Multimodal Adversarial Defense for Vision-Language Models by Leveraging One-To-Many Relationships

Futa Waseda[1,3,†] Antonio Tejero-de-Pablos[2] Isao Echizen[1,3]
[1]The University of Tokyo [2]CyberAgent [3]National Institute of Informatics
[†]futa-waseda@g.ecc.u-tokyo.ac.jp

Abstract

Pre-trained vision-language (VL) models are highly vulnerable to adversarial attacks. However, existing defense methods primarily focus on image classification, overlooking two key aspects of VL tasks: multimodal attacks, where both image and text can be perturbed, and the one-to-many relationship of images and texts, where a single image can correspond to multiple textual descriptions and vice versa (1:N and N:1). This work is the first to explore defense strategies against multimodal attacks in VL tasks, whereas prior VL defense methods focus on vision robustness. We propose multimodal adversarial training (MAT), which incorporates adversarial perturbations in both image and text modalities during training, significantly outperforming existing unimodal defenses. Furthermore, we discover that MAT is limited by deterministic one-to-one (1:1) image-text pairs in VL training data. To address this, we conduct a comprehensive study on leveraging one-to-many relationships to enhance robustness, investigating diverse augmentation techniques. Our analysis shows that, for a more effective defense, augmented image-text pairs should be well-aligned, diverse, yet avoid distribution shift—conditions overlooked by prior research. This work pioneers defense strategies against multimodal attacks, providing insights for building robust VLMs from both optimization and data perspectives.

1. Introduction

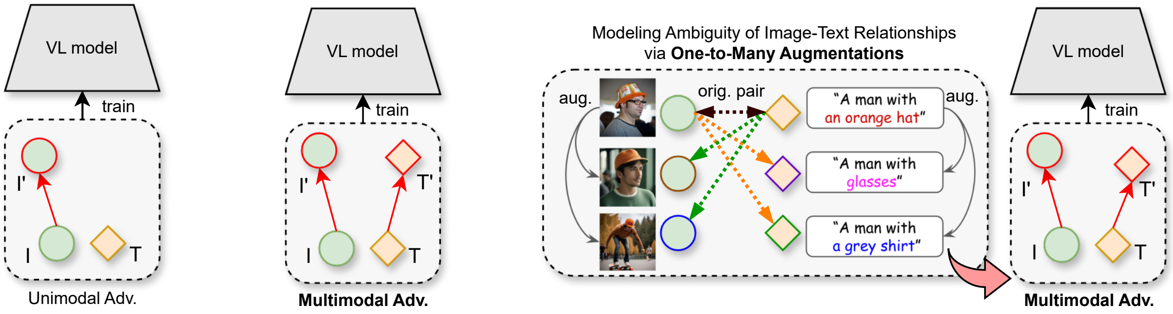
Vision-language (VL) tasks require modeling the relationships between images and texts; for example, image-text retrieval (ITR) retrieves the most relevant text given an image query, and vice versa. Recent VL models such as CLIP [25] achieve strong performance in these tasks; however, recent studies revealed that they are vulnerable to adversarial attacks [19, 37], which exploit nearly imperceptible input perturbations. Such vulnerabilities pose serious practical risks; for instance, attackers may manipulate images or descriptions to alter retrieval rankings, unfairly promoting or demoting entities in recommendation systems. As VL models are increasingly deployed, understanding and mitigating their adversarial vulnerabilities is urgent.

However, existing defense strategies for VL models [21, 27, 34] primarily focus on image attacks (e.g., robust zero-shot image classification), leaving other VL tasks unexplored. This is a considerable oversight for two reasons: (1) ***Multimodal manipulation***: Attackers can perturb both images and texts, requiring more complex defense strategies than image-only methods. (2) ***One-to-many (1:N)* cross-modal alignment**: Unlike classification with simple and deterministic labels (e.g., "a photo of a $\{class\}$"), VL tasks involve diverse, ambiguous sentences (e.g., "a man with glasses" vs. "a man wears an orange hat and glasses"), making robust image-text alignment more complicated. By overlooking these aspects, existing defenses are limited in their effectiveness beyond zero-shot image classification.

To address this gap, we pioneer defense strategies for VL models against multimodal attacks. Specifically, we study how to robustly fine-tune VL models for downstream VL tasks from both optimization and data perspectives, through extensive analysis.

First, we propose **M**ultimodal **A**dversarial **T**raining (MAT), which incorporates both image-text perturbations during training. Perturbing both modalities is non-trivial due to (1) the difficulty of simultaneous updates, (2) multiple objective choices for image-text attacks, and (3) increased computational complexity. Through extensive analysis, we designed MAT to be both effective and reasonably efficient, while also offering valuable insights for future work. MAT significantly enhances multimodal robustness, demonstrating the necessity of defense methods tailored for multimodal attacks, an aspect overlooked by vision-only unimodal strategies [21, 27].

Furthermore, we argue that MAT's performance is limited by the inadequate approximation of the real data distribution when using the deterministic (1:1) image-text pairs contained in the training data. To address this, we leverage the inherent one-to-many (1:N) relationships in VL data to enhance robustness. Inspired by works in cross-modal ambiguity modeling in ITR [14, 29], we explore augmentation



(a) Unimodal AT

(b) MAT (ours)

(c) MAT+ (ours)

Figure 1. **Comparison of adversarial training (AT) methods for robust VL models.** (a) **Unimodal AT**, such as TeCoA [21] and FARE [27], robustifies a single modality via unimodal adversarial examples (AEs). However, it overlooks two key aspects of VL tasks: *multimodal attacks*, where attackers perturb both modalities, and *one-to-many cross-modal alignment*, where an image has multiple valid descriptions, and vice versa. (b) **MAT** addresses multimodal attacks by generating multimodal AEs during AT. (c) **MAT+** further captures the inherent ambiguity in image-text relationships via one-to-many augmentations.

techniques to create diverse one-to-many (1:N) and many-to-one (N:1) image-text pairs. Our in-depth analysis reveals that augmentations are effective when pairs are well-aligned and diverse, without inducing distribution shift. Specifically, text augmentations outperform image augmentations, since the higher dimensionality of images makes distribution shift harder to avoid. Moreover, cross-modal augmentations (e.g., $image \rightarrow text$) outperform intra-modal ones (e.g., $text \rightarrow text$) by generating better-aligned pairs. These findings provide novel insights into multimodal robustness and complement the literature on unimodal adversarial training.

