DiMiG: Diverse Mismatch Guidance for Transferable Attacks on Vision–Language Models

Tianjie Ni[1], Ziyu Liu[1], Shuyuan Zhang[1], Yuxue Chen[1], Wenbo Fang[2], Junjiang He[1*]
[1] Sichuan University, China
[2] Southwest Minzu University, China
Email: netjay@stu.scu.edu.cn, ziyuliu@stu.scu.edu.cn, zhangshuyuan1@stu.scu.edu.cn,
chenyuxue299@stu.scu.edu.cn, fangwenbo@swun.edu.cn, hejunjiang@scu.edu.cn

***Abstract*—With the growing deployment of Vision-Language Pre-training (VLP) models, understanding their vulnerability to adversarial perturbations becomes increasingly important. Many multimodal adversarial attacks are effective in white-box settings yet degrade markedly under black-box transfer. Furthermore, many methods perturb both images and text simultaneously, while in many real-world scenarios, the text input is fixed or beyond the attacker's control. Based on this, we propose *DiMiG*, a transferable attack under an image-only threat model. First, we design a sample-conditional generator that generates adversarial image perturbations with only one forward pass. Then, a set of mismatched texts is mined for each image in the shared embedding space and used to provide diverse supervisory signals, thereby improving transferability. Finally, we devise a joint objective that pulls the adversarial image representation toward the mined mismatches and suppresses the matched text through weighted competition, while also enlarging the distance between adversarial and clean image representations in the same embedding space. Extensive experiments demonstrate that DiMiG achieves strong intermodel transferability while remaining highly efficient for adversarial evaluation. Under the image-only setting, with ALBEF as the surrogate, DiMiG improves the mean transferred ASR@1 across five victim models on Flickr30K and MSCOCO by 36.8 percentage points compared with representative prior methods. Our code is available at https://github.com/maskisbest/DiMiG.**

***Index Terms*—multimodal adversarial attack, generation-based adversarial attacks, transferable attack, vision–language models, black-box**

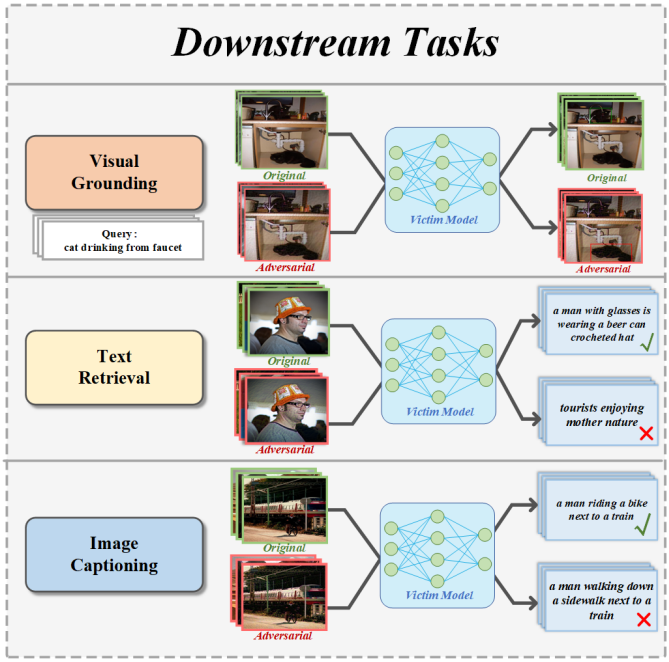


Fig. 1: Adversarial attack examples on different downstream tasks of a target model.

I. Introduction

Vision–Language pre-trained (VLP) models have attracted considerable attention in multimodal understanding and generation, supporting a wide range of downstream applications such as image-text retrieval, image captioning, and visual grounding. Most VLP architectures couple the two modalities by mapping images and texts into a shared representation space, typically through contrastive alignment objectives [1]–[3] and fusion-based interaction modules [4]–[6]. More recent generative frameworks further broaden this paradigm by unifying understanding and generation [7], [8]. However, the shared embedding space constitutes a potential attack surface that adversarial perturbations can leverage to disrupt multimodal alignment. Fig. 1 provides an example of adversarial attack.

Existing multimodal attacks against pre-trained (VLP) models are largely dominated by iterative, gradient-based optimization. The core idea is to conduct iterative gradient-based optimization on a surrogate model and progressively update the perturbations to disrupt cross-modal alignment. They typically aim to reduce the similarity of matched pairs, thereby misleading downstream tasks [9]–[13]. Since the perturbation is refined through per-sample iterations, this line of methods is often strong in white-box evaluations where gradients are accessible. Meanwhile, mainstream research adopts a more permissive threat model that allows perturbing both the image and the text, and use the adversarial change in one modality to cooperatively guide the update of the other [9]–[11].

However, in practical scenarios, attackers often face a *black-box* setting, in which the attacker may have full access to a surrogate model but cannot obtain gradients from the victim model. Under this setting, gradient-based attacks may suffer from a notable performance drop due to surrogate-specific overfitting and limited intermodel generalization [14], [15].

Moreover, the common assumption of jointly perturbing both modalities does not always hold in practice. Many systems (e.g., captioning and retrieval) take a single-modality input at inference time, and in visual grounding the query text is frequently provided by users and thus remains fixed.

