
		ours	**67.34**	**63.33**	**71.27**	**63.32**	45.06	46.99	**69.1**	**75.61**	**74.67**	**80.6**	86.76	85.03

C Attack Performance on MSCOCO

D Extended Evaluation on Downstream Tasks

D.1 Visual Grounding

Visual Grounding (VG) aims to localize a region in the image based on a textual description. This task requires fine-grained cross-modal alignment, and thus can only be conducted on models that support region-level vision-language fusion, such as ALBEF, TCL, and X-VLM.

Table 5: Visual grounding results on RefCOCO+ with ALBEF as the target.

Method	Source	Target	Val	TestA	TestB
Baseline	ALBEF	ALBEF	58.4	65.9	46.2
SGA	ALBEF	ALBEF	50.56	57.42	40.66
VLPT	ALBEF	ALBEF	**49.7**	56.32	40.54
Ours	ALBEF	ALBEF	49.9	**54.7**	**40.0**

We perform experiments on the RefCOCO+ dataset and evaluate the accuracy on three standard splits: Val, TestA, and TestB. We use ALBEF as the target model and generate universal perturbations using ALBEF. Additionally, we compare our method with two representative instance-specific adversarial methods: SGA(Lu et al., 2023) and VLPT(Gao et al., 2024), which are known for strong transferability despite being sample-specific.As shown in Table 5, our universal attack significantly degrades grounding accuracy across all splits. For example, when using ALBEF as the source model, accuracy drops from 58.4 to 49.9 on Val and from 46.2 to 40.0 on TestB, closely approaching or even outperforming instance-specific methods like SGA and VLPT. These results validate the strong attack capability of our universal perturbation in the visual grounding task, even under the more restrictive universal attack setting.

D.2 Cross-domain Transferability

Cross-domain scenarios. We evaluate the cross-domain robustness of our method by generating universal perturbations on MSCOCO and testing them on Flickr30K. As shown in Table 6, despite the domain shift, our method maintains strong ASR across models. Notably, when using TCL as the source, it achieves 83.90% (IR) on TCL and 69.75%(IR) on CLIP_{CNN}, confirming high transferability across both model and dataset boundaries.

E Extended Ablation Studies

Sensitivity to loss weights. We further investigate the influence of the loss weights α, β, and γ, which respectively control the contributions of the cosine semantic drift loss, the center shift loss, and the relational polarity loss in the overall objective. All experiments are conducted in a cross-model

Table 6: (ASR %) of cross-domain attacks on six models from MSCOCO to Flickr30k.

Source	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
ALBEF	79.86	84.52	53.52	60.33	21.85	38.85	46.55	64.56	57.64	69.47	41.32	55.80
TCL	47.89	57.61	85.40	83.90	19.31	39.15	46.92	62.81	54.66	69.75	45.01	57.33
X-VLM	29.80	49.89	34.99	52.85	75.91	77.00	46.80	66.57	58.94	72.76	37.33	54.35
CLIPvit	20.04	38.20	25.67	41.44	16.06	35.92	94.46	95.25	66.58	80.74	28.50	49.86
CLIPcnn	26.10	42.06	26.19	42.49	24.39	41.28	52.83	71.55	92.23	93.73	30.28	50.74
BLIP	37.00	48.33	42.65	48.69	17.99	34.82	47.29	64.27	57.77	72.69	57.52	66.79

transfer setting, where perturbations are optimized on ALBEF and evaluated on the CLIPcnn model. We vary one hyperparameter at a time while fixing the remaining two. As shown in Table 8, Table 9, and Table 10, the best transfer performance on CLIPcnn is consistently achieved when $\alpha = 1$, $\beta = 0.3$, and $\gamma = 1$. Deviating from this configuration leads to observable performance degradation, indicating that proper balancing between semantic drift, geometric regularization, and relational disruption is crucial for effective universal perturbation transfer. These results demonstrate that while the proposed loss components are complementary, their relative weighting plays a critical role in maximizing cross-model transferability.

Perturbation Budgets ϵ_v and ϵ_t. In our experiments, we adopt a standard visual perturbation budget $\epsilon_v = 12/255$, which is widely used in prior adversarial vision-language research. This value achieves a desirable balance between attack efficacy and imperceptibility — being large enough to encode transferable perturbation signals, while remaining imperceptible to human observers. As commonly reported, smaller values such as $4/255$ often result in poor generalization, and larger budgets like $16/255$ offer marginal improvements in attack success, while potentially compromising imperceptibility and increasing the risk of adversarial example detection.