Our contributions are summarized as follows:

- **First defense strategy against multimodal attacks in VL models:** We show that existing image-only defense methods are suboptimal for robust VL tasks and pioneer research in this new direction. Specifically, we investigate strategies from both optimization and data perspective.
- **Proposed Multimodal Adversarial Training (MAT):** We designed MAT to be both effective and efficient, through extensive analysis. MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL data.
- **Leveraging one-to-many relationships for robust VL tasks:** We leverage one-to-many (1:N) image-text relationships via augmentations to enhance robustness, an aspect overlooked in unimodal adversarial training, which assumes a deterministic image-to-label mapping.

2. Related work

Adversarial attacks on vision-language models. Adversarial attacks on VL models are categorized into unimodal and multimodal. Unimodal attacks, such as gradient-based image attacks [20] and BERT-Attack for text [17], perturb a single modality to mislead the models. In contrast, multimodal attacks, which perturb both image and text modalities, are significantly more effective [11, 19, 33, 37]. However, developing defense strategies against multimodal attacks for VL tasks remains largely unexplored.

Adversarial defense for vision-language models. Existing defense strategies for VL models mainly focus on vision robustness, in which adversarial attacks perturb only the image. For example, Mao et al. [21] and Wang et al. [34] approached zero-shot image classification on CLIP by proposing robust fine-tuning methods, which leverage unimodal adversarial training schemes to improve robustness. Schlarmann et al. [27] also focused on image attacks only, and proposed a method for fine-tuning CLIP's vision encoder to improve robustness in several VL tasks (*i.e.*, image classification, image-text retrieval). Unlike the previous work, ours is the first to investigate adversarial defense strategies against multimodal attacks. Specifically, we propose a multimodal adversarial training strategy to enhance robustness against such attacks in VL models and tasks. Furthermore, we leverage the one-to-many (1:N) image-text relationships to further improve adversarial robustness.

Leveraging the one-to-many (1:N) nature of image-text. Recent works aimed at modeling the ambiguity between image and text pairs; a sentence may have multiple visual interpretations and an image may be described in various ways, however, typically only one pair is used as ground truth. Such deterministic 1:1 pairing is inconsistent with the natural 1:N relationships in the data. To address this, prior studies propose representing image-text samples as probabilistic embeddings [5, 6], incorporating neighboring samples in the triplet loss [30], and generating multiple diverse representations for each pair [14, 29]. Inspired by these works, we leverage 1:N relationships to augment data in adversarial training and enhance VL robustness.

3. Preliminaries

Vision-language models. Many recent VL models (e.g., ALBEF [15] and BLIP [16]) are fundamentally built on

CLIP, which learns joint image-text representations via large-scale image-text contrastive learning. CLIP consists of an image encoder $f_{\theta_{\text{I}}} : \mathbb{R}^{d_{\text{I}}} \rightarrow \mathbb{R}^{d_{\text{E}}}$ and a text encoder $f_{\theta_{\text{T}}} : \mathbb{R}^{d_{\text{T}}} \rightarrow \mathbb{R}^{d_E}$, where θ_{I} and θ_{T} are their parameters, d_{I} and d_{T} are input dimensions, and d_{E} is the joint embedding dimension. Given an image $I \in \mathbb{R}^{d_{\text{I}}}$ and a text $T \in \mathbb{R}^{d_{\text{T}}}$, CLIP is trained to map them into a shared embedding space, maximizing the cosine similarity of image-text embeddings $S_{\theta_{\text{I,T}}}(I, T) = \cos(f_{\theta_{\text{I}}}(I), f_{\theta_{\text{T}}}(T))$ for correct image-text pairs while minimizing it for incorrect pairs. CLIP optimizes the InfoNCE loss for a batch of N image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$ as:

$$\mathcal{L}_{\text{CLIP-I}}(I, T) = -\sum_{i=1}^{N} \log \frac{\exp(S_{\theta_{\text{I,T}}}(I_i, T_i)/\tau)}{\Sigma_{j=1}^{N} \exp(S_{\theta_{\text{I,T}}}(I_i, T_j)/\tau)}, \quad (1)$$

where τ is the learnable temperature parameter. The overall loss is the average of the losses over images and texts, given by $\mathcal{L}_{\text{CLIP}} = (\mathcal{L}_{\text{CLIP-I}} + \mathcal{L}_{\text{CLIP-T}})/2$, where $\mathcal{L}_{\text{CLIP-T}}$ is the InfoNCE loss over texts.

Multimodal adversarial attacks. We aim to defend VL models against adversarial attacks, where both image and text modalities are perturbed. The objective of (untargeted) adversarial attacks on CLIP is to minimize the image-text similarity $S_{\theta_{I,T}}(I, T)$ for the correct image-text pairs (I, T) to mislead the models' predictions as:

$$(\delta_I^\star, \delta_T^\star) = \underset{\delta_I \in \Delta_I,\, \delta_T \in \Delta_T}{\arg\min} \; S_{\theta_{I,T}}\big(I + \delta_I,\, T + \delta_T\big), \quad (2)$$

$$(I', T') = \big(I + \delta_I^\star,\, T + \delta_T^\star\big). \quad (3)$$

where δ_{I} and δ_{T} are image and text perturbations, and Δ_{I} and Δ_{T} define the set of allowed image and text perturbations. Image attacks maintain perceptual similarity between I and I' typically using the L_p-norm. A common image attack strategy is projected gradient descent (PGD) [20], which iteratively updates I' by taking a small step in the direction of the gradient. Text attacks, such as BERT-Attack [17], modify N tokens in the text T to maximize the divergence between $f_{\theta_{\text{T}}}(T)$ and $f_{\theta_{\text{T}}}(T')$. Multimodal attacks perturb both image-text modalities to create (I', T'); Co-Attack [37] perturbs each modality sequentially, first perturbing the image and then the text; SGA [19] enhances Co-Attack by considering the set-level interaction between multiple images and texts.