To better match realistic constraints and enable efficient transferability assessment, we adopt an *image-only threat model* in a black-box setting, where the attacker perturbs only the image while keeping the text unchanged and evaluates the resulting adversarial images on victim models without accessing their gradients. Under this setting, a key challenge is to craft perturbations that are not overly dependent on surrogate-specific gradients and can still generalize to diverse victim architectures.

Two observations motivate our design. First, prior evidence suggests that increasing the *diversity* of training signals can improve intermodel generalization and transferability in adversarial settings [14]. Second, a text-conditioned perturbation generator naturally fits this image-only yet multimodal setting. It also reduces the dependence on gradients of the surrogate model, thereby improving transferability [16]. Motivated by these observations, we introduce diverse supervision to train a generator, where multiple texts guide the perturbation learning for improved transfer across VLP architectures.

Inspired by the above considerations, we propose **DiMiG**, a transferable attack under an *image-only* threat model in black-box settings. **First**, we train a lightweight, sample-conditioned generator that produces adversarial image perturbations in a single forward pass. **Then**, a set of mismatched texts is mined for each image in the shared embedding space and used to provide diverse supervisory signals, thereby improving transferability. **Finally**, we design a sample-competitive contrastive objective that leverages weighted competition among these mismatches to suppress the matched text and guide the adversarial image representation accordingly. After training, the generator enables efficient black-box attack across diverse victim architectures and downstream tasks. Our main contributions are summarized as follows:

- We propose a method to generate diverse supervision signals to facilitate transferable attack by mining mismatched texts in a shared embedding space.
- A sample-competitive contrastive objective is proposed to leverage these mismatches and improve intermodel transferability under the image-only constraint.
- Extensive experiments on image–text retrieval, image captioning, and visual grounding demonstrate consistently improved transferability and strong attack effectiveness under the same perturbation budget.

II. Related Work

A. Gradient-based adversarial attacks on VLP

Most existing multimodal attacks perform iterative, gradient-based optimization. They often perturb both image and text modalities to disrupt alignment more effectively. Co-Attack points out that independently attacking each modality may cause a "cancellation effect" and thus proposes cooperative optimization to drive both perturbations toward consistent directions in the embedding space [9]. To improve transferability, SGA introduces set-level guidance by constructing diverse aligned image–text pairs and using the other modality as supervision to disturb cross-modal interactions [10]. TMM further leverages cross-attention to guide perturbation allocation onto modality-consistent features and regularizes modality-specific factors to enhance transfer across architectures [11]. MAA combines RScrop (resizing + sliding crop) with MGSD (multi-granularity similarity disruption) to intensify fine-grained detail perturbations, thereby disrupting image–text alignment and improving intermodel transferability [13]. While effective in white-box settings, these approaches can become less reliable under transferable attack in black-box setting, and many of them assume a more permissive threat model where both modalities are perturbable.

B. Generative adversarial attacks on VLP

Generative adversarial attacks train a perturbation generator to produce adversarial examples *efficiently* at inference time, which is particularly appealing for transferable attacks in black-box settings across multiple victim models. C-PGC [16] learns *universal* perturbations with cross-modal conditioning, aiming to craft an input-agnostic perturbation that generalizes across many samples. Accordingly, it is not designed to tailor perturbations to the fine-grained semantics of each individual image. AnyAttack [17] adopts a pretrain–finetune pipeline to learn a noise generator and demonstrates transferability to commercial vision–language models. Under its targeted setup, the conditioning signals used during training are primarily determined by the targeted objective and available interfaces, rather than explicitly exploiting paired image–text interactions for each sample. These differences in problem settings motivate our focus on sample-conditioned, text-guided perturbation generation for intermodel transfer under the image-only constraint.

III. Proposed Attack

A. Problem Setup & Threat Model

Problem Setup. We consider three representative vision–language downstream tasks: image–text retrieval, image captioning, and visual grounding. Given an input image v and a text t, a vision–language pre-trained model F maps them into a shared embedding space via an image encoder f_I and a text encoder f_T, i.e., $e_v = f_I(v)$ and $e_t = f_T(t)$, and measures their alignment using a similarity function $sim(e_v, e_t)$. Retrieval returns a ranked list based on similarity, captioning produces a text sequence conditioned on v, while grounding predicts a localized region conditioned on v and the query text. In our setting, the attacker perturbs only the image modality while keeping the text input unchanged, and crafts an adversarial image $v^{adv} = \mathrm{clip}(v + \delta)$ under an ℓ_∞ budget $\|\delta\|_\infty \leq \epsilon$.