For text perturbations, we follow the convention of restricting the adversary to a single token modification ($\epsilon_t = 1$), as substituting multiple tokens can lead to unnatural or semantically implausible sentences. Empirically, we find that single-token replacement is sufficient to induce strong cross-modal interference while preserving fluency and stealthiness. Therefore, we fix both ϵ_v and ϵ_t throughout all our experiments.

Table 7: Ablation study on different loss components. We report TR (Text Retrieval) and IR (Image Retrieval) performance for CLIPcnn model under different ablation settings.

Cos	CS	RPL	TR	IR
✓	✓	✓	**54.4**	**75.77**
✓	✗	✗	52.85	74.3
✗	✓	✗	43.13	62.01
✗	✗	✓	51.42	68.42
✓	✓	✗	52.33	73.53
✓	✗	✓	54.15	73.42
✗	✓	✓	51.81	67.65

Table 8: Ablation study on the hyperparameter α. We evaluate the performance of CLIPcnn model on TR (Text Retrieval) and IR (Image Retrieval) under different α values.

α	TR	IR
0	51.81	67.65
0.1	53.63	69.71
1	**54.4**	**75.77**
10	51.04	73.88

Table 9: Ablation study on the hyperparameter β. We evaluate the performance of CLIPcnn model on TR (Text Retrieval) and IR (Image Retrieval) under different β values.

β	TR	IR
0.1	52.59	74.65
0.3	**54.4**	**75.77**
0.5	54.15	75.39
0.7	51.94	74.93
1	53.37	74.96

Table 10: Ablation study on the hyperparameter γ. We evaluate the performance of CLIPcnn model on TR (Text Retrieval) and IR (Image Retrieval) under different γ values.

γ	TR	IR
0	52.85	74.3
1	**54.4**	**75.77**

F More Experimental Results

Table 11: Attack success rates (%) on image-text retrieval tasks under R@5, comparing our method with CPGC and GAP on the Flickr30K dataset.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Flickr30k R@5	ALBEF	GAP	55.71	73.86	8.01	10.54	1.2	4.84	4.46	15.24	8.28	20.27	5.33	10.77
		CPGC	**83.67**	**80.02**	**41.84**	42.18	6.9	17.19	18.34	41.03	26.22	49.42	24.25	34.59
		ours	73.25	74.25	39.74	**46.73**	**9.2**	**27.81**	**19.07**	**45.27**	**26.96**	**52.96**	**33.5**	**46.33**
	TCL	GAP	17.64	20.09	77.89	74.53	0.9	4.2	4.25	15.48	8.6	20.25	8.65	13.16
		CPGC	29.76	35.62	**90.89**	**84.18**	3.2	13.65	20.93	42.06	25.27	49.1	16.5	30.32
		ours	**32.77**	**47.34**	78.08	76.52	**10.1**	**30.02**	**24.87**	**49.61**	**37.47**	**57.06**	**30.18**	**50.83**
	X-VLM	GAP	6.21	7.45	4.9	7.96	81.6	77.23	6.11	18.33	17.41	28.35	5.03	8.61
		CPGC	7.62	25.1	8.71	26.63	89.2	**85.84**	19.38	42.48	30.89	50.7	13.68	29
		ours	**9.92**	**26.5**	**14.71**	**29.04**	**91.7**	84.1	**20.52**	**44.4**	**31.74**	**51.12**	**16.3**	**30.73**
	CLIPvit	GAP	2.81	6.86	4.2	8	1.7	6.1	75.64	82.56	24.2	37.68	4.33	9.97
		CPGC	6.31	17.51	8.01	19.65	4.3	15.1	76.89	**85.2**	39.6	54.68	9.15	23.23
		ours	**8.92**	**18.54**	**12.11**	**21.33**	**6.7**	**17.83**	**77.41**	84.38	**48.62**	**63.09**	**13.18**	**27.34**
	CLIPcnn	GAP	2.81	7.41	5.81	9.27	2.4	7.01	9.33	19.64	57.63	69.33	4.02	9.76
		CPGC	3.01	19.09	5.11	**22.7**	3.3	**23.07**	17.41	41.17	61.57	74.32	6.74	25.16
		ours	**8.92**	**19.42**	**8.81**	19.55	**9.4**	19.62	**27.56**	**46.11**	**80.79**	**85.14**	**12.98**	**28.08**
	BLIP	GAP	3.41	7.01	3.7	7.45	1	4.4	4.46	14.43	6.79	18.67	39.13	68.02
		CPGC	14.43	21.67	13.91	21.59	**5.4**	**14.54**	18.03	36.26	23.89	44.79	**59.26**	**74.82**
		ours	**14.53**	**30.14**	**17.62**	**26.67**	4.4	12.77	**20.93**	**39.54**	**26.65**	**46.29**	44.06	59.79