Adversarial training for image classification. Adversarial training (AT) [20] is a widely used defense against adversarial attacks, where a model is trained on adversarial examples. It solves a min-max optimization problem:

$$\underset{\theta_{\text{C}}}{\arg\min}\, \mathbb{E}_{(x,y)\sim\mathcal{D}} \left(\max_{\delta_{\text{I}} \in \Delta_{\text{I}}} \left[f_{\theta_{\text{C}}}(x + \delta_{\text{I}}) \neq y \right] \right), \quad (4)$$

where θ_C represents the parameters of image classifier f_{θ_C}, and (x, y) denotes an image and its corresponding class label drawn from the data distribution $\mathcal{D}$.

Based on this, to improve CLIP's adversarial robustness in zero-shot image classification, TeCoA [21] adversarially fine-tunes CLIP's image encoder by minimizing the contrastive loss between adversarial images and the text expression of the corresponding class ("a photo of $\{*\}$"):

$$\underset{\theta_{\text{I}}}{\arg\min}\, \mathbb{E}_{(I,T)\sim\mathcal{D}} \left(\max_{\delta_{\text{I}} \in \Delta_{\text{I}}} \mathcal{L}_{\text{CLIP-I}}(I + \delta_{\text{I}}, T) \right). \quad (5)$$

However, TeCoA only defends against image attacks.

4. Methodology

We propose **M**ultimodal **A**dversarial **T**raining (MAT) by addressing the limitations of existing unimodal AT (Fig. 1). First, we describe its practical optimization strategies via image-text contrastive learning applicable to VL models. Then, we introduce a data-driven approach that leverages the one-to-many relationship between images and texts to enhance robustness by more accurately approximating the true multimodal data distribution.

4.1. Problem setup

To defend against multimodal attacks, we introduce MAT, which perturbs both images and texts during AT. Following the theory of unimodal AT (Eq. 4), we formulate MAT objective as:

$$\min_{\theta}\, \mathbb{E}_{(I,T)\sim\mathcal{D}^\star} \left(\max_{\substack{\delta_{\text{I}} \in \Delta_{\text{I}}, \\ \delta_{\text{T}} \in \Delta_{\text{T}}}} \mathcal{L}_{\text{VL}}(I + \delta_{\text{I}}, T + \delta_{\text{T}}) \right), \quad (6)$$

where $\mathcal{D}^\star$ represents the true data distribution, θ represents the parameters of a VL model, and $\mathcal{L}_{\text{VL}}$ refers to the VL training loss (e.g., $\mathcal{L}_{\text{CLIP}}$ for CLIP).

4.2. Practical multimodal min-max optimization

4.2.1. Multimodal inner maximization

Directly solving the inner-maximization in Eq. 6 is highly non-trivial due to the difficulty of updating both modalities simultaneously, and the increased computational cost.

To address these challenges, we emphasize that exact maximization is not always necessary; in fact, the basic PGD-AT defense method [20] approximates maximization with first-order adversaries. Our extensive analysis led us to design MAT to be both effective and reasonably efficient, while also offering insights for future multimodal defenses.

Specifically, we mainly adopt two practical strategies: (1) *Step-by-step perturbation:* we generate adversarial examples (I', T') by sequentially perturbing the text and image modalities, and (2) *Loss simplification:* replacing $\mathcal{L}_{\text{VL}}$ with effective yet efficient alternative.

Adversarial text generation. First, we generate adversarial texts using a representative text attack, BERT-attack [17]. BERT-attack identifies critical words contributing to the loss and replaces them with semantically similar, grammatically correct alternatives to maximize the loss. However, this discrete optimization process is computationally expensive, and directly maximizing $\mathcal{L}_{\mathrm{VL}}$ over multiple image-text pairs in a batch by repeatedly replacing words would be prohibitively time-consuming. To address this, we exploit the shared image-text alignment objective in VL models (e.g., CLIP, ALBEF, BLIP), and instead of direct loss maximization, we maximize the divergence between individual image-text embeddings as:

$$T' = T + \underset{\delta_{\mathrm{T}} \in \Delta_{\mathrm{T}}}{\arg\max} -\frac{f_{\theta_{\mathrm{I}}}(I) \cdot f_{\theta_{\mathrm{T}}}(T + \delta_T)}{\|f_{\theta_{\mathrm{I}}}(I)\| \|f_{\theta_{\mathrm{T}}}(T + \delta_T)\|}. \quad (7)$$

Adversarial image generation. Next, given adversarial texts T', we generate adversarial images I' using the widely adopted PGD attack [20]. Unlike the discrete optimization in text attacks, image attacks involve continuous optimization, where each update step requires a single back-propagation to maximize a loss function. Therefore, directly maximizing the loss $\mathcal{L}_{\mathrm{VL}}$ is practically feasible. For CLIP, instead of maximizing $\mathcal{L}_{\mathrm{CLIP}}$, we adopt a simple approach by minimizing the cosine similarity between the image-text embeddings as:

$$I' = I + \underset{\delta_{\mathrm{I}} \in \Delta_{\mathrm{I}}}{\arg\max} -\frac{f_{\theta_{\mathrm{I}}}(I + \delta_{\mathrm{I}}) \cdot f_{\theta_{\mathrm{T}}}(T')}{\|f_{\theta_{\mathrm{I}}}(I + \delta_{\mathrm{I}})\| \|f_{\theta_{\mathrm{T}}}(T')\|}. \quad (8)$$

For ALBEF and BLIP, we directly maximize their downstream objective functions, as they involve advanced techniques that cannot be trivially simplified. Please see Appendix A.1 for details.

4.2.2. Multimodal outer minimization

To update the model parameters, we minimize the train loss function $\mathcal{L}_{\mathrm{VL}}$ using the generated adversarial images and texts. For CLIP, we minimize the InfoNCE loss between the adversarial image I' and text T': $\min_\theta \mathcal{L}_{\mathrm{CLIP}}(I', T')$. For ALBEF and BLIP, we fine-tune the models with their original downstream-task-specific objectives.