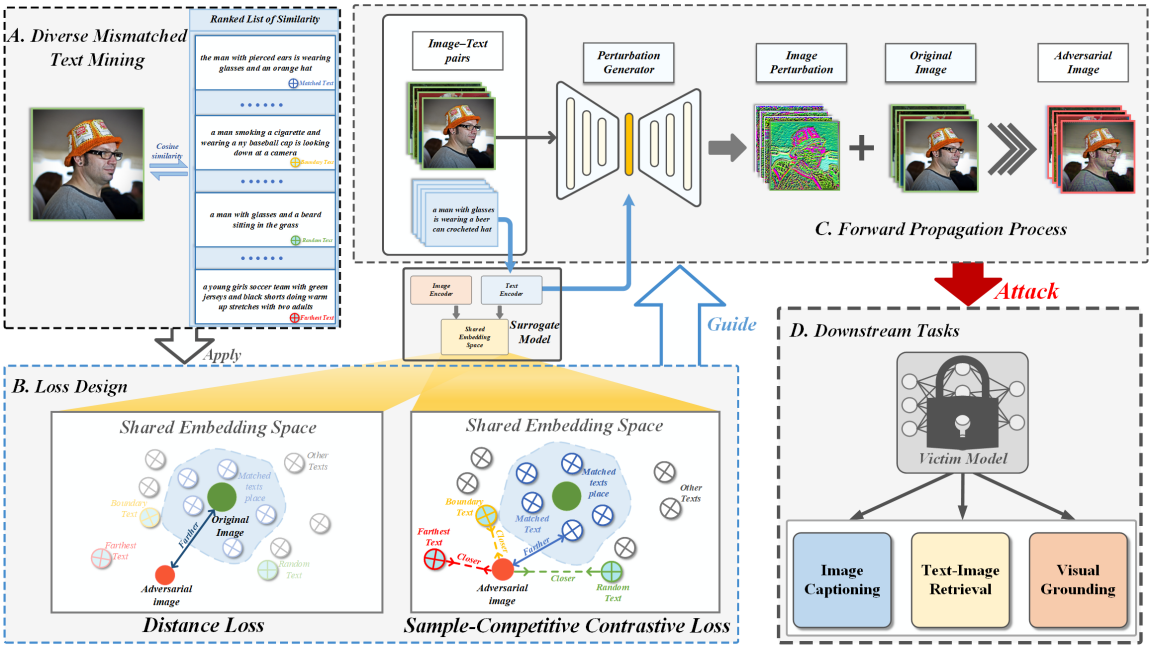


Fig. 2: Overview of the proposed image-only transferable attack. We mine farthest/boundary/random mismatched texts via the surrogate model to guide training of a sample-conditioned generator, which produces adversarial perturbations in a single forward pass. The generated adversarial images are transferred to victim models for downstream tasks.

B. Overview of the Proposed Attack

Our method has three key steps. First, we mine a set of three mismatched texts per image based on the surrogate model. Second, we design a sample-competitive contrastive loss that suppresses the matched text via competition within the candidate pool. Third, we train a lightweight sample-conditioned perturbation generator, enabling one-pass adversarial image generation at inference time. Fig. 2 illustrates the overall pipeline.

Threat Model. We adopt a black-box setting where the attacker has full access to a surrogate (source) model F_s but cannot access gradients of the target (victim) model F_t. The goal is to induce task failures on F_t with imperceptible image perturbations, which can be written as

$$\mathcal{T}\big(F_t(v^{adv}, t)\big) \neq \mathcal{T}(F_t(v, t)), \tag{1}$$

where $\mathcal{T}(\cdot)$ denotes the task-specific decision or output. For captioning, t can be omitted.

C. Diverse Mismatched Text Mining

Overview. Let $\mathcal{D} = \{(v_i, \mathcal{T}_i)\}_{i=1}^{N}$ denote the training set, where v_i is an image and $\mathcal{T}_i$ is its paired text set. Let $\mathcal{T} = \bigcup_{i=1}^{N} \mathcal{T}_i$ be the entire text corpus in $\mathcal{D}$. For each image v_i, we mine three mismatched texts from the non-paired pool $\mathcal{T} \setminus \mathcal{T}_i$ in an offline manner and store $(t_i^{far}, t_i^{bd}, t_i^{rand})$ for training. Using the surrogate model F_s, we extract ℓ_2-normalized embeddings $e_{v_i} = f_I(v_i)$ and $e_t = f_T(t)$, and measure cross-modal similarity by

$$\mathrm{sim}(e_{v_i}, e_t) = e_{v_i}^{\top} e_t. \tag{2}$$

Overall, t_i^{far}, t_i^{bd}, and t_i^{rand} respectively provide global farthest, near-boundary, and diverse signals for the subsequent competitive objective.

Farthest mismatched text. The *Farthest* mismatched text serves as a global "far anchor" to encourage a cross-modal displacement. We define it as the least similar non-paired text in the shared embedding space as

$$t_i^{far} = \arg\min_{t \in \mathcal{T} \setminus \mathcal{T}_i} \mathrm{sim}(e_{v_i}, e_t). \tag{3}$$

Boundary mismatched text. The *boundary* mismatched text is designed to capture near-boundary *confusable* cases, i.e., non-paired texts that are highly similar to v_i in the shared embedding space. We first form a candidate set $\mathcal{N}_k(i) \subset (\mathcal{T} \setminus \mathcal{T}_i)$ by selecting the top-k most similar non-paired texts to v_i. Then, we choose t_i^{bd} from $\mathcal{N}_k(i)$ by enforcing a mild directional constraint with respect to the far anchor. Specifically, we define the displacement direction

$$\mathbf{d}(t) = e_t - e_{v_i}. \tag{4}$$

We select a candidate whose direction is moderately aligned with $\mathbf{d}(t_i^{far})$ while avoiding excessive collinearity:

$$\theta_1 < \angle\big(\mathbf{d}(t), \mathbf{d}(t_i^{far})\big) < \theta_2, \quad t \in \mathcal{N}_k(i). \tag{5}$$

This constraint is used to avoid conflicts with the far-anchor direction and to maintain difference. The values of k and (θ_1, θ_2) are specified in the implementation details. If no candidate in $\mathcal{N}_k(i)$ satisfies the constraint, we return to the most similar non-paired text as in t_i^{bd}.