Table 12: Attack success rates (%) on image-text retrieval tasks under R@5, comparing our method with CPGC and GAP on the MSCOCO dataset.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
MSCOCO R@5	ALBEF	GAP	74.43	78.62	37.99	30.08	5.56	7.19	14.26	17.11	15.58	21.62	23.73	23.26
		CPGC	**93.36**	**91.56**	70.76	62.31	19.97	30.46	41.58	51.23	44.14	55.98	41.08	49.22
		ours	87.93	89.86	65.59	58.54	**27.87**	**35.17**	**53.27**	**58.95**	**57.66**	**64.93**	**50.2**	**54.04**
	TCL	GAP	41.48	32.59	92.54	87.81	6.46	8.08	16.09	18.47	17.98	24.3	29.9	28.64
		CPGC	60.62	56.21	**94.89**	**90.33**	22.08	30.38	53.14	64.98	58.85	70.77	45.28	53.55
		ours	**65.92**	**62.45**	94.39	89.91	**28.16**	**37.58**	51.83	57.56	55.04	63.11	**61.31**	**64.48**
	X-VLM	GAP	12.29	11.64	13.43	10.99	90.8	83.05	20.02	23.4	37.72	40.09	12.64	12.04
		CPGC	31.59	48.69	32.1	48.11	**96.7**	**91.66**	49.53	60.82	59.83	69.59	37.4	52.5
		ours	**38.06**	**53.1**	**43.42**	**54.09**	94.79	89.56	**50.68**	**62.23**	**59.94**	**69.62**	**43.29**	**55.78**
	CLIPvit	GAP	18.78	17.46	20.38	17.37	16.81	15.15	95.21	93.04	62.77	62.62	18.48	19.48
		CPGC	25.69	35.95	24.69	33.14	21.37	**31.38**	96.7	**96.49**	70.76	77.86	28.72	42.01
		ours	**26.73**	34.11	**26.02**	32.67	**22.43**	30.72	**96.86**	95.82	**76.58**	**80.85**	**33.77**	**46.66**
	CLIPcnn	GAP	13.54	14.09	14.42	14.14	11.02	12.03	25.27	24.98	88.67	88.83	12.8	14.98
		CPGC	16.25	30.07	20.96	**34.15**	16.58	30.23	48.04	56.66	91.54	88.96	21.37	38.85
		ours	**32.95**	**36.01**	**27.71**	32.23	**28.83**	**34.1**	**60.73**	**68.95**	**93.23**	**94.54**	**37.9**	**47.56**
	BLIP	GAP	23.62	24.22	22.96	18.43	9.93	9.75	19.2	22.24	24.99	30.19	62.75	66.88
		CPGC	42.56	43.73	41.72	41.8	31.05	35.63	44.37	57.9	54.47	66.01	**81.71**	**81.91**
		ours	**48.29**	**49.73**	**53.39**	**46.76**	24.61	29.7	**53.04**	**62.44**	**60.54**	**68.87**	80.42	77.36