4.3. Leveraging one-to-many relationships for robustness generalization

While MAT addresses the multimodality of VL tasks, the robustness can be further improved by leveraging also the one-to-many nature of image-text data. Thus, we propose Multimodal Augmented Adversarial Training (MAT+), which incorporates one-to-many augmentations to better approximate the true multimodal distribution of images and texts. Notably, this data-driven strategy for enhancing robustness is applicable to any VL model trained on image-text pairs.

4.3.1. Approximating the multimodal distribution by modeling ambiguity via augmentations

The true distribution of the data $\mathcal{D}^\star$ is inaccessible, but the dataset $\hat{\mathcal{D}}$ serves as an approximation, and its quality affects robustness significantly. In the field of image classification, Gowal et al. [10] proved that adding image augmentations to the training data $\hat{\mathcal{D}}_{\mathrm{tr}}$ produces a $\hat{\mathcal{D}}$ closer to $\mathcal{D}^\star$. This improves robustness under the assumption of a deterministic closed-set image-to-label mapping. However, their theory does not consider the ambiguity of VL data, where a single image can have multiple valid descriptions and vice versa (Fig. 1). Thus, for VL data, multimodal AT with (1:1) image-text pairs only provides a limited $\hat{\mathcal{D}}$. To model this ambiguity, we generate diverse one-to-many (1:N) and many-to-one (N:1) image-text pairs through augmentations, improving the approximation of Eq. 6.

Modeling the ambiguity of text descriptions. Let $\hat{\mathcal{D}}_{\mathrm{tr}}$ be a dataset of deterministic (1:1) image-text pairs (I, T). Relying solely on these pairs produces a weak $\hat{\mathcal{D}}$ because the original text T often provides only a partial or subjective depiction of the image I (Fig. 1). In other words, $\hat{\mathcal{D}}_{\mathrm{tr}}$ approximates the true image-to-text mapping $g^\star_{\mathrm{T}} : \mathbb{R}^{d_{\mathrm{I}}} \rightarrow \mathbb{R}^{d_{\mathrm{T}}}$, a “perfect” human annotator capable of generating detailed descriptions. However, striving for a true $g^\star_{\mathrm{T}}$ is neither desirable nor practical; language compresses visual information by focusing on human interests, naturally involving ambiguity. Achieving $g^\star_{\mathrm{T}}$ would require pixel-level descriptions, which are linguistically unrealistic.

Thus, instead of relying on a single caption per image, we embrace text ambiguity by generating one-to-many (1:N) image-text pairs. In practice, given $\hat{\mathcal{D}}_{\mathrm{tr}}$ with ground-truth pairs (I, T), we can generate new aligned texts using I, T, or both. Consequently, Eq. 6 is approximated as:

$$\min_\theta \mathbb{E}_{\substack{(I,T)\sim\hat{\mathcal{D}}_{\mathrm{tr}}, \\ \phi\sim\Phi}} \left(\max_{\substack{\delta_{\mathrm{I}}\in\Delta_{\mathrm{I}}, \\ \delta_{\mathrm{T}}\in\Delta_{\mathrm{T}}}} \mathcal{L}_{\mathrm{VL}}(I + \delta_{\mathrm{I}}, \hat{g}_{\mathrm{T},\phi}(I, T) + \delta_{\mathrm{T}}) \right), \quad (9)$$

where $\hat{g}_{\mathrm{T},\phi}$ is the approximated image-to-text model, with ϕ capturing the inherent randomness (ambiguity) in generation, and Φ is the set of all possible random variables.

Modeling ambiguity of images. Similarly, a text description can correspond to multiple images, given the high dimensionality and degree-of-freedom of the image data space. For example, the description “a man with an orange hat” could be simultaneously paired with various images, such as a man with a green shirt or engaging in different activities (Fig. 1).

Thus, as in the text case, we introduce image augmentations to create many-to-one image-text pairs to approximate $\mathcal{D}^\star$. Formally, Eq. 6 is approximated as:

$$\min_{\theta} \mathbb{E}_{\substack{(I,T)\sim\hat{\mathcal{D}}_{\text{tr}},\\ \psi\sim\Psi}} \left(\max_{\substack{\delta_{\text{I}}\in\Delta_{\text{I}},\\ \delta_{\text{T}}\in\Delta_{\text{T}}}} \mathcal{L}_{\text{VL}}(\hat{g}_{\text{I},\psi}(I,T)+\delta_{\text{I}}, T+\delta_{\text{T}}) \right), \tag{10}$$

where $\hat{g}_{\text{I},\psi}$ is the approximated text-to-image model, with ψ capturing the inherent randomness, and Ψ denoting the set of all possible random variables.

4.3.2. Conditions for effective augmentations

Not all Φ and Ψ produce a model that can generate valid image-text pairs $\mathcal{I}\times\mathcal{T}\subseteq\mathbb{R}^{d_{\text{I}}}\times\mathbb{R}^{d_{\text{T}}}$, that is, paired data that effectively refine $\hat{\mathcal{D}}$ toward $\mathcal{D}^{\star}$. We can develop the theory of Gowal et al. [10] for image classification to address multimodal VL augmentations. Let p be the probability measure corresponding to $\mathcal{D}^{\star}$, for which every valid pair $(I,T)\in\mathcal{I}\times\mathcal{T}$ has non-zero probability: $p(I,T)>0$. Similarly, let $\hat{p}$ be the probability measure corresponding to $\hat{\mathcal{D}}$. The sufficient condition for $\hat{\mathcal{D}}$ to be an effective approximation is as follows:

Condition 1. (Accurate approximation) The approximated data distribution $\hat{\mathcal{D}}$ and true data distribution $\mathcal{D}^{\star}$ must be equivalent: $\forall(I,T)\in\mathcal{I}\times\mathcal{T}, p(I,T)=\hat{p}(I,T)$.