Random mismatched text. Finally, we sample a *random* mismatched text t_i^{rand} from the remaining non-paired text

pool, excluding the already selected t_i^{far} and t_i^{bd}, to introduce diversity.

D. Loss Function Design

Candidate pool. For each training image v_i, we first sample one paired text $t_i^- \in \mathcal{T}_i$ as the matched description. Together with the three mined mismatched texts $(t_i^{far}, t_i^{bd}, t_i^{rand})$ from Sec. III-C, we form a candidate pool

$$\mathcal{C}_i = \{t_i^-,\, t_i^{far},\, t_i^{bd},\, t_i^{rand}\}. \tag{6}$$

Sample-Competitive Contrastive Loss. Following CPGC [16], we adopt an InfoNCE-type objective with sample competition on the pool $\mathcal{C}_i$. Given an adversarial image embedding $e_{v_i^{adv}}$, we define the temperature-scaled similarity score by

$$s_i(t) = \mathrm{sim}(e_{v_i^{adv}}, e_t)/\tau. \tag{7}$$

We treat the three mismatches as a "positive" set (attack targets) and aggregate their weighted responses into

$$\begin{aligned} A_i^+ = \; & \exp\big(w_{far}\, s_i(t_i^{far})\big) + \exp\big(w_{bd}\, s_i(t_i^{bd})\big) \\ & + \exp\big(w_{rand}\, s_i(t_i^{rand})\big). \end{aligned} \tag{8}$$

while the matched text t_i^- is regarded as a "negative" term to be suppressed as

$$A_i^- = \exp\big(s_i(t_i^-)\big). \tag{9}$$

Let B denote the batch size. We then formulate the sample-competitive contrastive loss in the form of

$$\mathcal{L}_{\mathrm{SCC}} = -\sum_{i=1}^{B} \log \frac{A_i^+}{A_i^+ + A_i^-}, \tag{10}$$

$\mathcal{L}_{\mathrm{SCC}}$ exhibits a two-level competition mechanism. At the pool level, A_i^+ and A_i^- share the normalizer $A_i^+ + A_i^-$, forcing the mismatched set and the matched term to compete for probability mass and thus explicitly suppressing matched alignment. At the within-set level, A_i^+ is constructed as a weighted sum of exponential similarities, which induces a soft competition among the three mismatched types so that the most effective one receives larger gradients and adaptively dominates the optimization. Overall, this design upgrades the supervision from aligning with a single fixed mismatch to competitively suppressing the matched text against a set of mismatches, enhancing intermodel transferability.

Distance Loss. In addition to the sample-competitive contrastive loss above, we introduce a uni-modal term in the shared vision–language embedding space. This term encourages the adversarial image representation to deviate from the original one. Specifically, let $e_{v_i} = f_I(v_i)$ and $e_{v_i^{adv}} = f_I(v_i^{adv})$, where f_I is the image encoder of the surrogate VLP model. We define

$$\mathcal{L}_{\mathrm{DIS}} = -\sum_{i=1}^{B} \left\| e_{v_i^{adv}} - e_{v_i} \right\|_2^2, \tag{11}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{SCC}} + \lambda\, \mathcal{L}_{\mathrm{DIS}}. \tag{12}$$

E. Sample-conditioned Image Perturbation Generator

We employ a sample-conditioned generator $G(\cdot;\omega)$ to produce an image perturbation:

$$\delta_i = G(v_i, e_t; \omega), \qquad v_i^{adv} = \mathrm{clip}(v_i + \delta_i). \tag{13}$$

Architecturally, G adopts an encoder–cross-attention–decoder design. Three convolutional layers first encode v_i, a cross-attention module injects the textual condition at the intermediate feature level, and three deconvolutional layers then decode the modulated features to δ_i. The text embedding is obtained from the surrogate text encoder, i.e., $e_t = f_T(t)$, and is used as the conditioning signal.

Optimization. We optimize ω by minimizing Eq. (12) under the ℓ_∞ constraint $\|\delta_i\|_\infty \le \epsilon$ (default $\epsilon = 12/255$). After training, G generates δ_i given (v_i, e_t) in a single forward pass.

IV. Experiments

In this section, we evaluate the proposed multimodal attack under a black-box setting on three representative vision–language downstream tasks: image–text retrieval, image captioning, and visual grounding. We present intermodel transferability results across diverse victim architectures under an image-only threat model, followed by ablations on the weighting factors $(w_{\text{far}}, w_{\text{bd}}, w_{\text{rand}})$, the perturbation budget ϵ, and the trade-off coefficient λ. Since our approach follows a generator-based paradigm, adversarial images can be produced in one forward pass after training, enabling efficient evaluation at scale. We further provide brief analyses on the effects of diverse mismatched-text supervision.

A. Experimental Setup

1) Tasks and Datasets:

Tasks. We consider three representative vision–language downstream tasks

- **Image–text retrieval** [18] ranks a gallery of texts (or images) for a given query image (or text) according to cross-modal similarity in a shared embedding space. An attack is successful if it significantly degrades the retrieval ranking of the correct match.
- **Visual grounding** [19] localizes an image region that corresponds to a given textual query. An attack is successful if the predicted region no longer overlaps with the ground-truth object region.
- **Image captioning** [20] generates a natural-language description conditioned on an input image. A successful attack induces semantically incorrect or irrelevant captions while keeping the perturbation imperceptible.