Table 13: Attack success rates (%) on image-text retrieval tasks under R@10, comparing our method with CPGC and GAP on the Flickr30k dataset.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Flickr30k R@10	ALBEF	GAP	51.6	71.17	5.8	6.65	0.6	2.7	1.42	9.82	4.19	13.81	3.71	7.01
		CPGC	**80.5**	**75.17**	34.8	34.28	4.2	11.72	9.83	31.14	16.87	39.4	18.76	27.08
		ours	68.1	69.46	33.2	**37.53**	**4.5**	**21.33**	**11.75**	**35.69**	**18.3**	**42.41**	**25.28**	**39.33**
	TCL	GAP	14.9	14.29	73.26	70.49	0.6	2.22	2.13	9.73	4.29	13.45	5.92	9.51
		CPGC	24.2	27.32	**89.2**	**80.73**	2.1	9.33	12.77	32.4	16.97	38.63	12.54	24.1
		ours	**24.2**	**39.58**	72.2	71.34	**6.1**	**23.27**	**14.49**	**38.96**	**27.2**	**47.25**	**22.57**	**43.16**
	X-VLM	GAP	4.1	4.81	2.7	4.51	76.5	72.58	3.65	11.63	10.84	18.91	3.01	5.29
		CPGC	4.1	17.79	4.6	19.27	86.3	82.94	11.14	31.49	21.06	40.3	7.32	21.81
		ours	**6.4**	**18.99**	**9.6**	**20.95**	**88.9**	79.97	**12.87**	**34.63**	**22.8**	**40.95**	**11.43**	**23.41**
	CLIPvit	GAP	2.1	4.04	2.6	4.81	1	3.79	63.29	77.83	17.89	28.11	2.31	6.12
		CPGC	4.2	11.45	4.6	13.08	2.8	10.15	67.98	**79.46**	29.75	45.56	5.52	17.23
		ours	**5.3**	**12.96**	**7.4**	**14.64**	**4.5**	**12.32**	**71.33**	78.92	**38.24**	**55.12**	**8.43**	**20.62**
	CLIPcnn	GAP	2.3	4.23	3.3	5.56	1.4	4.41	4.56	12.35	49.86	62.17	2.21	6.63
		CPGC	2.5	**13.56**	3.7	**16.87**	1.7	**17.38**	9.93	32.4	53.27	66.34	3.91	19.62
		ours	**5.7**	12.74	**5.0**	13.27	**6.7**	13.6	**18.14**	**35.63**	**75.66**	**80.84**	**8.83**	**21.37**
	BLIP	GAP	2	3.98	1.5	4.17	0.2	2.24	1.93	8.38	3.68	12.48	36.81	67.22
		CPGC	**11.2**	15.49	9.1	14.14	2.8	10.19	9.83	27.46	14.83	34.43	**53.46**	**72**
		ours	9.9	**26.06**	**13.1**	**18.97**	2.6	8.75	**12.36**	**29.18**	**17.59**	**36.27**	39.52	55.08

Table 14: Attack success rates (%) on image-text retrieval tasks under R@10, comparing our method with CPGC and GAP on the MSCOCO dataset.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIPvit		CLIPcnn		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
MSCOCO R@10	ALBEF	GAP	69.78	76.24	32.04	23.7	3.16	4.91	10.07	13.16	12.55	16.6	19.1	19.96
		CPGC	**91.58**	**89.62**	64.5	55.3	13.81	23.34	33.3	43.75	35.82	48.38	33.77	43.11
		ours	85.12	87.88	57.57	51.39	**20.68**	**27.45**	**45.14**	**51.55**	**49.09**	**57.57**	**43.34**	**47.72**
	TCL	GAP	34.36	25.76	90.65	85.27	4.01	5.44	11.77	14.88	13.66	18.89	24.91	24.67
		CPGC	52.59	49.09	**93.63**	**88.53**	15.04	23.25	44.22	58.02	50.16	63.95	37.77	47.26
		ours	**58.5**	**55.14**	92.78	87.31	**20.44**	**29.85**	43.7	50.28	**46.86**	**55.24**	**54.04**	**58.68**
	X-VLM	GAP	7.66	7.65	8.07	7.27	88.3	79.85	15.63	18.77	31.79	33.54	8.24	9.37
		CPGC	23.01	40.39	23.15	40.07	94.97	88.95	40.24	53.74	52	62.7	30.43	45.67
		ours	**28.9**	**44.76**	**34.03**	**46.0**	93.52	86.67	**41.9**	**54.85**	51.3	62.32	**35.08**	**48.8**
	CLIPvit	GAP	13.13	12.39	14.05	12.13	10.68	10.89	93.72	91.51	57.39	56.47	13.35	15.64
		CPGC	17.87	28.52	17.48	26.09	14	24.67	95.55	**95.31**	64.04	72.75	22.05	35.65
		ours	**18.81**	26.64	**18.45**	25.21	**14.71**	**24.07**	95.38	94.24	**71.39**	**76.25**	**26.19**	**39.79**
	CLIPcnn	GAP	9.02	10.08	9.06	9.89	6.97	8.25	18.78	20.11	87.6	83.92	8.53	11.91
		CPGC	10.68	22.98	14.19	**27.21**	10.62	**23.96**	39.89	49.62	88.74	85.18	15.97	33.25
		ours	**24.37**	**28.31**	**19.82**	25.01	**20.99**	26.91	**52.63**	**62.39**	**91.39**	**93.1**	**30.87**	**40.99**
	BLIP	GAP	17.76	18.98	16.32	13.13	6.21	6.44	13.97	17.39	19.73	24.37	57.99	65.49
		CPGC	33.64	36.14	32.15	33.8	**22.64**	**28.52**	36.07	50.3	47	59.07	**78.39**	**78.98**
		ours	**40.39**	**44.46**	**44.29**	**39.12**	17.52	23.2	**44.29**	**55.72**	**53.68**	**62.19**	76.69	73.78