Ineffective one-to-many augmentations. Let us denote augmentations for image-text pairs $(I,T)\in\hat{\mathcal{D}}_{\text{tr}}$ as $I_{\text{aug}}=\hat{g}_{\text{I},\psi}(I,T)$ and $T_{\text{aug}}=\hat{g}_{\text{T},\phi}(I,T)$. Then, when adding an augmented pair (I_{aug},T) or (I,T_{aug}) to $\hat{\mathcal{D}}_{\text{tr}}$, we define three cases that result into ineffective augmentations, violating Cond. 1.

Case 1. (Semantic mismatch) If augmented image-text pairs are not semantically aligned, $\hat{p}$ diverges from p by generating data in regions of zero probability; that is, $\hat{p}(I_{\text{aug}},T)\neq p(I_{\text{aug}},T)=0$ or $\hat{p}(I,T_{\text{aug}})\neq p(I,T_{\text{aug}})=0$, meaning that the distributions $\mathcal{D}^{\star}$ and $\hat{\mathcal{D}}$ differ.

Case 2. (Limited diversity) Even if augmented image-text pairs are semantically aligned, trivial augmentations $I\simeq I_{\text{aug}}$ or $T\simeq T_{\text{aug}}$ do not provide enough variation to refine $\hat{\mathcal{D}}$ and capture the inherent ambiguity.

Case 3. (Distribution shift) Even if augmented image-text pairs are aligned, excessive augmentations cause distribution shifts. This also applies to the unimodal case [10].

Our results in Sec. 6 show that MAT+ with well-chosen augmentations avoiding these three cases significantly enhances robustness.

5. Experimental settings

We evaluate the multimodal robustness of defense methods on VLMs (i.e., CLIP and ALBEF) in their original tasks they were evaluated in, image-text retrieval (ITR) and visual grounding (VG). We also include BLIP, a VLM capable of text generation for image captioning (IC).

Datasets. We use the commonly used Flickr30k [24] and the COCO [3] datasets for ITR. While these datasets contain five captions per image, our baseline training uses 1:1 image-text pairs (taking the first annotated caption), reflecting the typical setup when fine-tuning with real-world data, such as data collected from the internet, where 1:1 pairings are more common. For VG, we use RefCOCO+ [13], and for IC, we use COCO. See Appendix A for details.

Evaluation metrics. We use the recall@k (R@k) for evaluating ITR, including both image-to-text and text-to-image retrieval. For VG we measure the accuracy in localizing the regions corresponding to the text descriptions, while IC evaluates the quality of generated captions using diverse metrics. Details are in Appendix A.

Adversarial attacks. We evaluate defense methods against the multimodal adversarial attack SGA [19], with perturbation constraints of $\epsilon_I=2/255$ (L_∞-norm) for images and one token for text, following [19, 37]. Results on unimodal attacks (PGD, BERT-Attack) and Co-Attack [37] are in Appendix B.

Baseline defense methods. Since there do not exist defense methods tailored for multimodal attacks, we compare our method with the unimodal (image-only) AT methods closest to our setting proposed for CLIP:

- *FARE* [27]: An unsupervised (unimodal) adversarial fine-tuning scheme for CLIP, which focuses on obtaining a robust CLIP vision encoder. Since it is unsupervised, we can apply it to our ITR setting as-is.
- *TeCoA-ITR*: Since the original TeCoA [21] is text-guided, we extend it to be used in ITR. While the original TeCoA fine-tunes the vision encoder with image classification loss, TeCoA-ITR fine-tunes all parameters using cross-modal objective of Eq. 8 to generate adversarial images.

Training details. We fine-tune the pre-trained CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B models using MAT. Adversarial images are generated via 2-step-PGD (perturbation size of 2/255 in l_∞-norm), and adversarial texts using BERT-attack (1-token perturbation). Please see the detailed hyperparameters in Appendix A.4. Computational cost details are in Appendix A.5.

Augmentation strategies. We consider two types of augmentations: *intra-modal* and *cross-modal*. Intra-modal augmentation enhances data points without considering image-text interactions (text → text, image → image), while cross-modal augmentation enhances data points by leveraging the other modality (image ↔ text). Specifically, we explore the following augmentation techniques:

- Text augmentations:
  - ○ *Intra-modal*: EDA [35] for basic word-level edits. LLM-based rewriting [8] is shown in Appendix B.
  - ○ *Cross-modal*: Image-to-text (I2T) generation using di-

Method	Img aug.	Text aug.	Flickr30k				COCO			
			Clean		SGA		Clean		SGA	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**92.1**	**77.2**	0.6	0.6	**66.6**	**50.1**	0.1	0.1
FARE			75.9	61.0	27.1	21.0	45.2	32.3	9.1	6.9
TeCoA-ITR			83.1	68.2	27.5	17.6	58.0	41.6	9.6	6.2
(ours) MAT			84.6	67.7	36.4	24.9	55.8	40.7	17.7	12.3
(ours) MAT+		Basic(EDA)	85.4 ↑0.8	69.5 ↑1.8	39.1 ↑2.7	27.5 ↑2.6	55.9↑0.2	40.2↓0.5	18.4↑0.7	12.9↑0.6
		I2T(div-Caps)	84.7 ↑0.1	69.2 ↑1.5	40.3 ↑3.9	27.8 ↑2.9	56.7↑0.9	39.9↓0.8	18.9↑1.2	12.5↑0.2
		I2T(Human)	85.7 ↑1.1	71.9 ↑4.2	**45.6** ↑9.2	**32.2** ↑7.3	58.9↑3.1	43.1↑2.4	**21.3**↑3.6	**14.5**↑2.2
	Basic(RandAug)		84.1 ↑0.5	67.1 ↓0.6	35.6 ↓0.8	24.4 ↓0.5	55.9↑0.1	40.7↑0.0	18.3↑0.6	12.6↑0.3
	T2I(SD)		83.3 ↓1.3	68.4 ↑0.7	37.9 ↑1.5	25.2 ↑0.3	54.2↓1.6	38.3↓2.4	17.2↓0.5	11.7↓0.6
	T-I2I(SD)		83.8 ↑0.8	69.1 ↑1.4	39.3 ↑2.9	25.8 ↑0.9	57.1↑1.3	41.7↑1.0	18.8 ↑1.1	12.6 ↑0.2