Datasets. For image–text retrieval, we use **Flickr30K** [21] and **MSCOCO** [22] as standard benchmarks, for image captioning we use **MSCOCO** [22], and for visual grounding we use **RefCOCO+** [23], following common practice in prior works [16].

TABLE I: DiMiG ASR@1 (%) on image–text retrieval (Flickr30K/MSCOCO). TR denotes image→text retrieval and IR denotes text→image retrieval. Gray-shaded cells indicate white-box results for reference.

Source	Dataset	Perturbed modality	Method	ALBEF		TCL		X-VLM		CLIP_{ViT}		CLIP_{CNN}		BLIP	
				TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
ALBEF	Flickr30K	multi-modal	Co-Attack	95.64	96.52	21.87	32.19	13.43	27.45	21.13	33.23	23.43	35.26	25.39	42.84
			SGA	97.11	96.72	43.95	48.83	26.54	39.11	33.83	43.57	34.39	46.68	39.65	53.65
			TMM	97.53	97.51	64.97	69.60	47.14	55.49	52.90	60.90	56.61	62.97	59.99	66.01
		image-only	C-PGC	86.74	86.3	61.49	67.76	14.43	27.48	23.28	39.3	33.42	48.25	31.55	36.73
			MAA	**100.00**	**100.00**	70.6	73.31	11.59	24.05	26.85	43.16	33.81	51.09	49.0	57.01
			Ours	98.77	98.10	**95.34**	**95.52**	**63.72**	**65.07**	**51.48**	**60.10**	**69.17**	**71.57**	**88.54**	**88.25**
	MSCOCO	multi-modal	Co-Attack	94.63	96.59	33.07	42.11	20.43	33.35	39.85	48.05	43.47	52.11	44.31	55.61
			SGA	96.19	97.13	54.04	58.82	32.70	44.89	51.76	60.25	54.82	62.81	55.13	64.72
			TMM	96.79	97.73	70.19	74.02	52.97	58.23	68.37	75.34	70.97	76.88	71.29	74.21
		image-only	C-PGC	89.06	89.01	54.42	44.46	11.07	11.86	12.22	14.92	19.0	20.14	19.97	19.63
			MAA	**99.87**	**99.67**	48.91	51.45	6.93	10.85	13.52	13.22	23.8	31.14	30.81	37.07
			Ours	92.23	92.48	**77.51**	**76.13**	**56.94**	**56.27**	**59.03**	**67.97**	**72.79**	**78.43**	**73.81**	**75.50**
TCL	Flickr30K	multi-modal	Co-Attack	22.88	33.40	95.97	96.71	14.52	28.54	22.63	35.40	24.04	39.09	29.78	44.78
			SGA	47.69	56.04	97.37	96.95	30.09	43.67	36.71	47.32	38.20	49.32	43.30	55.76
			TMM	68.10	72.30	97.87	97.60	49.17	57.71	54.63	63.52	58.87	65.71	62.69	68.87
		image-only	C-PGC	37.72	42.15	92.55	88.29	10.57	21.36	21.92	38.6	32.51	47.13	20.72	31.11
			MAA	**84.58**	**73.05**	**100.00**	**100.00**	13.31	25.67	25.86	43.0	34.46	52.59	47.84	57.19
			Ours	58.17	58.10	95.13	93.91	**24.90**	**32.88**	**50.37**	**57.19**	**61.92**	**67.89**	**54.15**	**57.47**
	MSCOCO	multi-modal	Co-Attack	34.65	44.56	94.82	96.80	21.38	35.76	41.63	51.20	44.97	54.33	45.66	57.17
			SGA	58.29	68.80	96.73	97.32	37.04	47.81	54.62	62.36	57.63	64.95	56.85	65.15
			TMM	73.62	78.38	97.00	97.92	54.56	60.56	70.60	78.98	73.27	80.21	73.60	77.43
		image-only	C-PGC	42.11	35.76	90.52	86.34	13.63	15.57	18.15	19.56	29.6	29.61	24.15	21.9
			MAA	**63.87**	**68.63**	**100.00**	**100.00**	7.18	12.4	13.7	15.23	26.0	32.86	29.11	35.93
			Ours	50.41	54.58	89.03	85.62	**32.22**	**38.67**	**56.52**	**68.76**	**64.79**	**74.63**	**46.32**	**55.57**

2) Surrogate and Victim Models:

Surrogate models. We consider **ALBEF** or **TCL** as the surrogate (source) vision–language model F_s to train the proposed sample-conditioned perturbation generator. During training, the generator optimizes the sample-competitive objective using the surrogate encoders and does not access any information from target(victim) models.

Victim models. To evaluate transferability in a black-box setting, we test adversarial images on a diverse set of victim models, including **ALBEF**, **TCL**, **BLIP**, **CLIP (CNN backbone)**, **CLIP (ViT backbone)**, and **X-VLM**. This set covers different model families and architectures, enabling a comprehensive assessment of transferability.

3) Implementation Details:

We train the sample-conditioned generator using the surrogate model. Unless otherwise specified, we set the perturbation budget to $\epsilon = 12/255$, the trade-off coefficient in Eq. (12) to $\lambda = 0.1$, and the temperature in Eqs. (7) to $\tau = 0.1$. For boundary-text mining, we set $k = 64$ and $(\theta_1, \theta_2) = (15°, 90°)$ in Eq. (5). For the diverse mismatched texts, we set $(w_{\text{far}}, w_{\text{bd}}, w_{\text{rand}}) = (1.0, 0.5, 0.1)$. These hyper-parameters are kept fixed for all experiments and are varied only in the ablation studies. Task-specific evaluation metrics are introduced together with the corresponding main results.