G Visualization

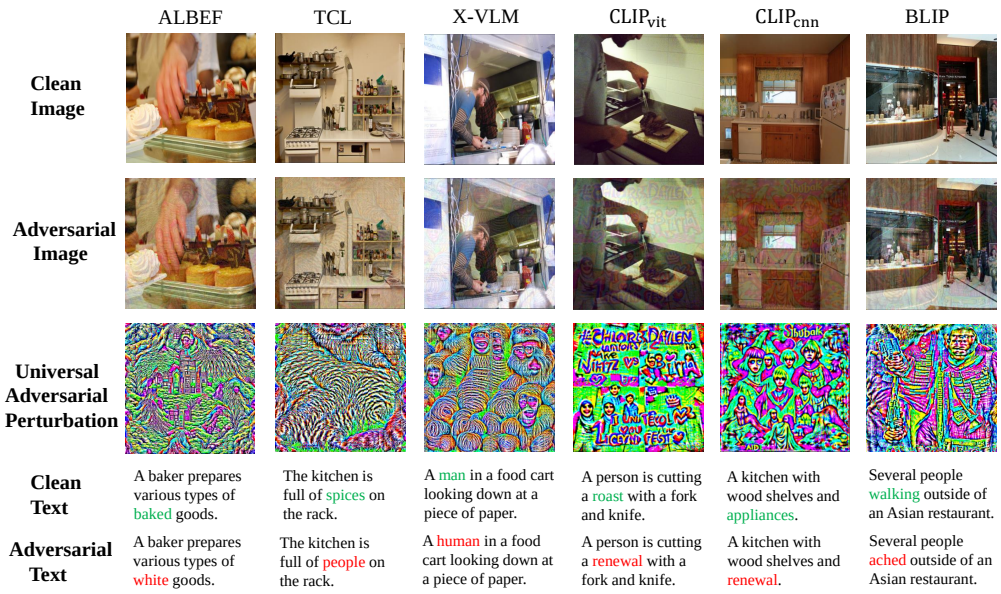

Figure 4: Visualization results of adversarial perturbations on multiple VLP models.

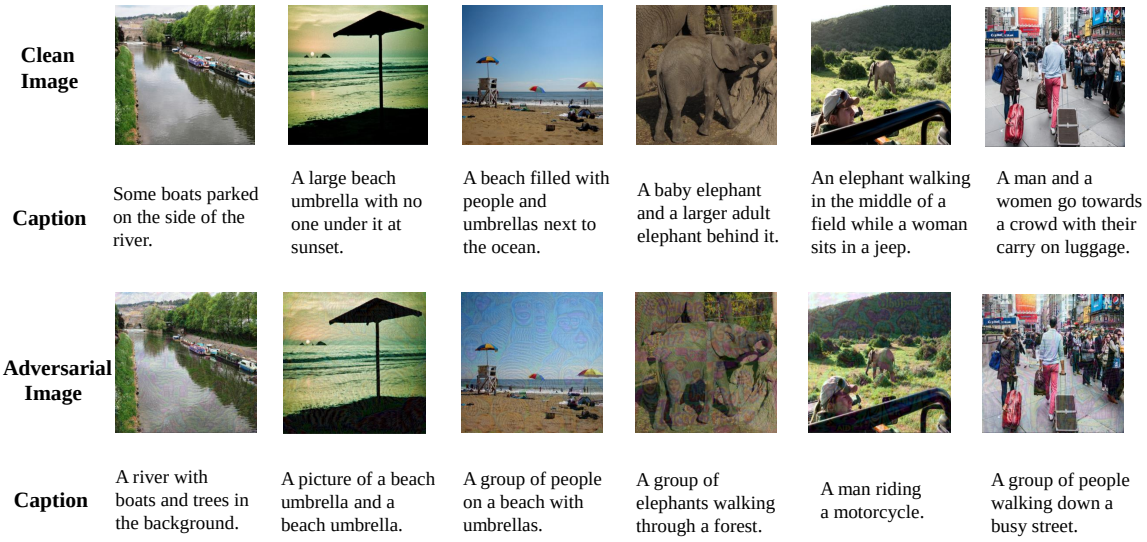

Figure 5: Visualization of image captioning (IC) results. The adversarial perturbations significantly alter the generated descriptions across models, highlighting the impact of our universal attack.