Table 1. **Comparison of CLIP trained on the Flickr30k and COCO datasets for image-text retrieval (ITR)** under the no-attack scenario (Clean) and multimodal attack (SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

Method	Img aug.	Text aug.	Flickr30k				COCO			
			Clean		SGA		Clean		SGA	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**89.5**	**77.7**	2.5	1.3	**69.9**	**53.6**	1.0	0.7
TeCoA-ITR			85.4	69.3	35.5	21.9	64.8	48.6	14.2	9.5
(ours) MAT			82.0	66.3	47.1	32.9	63.9	46.2	31.2	21.2
(ours) MAT+		Basic(EDA)	82.2 ↑0.2	67.9 ↑1.6	44.6 ↓2.5	31.2 ↓1.7	63.9 ↑0.0	46.8 ↑0.6	31.5 ↑0.3	20.9 ↓0.3
		I2T(div-Caps)	85.6 ↑3.6	71.0 ↑4.8	48.8 ↑1.7	35.0 ↑2.1	66.0 ↑2.0	49.9 ↑3.7	35.5 ↑4.3	20.3 ↓0.9
		I2T(Human)	85.8 ↑3.8	72.8 ↑6.5	52.9 ↑5.8	**38.8** ↑5.9	68.5 ↑4.5	49.1 ↑3.0	**36.2** ↑4.9	**23.5** ↑2.2
	Basic(RandAug)		82.2 ↑0.2	67.2 ↑0.9	48.3 ↑1.2	33.4 ↑0.5	63.1 ↓0.9	48.3 ↑2.1	30.3 ↓1.0	21.7 ↑0.4
	T2I(SD)		83.7 ↑1.7	68.3 ↑2.1	52.0 ↑4.9	36.2 ↑3.3	61.5 ↓2.4	46.0 ↓0.1	25.3 ↓5.9	18.6 ↓2.6
	T-I2I(SD)		85.1 ↑3.1	69.8 ↑3.6	**55.2** ↑8.1	37.6 ↑4.8	64.9 ↑0.9	47.1 ↑1.0	33.2 ↑2.0	22.4 ↑1.2

Table 2. **Comparison of ALBEF trained on Flickr30k and COCO datasets for image-text retrieval (ITR)** under the no-attack scenario (Clean) and a multimodal attack (SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

verse prompts with InternVL [4] ("I2T(div-Caps)"), and human-generated captions from the remaining 4 annotations of Flickr30k and COCO ("I2T(Human)").

- Image augmentations:
  - *Intra-modal*: RandAug [7], randomly applying affine transformations and color distortions.
  - *Cross-modal*: Stable Diffusion (SD) for text-to-image (T2I) [26] and text-guided image-to-image (T-I2I) [22].

Each original data point is augmented four times, leading to a ×5 expansion, except for T-I2I(SD), where using only two augmentations yielded better results. Please see Appendix A.3 for the detailed augmentation settings.

6. Results

6.1. Image-text retrieval (ITR)

6.1.1. Effectiveness of Multimodal Adversarial Training

Defending against multimodal attacks requires multimodal perturbations. With multimodal adversarial perturbations, MAT consistently achieves significantly greater robustness against multimodal attacks than the unimodal AT methods, FARE and TeCoA-ITR, which focus solely on image perturbations during AT. The improvements are substantial and consistent for CLIP on Flickr30k and COCO (Tab. 1), as well as ALBEF on both datasets (Tab. 2). These results highlight the necessity of defense strategies tailored for multimodal perturbations.

Extensive analysis towards efficient and effective multimodal defense. Solving the inner-maximization in Eq. 6 is non-trivial, since it requires updating both modalities and involves a high computational cost. We conducted ablation studies on key design factors—objective functions, perturbation order, and perturbation strength—to make our defense efficient and effective (Tab. 3).

- **Objective functions:** Crossmodal loss (cosine similarity between image-text pairs) outperforms unimodal loss (similarity between orig. and adv. samples), with image-side objective being particularly critical since the image modality is more vulnerable than text (MAT vs. (1-1)/(1-2)). For text perturbations, optimizing the full CLIP loss in BERT-Attack is prohibitively slow ((1-3)); disrupting a single image-text pair suffices with much lower cost.
- **Perturbation order:** The order of T→I or I→T has little effect, while T→I introduced slightly better multimodal robustness (MAT vs. (2-1)). More sequences (e.g., T→I→T) slightly improves multimodal robustness but at a higher cost ((2-2), (2-3)).

	Adversarial Training Config.				Clean		PGD (image)		BERT-Attack (text)		**SGA (multimodal)**		**Time cost**	
	Order	Image Attack (Obj., Optim.)	Text Attack (Obj., Optim.)	Trained params.	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	(sec/iter)	(relative)
Finetune	-	-	-	All	92.1	77.2	11.9	10.1	75.4	53.1	0.6	0.6	1.13	× 0.13
FARE	I	(Uni, PGD-10)	-	Vision	75.9	61.0	69.7	55.1	53.2	40.2	27.1	21.0	7.81	× 0.89
TeCoA-ITR	I	(Cross, PGD-10)	-	All	83.1	68.2	77.7	61.9	64.7	42.7	27.5	17.6	10.29	× 1.17
MAT	T → I	(Cross, PGD-2)	(Cross, BERT)	All	83.7	67.5	77.4	61.4	72.2	51.1	37.5	24.8	8.79	-
Objective ablation														
(1-1)	T → I	(Uni, PGD-2)	(Cross, BERT)	All	90.6	76.5	70.7	57.0	80.4	59.4	16.2	11.6	8.45	× 0.96
(1-2)	T → I	(Cross, PGD-2)	(Uni, BERT)	All	83.4	66.8	79.3	64.2	71.0	49.4	33.3	22.8	8.74	× 0.99
(1-3)	T → I	(CLIP, PGD-2)	(CLIP, BERT)	All	-	-	-	-	-	-	-	-	460.7	× 52.0
Perturbation order ablation														
(2-1)	I → T	(Cross, PGD-2)	(Cross, BERT)	All	84.4	68.7	80.5	65.0	74.8	51.2	36.3	24.6	8.79	× 1.00
(2-2)	T → I → T	(Cross, PGD-2)	(Cross, BERT)	All	83.9	67.3	79.8	64.0	74.6	53.1	38.1	25.9	13.97	× 1.59
(2-3)	I → T → I	(Cross, PGD-2)	(Cross, BERT)	All	82.3	65.7	79.9	62.7	69.3	49.1	37.4	25.1	11.10	× 1.26
Perturbation strength ablation														
(3-1)	T → I	(Cross, PGD-2)	EDA	All	86.2	71.5	82.5	67.4	72.3	48.6	35.5	22.5	3.58	× 0.41
(3-2)	T → I	(Cross, PGD-10)	EDA	All	84.3	67.8	80.9	65.7	68.4	45.7	36.0	23.7	12.18	× 1.39
(3-3)	T → I	(Cross, PGD-10)	(Cross, BERT)	All	79.9	63.8	78.0	61.6	68.2	47.7	38.5	24.7	17.22	× 1.96