B. Main Results of DiMiG

We demonstrate the main results on three representative vision–language downstream tasks, including image–text retrieval, image captioning, and visual grounding. Following the image-only threat model, adversarial perturbations are generated on a source model and directly evaluated on target models.

Image–Text Retrieval. We evaluate transferability on Flickr30K and MSCOCO, reporting ASR@1 for both retrieval directions, i.e., image→text (TR) and text→image (IR). Table I summarizes the results across diverse victim architectures (gray-shaded cells show white-box numbers for reference). Under the image-only setting, DiMiG yields strong and consistent transfer gains across five victim models. The mean transferred ASR@1 improves by 30.8 points on Flickr30K and 42.7 points on MSCOCO over a strong prior image-only baseline (MAA). The advantage is particularly evident on architecturally distant victims (e.g., X-VLM and CLIP/BLIP families), indicating better cross-architecture generalization. We attribute the improved transferability to (i) the diverse mismatched-text supervision that provides richer cross-modal guidance during training, and (ii) the sample-conditioned generator, which avoids per-instance iterative optimization at attack time and thus reduces surrogate-specific bias. We note that DiMiG does not always achieve the highest white-box ASR on the surrogate model. This reflects a trade-off between surrogate fitting and intermodel transferability. While methods such as MAA are better positioned to fit the surrogate's local decision boundary and saturate white-box performance, DiMiG is designed to learn more transferable perturbation di-

TABLE II: ASR@K (%) of cross-dataset attacks on image–text retrieval under Flickr30K → MSCOCO and MSCOCO → Flickr30K. The generator is trained on *ALBEF* as the surrogate model on the source dataset and transferred to attack victim models on the target dataset. TR denotes image→text retrieval and IR denotes text→image retrieval.

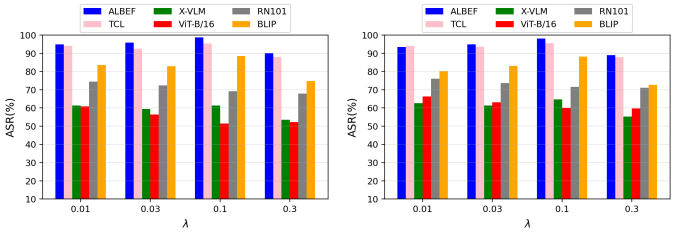
Setting	Metric	ALBEF		TCL		BLIP		X-VLM		$\text{CLIP}_{\text{ViT-B}}$		$\text{CLIP}_{\text{RN101}}$	
		TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Flickr30K → MSCOCO	ASR@1	77.35	73.37	60.18	60.29	59.53	58.54	40.15	33.95	44.44	47.13	59.40	55.71
	ASR@5	64.27	66.90	45.31	43.11	47.46	51.45	23.73	19.70	25.39	25.52	43.78	37.10
	ASR@10	58.57	61.15	38.40	37.32	42.53	46.83	19.55	16.18	19.11	19.07	35.83	29.84
MSCOCO → Flickr30K	ASR@1	58.99	61.78	66.15	72.40	47.42	51.41	25.71	33.35	51.11	59.17	67.23	68.94
	ASR@5	46.19	46.12	49.05	56.20	34.00	34.47	13.80	15.34	27.77	34.09	42.14	45.71
	ASR@10	40.60	40.48	42.10	49.93	28.59	27.51	10.80	11.24	21.18	25.52	33.44	35.98

TABLE III: Visual grounding on RefCOCO+. We provide $\text{Acc@IoU}_{0.5}$ on clean inputs (Baseline) and on adversarial inputs transferred from different surrogate models. "ALBEF (Source)" and "TCL (Source)" denote adversarial images generated by the corresponding surrogate-trained generator and evaluated on each victim model. Lower adversarial accuracy indicates a stronger attack effect.

Target	Baseline			ALBEF (Source)			TCL (Source)		
	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB
ALBEF	48.8	53.7	41.4	37.9	40.8	32.9	42.0	44.7	37.0
TCL	46.4	53.8	39.6	31.9	34.2	29.4	31.3	32.8	29.8
X-VLM	77.4	84.0	68.8	67.7	73.7	57.9	68.5	75.3	58.5

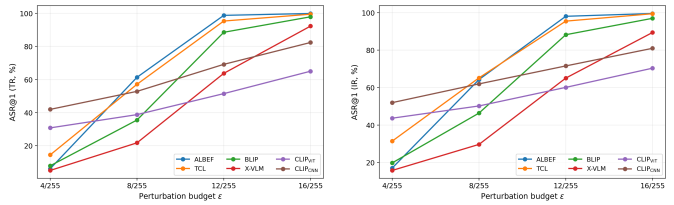
TABLE IV: Image captioning on MSCOCO evaluated on BLIP (victim). "Baseline" uses clean images, and each "Source" corresponds to adversarial images produced by a generator trained on the specified source model.