Table 3. **Analysis of MAT's strategies.** Ablation study of CLIP trained on Flickr30k for image-text retrieval (ITR), comparing attack objectives (Cross- vs. Uni-modal), perturbation order ("Order"), and attack strength (PGD-2 vs. PGD-10, EDA vs. BERT).

Method	**Aug.**	**Clean**		**SGA**	
		TR@1	IR@1	TR@1	IR@1
Finetune		**72.9**	**57.5**	1.2	1.1
TeCoA-ITR		64.6	51.8	20.2	13.9
(ours) MAT		66.9	49.9	31.3	21.0
(ours) MAT+	I2T(Human)	71.0 ↑4.1	54.3 ↑4.5	**35.6** ↑4.3	**25.7** ↑4.7
	T-I2I(SD)	68.2 ↑1.3	50.5 ↑0.6	33.5 ↑2.2	22.9 ↑1.9

Table 4. **Comparison of BLIP trained on COCO for ITR** under the no-attack scenario (Clean) and the SGA attack, reporting R@k.

- **Perturbation strength:** *Image:* Strength is controlled by PGD iterations. While FARE and TeCoA-ITR use a ten-steps PGD (PGD-10), MAT achieves comparable efficiency by using PGD-2, sufficient for multimodal robustness. Using PGD-10 further improves robustness, at the cost of clean accuracy and efficiency ((3-3)). *Text:* EDA (simple word edits) is an efficient alternative to BERT-Attack ((3-1), (3-2)), since single-token perturbations in the discrete space are largely covered by simple edits.

6.1.2. Effectiveness of one-to-many augmentations

Through our comprehensive analysis, we reveal that some augmentations consistently enhance robustness, whereas some do not. To identify key factors for effective augmentation, we analyze *alignment*, *diversity*, and *distribution gap* based on our hypothesis for approximating the true image-text distribution (Sec.4.3.2). To analyze this, we introduce three measurable properties:

- **Alignment**: The semantic similarity between augmented image-text pairs: $S_{\theta_{\mathrm{I,T}}}(I_{\mathrm{aug}}, T)$ or $S_{\theta_{\mathrm{I,T}}}(I, T_{\mathrm{aug}})$.
- **Diversity**: The Kullback-Leibler (KL) divergence between the original and augmented samples: $d_{\mathrm{KL}}(I, I_{\mathrm{aug}})$ or $d_{\mathrm{KL}}(T, T_{\mathrm{aug}})$.
- **Distribution gap**: The Fréchet distance [9] between the distributions of the original and augmented samples, $d_{\mathrm{F}}(\hat{\mathcal{D}}_{\mathrm{tr}}, \hat{\mathcal{D}}_{\mathrm{aug}})$, where $\hat{\mathcal{D}}_{\mathrm{aug}} = (I_{\mathrm{aug}}, T), (I, T_{\mathrm{aug}})_{i=1}^{N}$. This metric, widely used for evaluating generative models (e.g., FID [12]), quantifies the distribution gap.

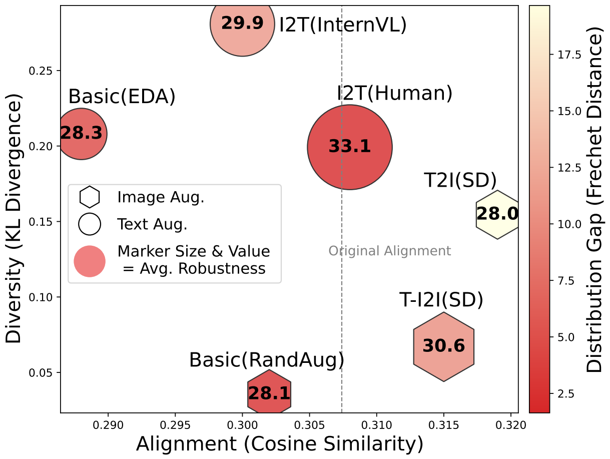
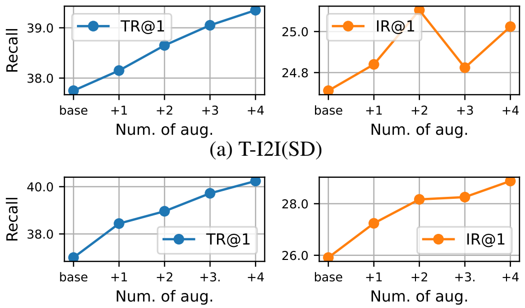


Figure 2. The relationships between the three properties of augmentations, (i) *alignment*, (ii) *diversity*, and (iii) *distribution gap*, versus Robust Accuracy against multimodal attack (the overall average of IR@1/TR@1 for CLIP/ALBEF on Flickr/COCO).

High *alignment* reduces semantic mismatches (Case 1), sufficient *diversity* captures image-text ambiguity, preventing Case 2, and minimal *distribution gap* avoids generating out-of-distribution samples, preventing Case 3.