Source	B@4	METEOR	ROUGE_L	CIDEr	SPICE
Baseline	39.7	31.0	60.0	133.3	23.8
ALBEF	16.2	16.4	38.7	48.1	10.2
TCL	26.3	23.2	49.2	85.1	16.4



(a) TR (ASR@1). (b) IR (ASR@1).

Fig. 3: Sensitivity of TR/IR directions to λ (ALBEF surrogate and DiMiG ASR@1 on victim models in legend).



(a) TR (ASR@1). (b) IR (ASR@1).

Fig. 4: Effect of ϵ (ALBEF surrogate and DiMiG ASR@1 on victim models in legend).

rections under an image-only black-box setting. Consequently, DiMiG may sacrifice some white-box attack strength, but this trade-off leads to stronger transfer performance on unseen victim models.

Co-Attack, SGA, and TMM are **multi-modal** attacks evaluated through **joint image-text perturbations**. These methods co-optimize the perturbation in both modalities. And each modality's perturbation is optimized under the guidance of the other modality's adversarial counterpart. Adapting them to image-only variants would require modifying their core optimization methods. Such an image-only adaptation would not represent the methods described in the original papers. Therefore, we do not force these multi-modal attacks into the image-only setting. In contrast, C-PGC and MAA follow the same **image-only** setting and constitute the primary apples-to-apples comparisons in our experiments. We demonstrate their results to contextualize attack strength. Direct comparisons should be interpreted with this difference in mind, while C-PGC and DiMiG follow the same image-only setting.

We further evaluate cross-dataset generalization under Flickr30K → MSCOCO and MSCOCO → Flickr30K in Tab. II. Unless otherwise specified, the generator is trained with *ALBEF* as the source model on the source dataset and directly transferred to attack diverse victim models on the target dataset.

Visual Grounding. We also assess transferability on visual grounding, which localizes the referred region in an image given a natural-language query. Experiments are conducted on RefCOCO+, where we train the generator on ALBEF or TCL as the source and evaluate the crafted adversarial images on three target models (ALBEF/TCL/X-VLM). As summarized in Table III, adversarial images consistently reduce $\text{Acc@IoU}_{0.5}$

TABLE V: Ablation on the weighting factors $(w_{\text{far}}, w_{\text{bd}}, w_{\text{rand}})$ in Eq. (8) on Flickr30K (surrogate is ALBEF). We show ASR (%). AVG denotes the average ASR over all reported victim models.

$(w_{\text{far}}, w_{\text{bd}}, w_{\text{rand}})$	ALBEF		TCL		X-VLM		CLIP_{ViT}		CLIP_{CNN}		BLIP		AVG
	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	
$(1, 0.5, 0.1)$	**98.77**	**98.10**	**95.34**	**95.52**	**63.72**	**65.07**	51.48	60.10	69.17	71.57	**88.54**	**88.25**	**79.80**
$(0, 0.5, 0.1)$	93.32	91.01	91.10	91.03	53.56	56.50	56.65	62.65	72.15	73.56	75.39	74.28	74.77
$(1, 0, 0.1)$	92.39	91.95	91.72	91.10	53.35	56.44	56.16	63.09	71.24	73.84	78.76	79.67	74.97
$(1, 0.5, 0)$	91.88	89.63	91.93	91.87	49.59	53.68	**57.27**	**63.76**	72.28	73.91	74.87	72.57	73.60
$(1, 0, 0)$	96.09	95.31	92.55	94.31	55.79	58.83	54.43	61.40	**73.58**	**75.07**	81.60	82.61	76.13
$(1, 1, 1)$	94.86	93.57	92.96	93.10	58.03	60.57	57.76	63.64	73.19	75.25	82.23	78.84	76.50

across Val/TestA/TestB splits. Notably, the degradation is evident not only when the surrogate and target coincide, but also under intermodel transfer, indicating non-trivial black-box transferability of our perturbations to grounding models.

Image Captioning. We further evaluate transferability on image captioning using MSCOCO. Although captioning is image-only at inference, we condition the generator on a ground-truth caption from MSCOCO to craft v^{adv}, and feed only v^{adv} to the victim captioner. We consider BLIP as the victim captioner and report standard COCOEvalCap metrics (BLEU-4, METEOR, ROUGE_L, CIDEr, and SPICE) on clean images and on adversarial images transferred from different source models. In Table IV, "Baseline" corresponds to clean images, while each "Source" indicates the surrogate model used to train the generator for crafting adversarial images. As shown, adversarial perturbations lead to consistent degradation across all metrics compared with the clean baseline, demonstrating effective black-box transferability to captioning.

C. Ablation Study

We conduct ablation studies to analyze key design choices of our method. Unless otherwise stated, all ablations are performed under the black-box setting with the image-only perturbation budget $\epsilon = 12/255$.

Effect of weighting factors $(w_{\text{far}}, w_{\text{bd}}, w_{\text{rand}})$. We first study how the mismatched-text supervision contributes to transferability by varying the weighting factors in Eq. (8). Table V reports the results. Overall, assigning a dominant weight to the far anchor while keeping moderate boundary guidance and a small random component yields the best transfer performance, indicating that the three mismatches provide diverse signals.

Sensitivity to the trade-off coefficient λ. We vary λ in Eq. (12) to balance the sample-competitive contrastive term and the distance term. Unless otherwise specified, the generator is trained on ALBEF as the source, and DiMiG ASR@1 is evaluated on the target models shown in the legend for both retrieval directions (TR/IR). As shown in Fig. 3a–3b, the method is relatively robust when λ lies in a moderate range, while overly large λ may degrade transferability. We therefore choose $\lambda = 0.1$, which achieves the best performance (highest mean ASR@1 across victim models).

Impact of perturbation budget ϵ. We further evaluate different ℓ_∞ budgets ϵ. As shown in Fig. 4a and Fig. 4b, increasing ϵ generally improves ASR@1, with a rapid gain from $4/255$ to $12/255$, whereas the improvement beyond $12/255$ becomes marginal on most victims, suggesting diminishing returns. Following common practice under an ℓ_∞ constraint, we set $\epsilon = 12/255$ as a good trade-off between attack success and perceptual imperceptibility.

D. Analysis & Discussion

Transferability. Our transfer gain mainly comes from the diverse mismatched-text supervision, which provides varied guidance at different geometric scales in the shared embedding space. The farthest mismatched text provides a global displacement anchor. It encourages the adversarial representation to move away from the original matched text in a more model-agnostic manner, rather than being overly tied to a specific local neighborhood around the surrogate decision boundary. The boundary mismatched text offers near-boundary guidance. It constructs an ambiguous neighborhood where the match decision is easier to flip. In this region, a small representation shift is more likely to cross the decision boundary. This yields reliable local directions that generalize better across models. Finally, the random mismatched text simply adds diversity. It helps reduce early overfitting to a single mismatched sample and improves robustness when transferring to victim architectures. Together, these diverse signals form an effective supervision set for the sample-competitive objective, resulting in stronger adversarial transferability.

Generative paradigm and efficiency. Our attack trains a lightweight sample-conditioned generator on a surrogate model and produces perturbations in a single forward pass at test time. This approach avoids sample-by-sample iterative optimization in gradient-based attacks, thereby reducing inference time overhead when evaluating across multiple victim models and datasets. More importantly, the generator amortizes the optimization over the training distribution. This reduces reliance on gradients of the surrogate model and helps learn perturbation patterns that are more consistent and generalizable under the black-box setting.

Failure cases and limitations. Although our method improves transferability, a generator-based transfer attack runs under a more constrained setting than strict white-box iterative attacks. It does not explicitly optimize the perturbation for each test input using victim gradients, and thus may not precisely fit the surrogate model's local decision boundary for a specific

instance. As a result, there can be a performance gap compared with strong gradient-based white-box methods that directly exploit fine-grained gradient feedback.

V. Conclusion

We propose a transferable adversarial attack method in a black-box setting under an image-only threat model, where the attacker can only modify the input image and does not rely on gradients of the victim model. To improve intermodel transferability, the proposed sample-conditioned generative framework combines diverse mismatched-text supervision with a sample-competitive contrastive objective, providing stable and generalizable guidance in the shared embedding space. Extensive experiments on image–text retrieval, image captioning, and visual grounding demonstrate that our method achieves stronger and more stable transfer performance than representative baselines, while enabling efficient one-pass perturbation generation. A key limitation is that, compared with white-box iterative gradient attacks that explicitly optimize each test instance, the generator-based transferable attack is more constrained and may be less effective at tightly fitting the surrogate's local decision boundary for hard samples. Future work includes consolidating generators trained with multiple surrogate models into a single model via distillation or fusion, and developing more adaptive mismatched-text mining and weight assignment strategies to further improve robustness and the transfer upper bound.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62402330), in part by the Sichuan Provincial Science and Technology Department Regional Innovation Cooperation Key Project (Grant No. 2025YFHZ0265).

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[3] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[4] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[5] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," *arXiv preprint arXiv:2111.08276*, 2021.

[6] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 671–15 680.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[8] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[9] J. Zhang, Q. Yi, and J. Sang, "Towards adversarial attack on vision-language pre-training models," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5005–5013.

[10] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, and F. Zheng, "Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 102–111.

[11] H. Wang, K. Dong, Z. Zhu, H. Qin, A. Liu, X. Fang, J. Wang, and X. Liu, "Transferable multimodal attack on vision-language pre-training models," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 1722–1740.

[12] Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, and F. Ma, "Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 52 936–52 956, 2023.

[13] P.-F. Zhang, G. Bai, and Z. Huang, "Maa: Meticulous adversarial attack against vision-language pre-trained models," *arXiv preprint arXiv:2502.08079*, 2025.

[14] W. Wu, Z. Liu, Y. Chen, C. Su, D. Peng, and X. Wang, "Improving the transferability of adversarial examples by inverse knowledge distillation," *arXiv preprint arXiv:2502.17003*, 2025.

[15] Z. Qin, Y. Fan, Y. Liu, L. Shen, Y. Zhang, J. Wang, and B. Wu, "Boosting the transferability of adversarial attacks with reverse adversarial perturbation," *Advances in neural information processing systems*, vol. 35, pp. 29 845–29 858, 2022.

[16] H. Fang, J. Kong, W. Yu, B. Chen, J. Li, H. Wu, S.-T. Xia, and K. Xu, "One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4090–4100.

[17] J. Zhang, J. Ye, X. Ma, Y. Li, Y. Yang, Y. Chen, J. Sang, and D.-Y. Yeung, "Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 900–19 909.

[18] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[19] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, "Learning to compose and reason with language tree structures for visual grounding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 684–696, 2019.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[21] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[23] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European conference on computer vision*. Springer, 2016, pp. 69–85.