Fig. 2 reveals that all properties, high alignment, high diversity, and small distribution gap, is crucial for effective augmentations.

Cross-modal augmentation outperforms intra-modal augmentation due to higher image-text alignment. We observe that cross-modal augmentations yield better robustness than intra-modal ones, by considering the other modality and generating well-aligned image-text pairs. For example, the alignment scores of RandAug and EDA are lower than those of the original pairs (Fig. 2), indicating *semantic mismatch* (Case. 1). Intra-modal augmentations struggle to balance diversity and alignment: increasing diversity with-



(a) T-I2I(SD)

(b) I2T(Human)

Figure 3. Analysis of the number of augmentations and robustness against SGA attacks for: (a) image augmentations using T-I2I(SD), and (b) text augmentations using I2T(Human).

out considering image-text relationships disrupts alignment. Fig. 2 shows that intra-modal augmentations appear in the lower-left, achieving only one of alignment or diversity.

Achieving sufficient diversity while keeping the distribution gap minimal is crucial. Figure 2 indicates that even if augmentations are well-aligned as well as sufficiently diverse to expand the training data, they do not enhance robustness when there is a large distribution gap from the original samples. For example, T2I(SD) introduces a good diversity but also introduces a large distribution gap, resulting in minimal robustness improvement.

Appendix D provides qualitative examples of the augmentations generated for each technique.

Image augmentations are challenging. While both image and text augmentations proved effective, text performs better. This is because the high-dimensional image space makes it difficult to generate augmentations that are diverse yet distribution-consistent, whereas the structured nature of text allows controlled augmentation. Figure 2 shows that text augmentations can introduce diversity while maintaining a small distribution gap, unlike image augmentations. Additionally, Fig. 3 analyzes the number of augmentations. Increasing text augmentations with I2T(Human) improves robustness, as more captions better approximate the data distribution, however, image augmentations using T-I2I(SD) enhance robustness up to two additional augmentations but can degrade it beyond that due to distribution shift. These results suggest that developing image augmentations with high diversity yet minimal distribution gap is a promising direction for future research.

Many-to-many augmentations (N:N). Appendix B shows that combining image and text augmentations (N:N) did not enhance performance over text-only augmentations. With two augmented images and four texts, original pairs constitute only a 12.5% of the data, which makes (N:N) augmentations prone to distorting the data distribution if not designed carefully. Thus, while theoretically promising, many-to-many augmentations require a dedicated methodology, which falls out of the scope of this work.

Method	**Aug.**	**SGA**		
		Val	Test-A	Test-B
Fine-tune		32.9	36.3	29.6
TeCoA-ITR		38.8	44.8	32.9
(ours) MAT		40.4	44.2	35.1
(ours) MAT+	I2T(div-Caps)	**43.2** ↑2.8	**48.1** ↑3.9	**37.1** ↑2.1
	I2T(Human)	42.0 ↑1.6	46.5 ↑2.3	35.7 ↑0.6
	T-I2I(SD)	41.1 ↑0.7	46.0 ↑1.8	34.8 ↓0.3

Table 5. Accuracy comparison using ALBEF on the RefCOCO+ dataset for visual grounding.

Method	**Aug.**	**SGA**		
		BLEU-4	ROUGE	CIDEr
Finetune		11.1	35.6	35.5
FARE		23.6	49.0	77.8
(ours) MAT		21.8	46.8	74.4
(ours) MAT+	I2T(Human)	**25.3** ↑3.5	**49.9** ↑3.1	**83.6** ↑9.2

Table 6. Generated text quality comparison using BLIP evaluated on the COCO dataset for IC.

Ablation study: MAT+ vs. naively increasing data samples. Appendix B.3 (Tab. 14) presents controlled experiments to disentangle the effect of data size from the one-to-many (1:N) strategy. With the same number of data points, MAT+ outperforms the naive 1:1 settings.

6.2. Additional tasks

Visual grounding (VG). We evaluate MAT in VG using the model and dataset originally proposed for this task, ALBEF and RefCOCO+ (Tab. 5). MAT proves to be the most effective in this task as well, while MAT+ further improves robustness. This demonstrates the effectiveness of one-to-many augmentations in VL tasks beyond ITR.

Image captioning (IC). We verify the effectiveness of one-to-many augmentations in IC using the model and dataset originally proposed for this task, BLIP and COCO (Tab. 6). Since IC involves only image inputs, MAT does not surpass FARE, the unimodal AT method tailored for image attacks. Nevertheless, augmentations still enhance robustness in this task, with MAT+ outperforming FARE.

7. Conclusions

This work is the first to study adversarial defense for vision-language (VL) models against multimodal attacks in VL tasks. Existing defenses fail against multimodal attacks as they focus on image-only perturbations and overlook the one-to-many nature of image-text pairs. To address this, we proposed MAT, a novel defense framework that leverages multimodal perturbations and one-to-many augmentations. Through comprehensive analysis, we design efficient yet effective multimodal defenses from an optimization perspective and introduce a data-driven approach to further enhance robustness. Our findings reveal key challenges, such as the difficulty in complex optimization and the need for diverse yet in-distribution augmentations, guiding future research.

Acknowledgment This work was partially supported by JSPS KAKENHI Grants JP21H04907 and JP24H00732, by JST CREST Grants JPMJCR20D3 and JPMJCR2562 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, and by JST K Program Grant JPMJKP24C2 Japan.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 12

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 12

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5, 12

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6, 12

[5] Sanghyuk Chun. Improved probabilistic image-text representations. In *Proceedings of the International Conference on Learning Representations*, 2024. 2

[6] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 2

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 6, 13

[8] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 12

[9] Maurice Fréchet. Sur le coefficient de linéarité, dit de corrélation. *Revue de l'Institut International de Statistique*, pages 365–379, 1936. 7

[10] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021. 4, 5

[11] Dongchen Han, Xiaojun Jia, Yang Bai, Jindong Gu, Yang Liu, and Xiaochun Cao. Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. *arXiv preprint arXiv:2312.04403*, 2023. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5, 12

[14] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23422–23431, 2023. 1, 2

[15] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 11, 12, 13

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 11

[17] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020. 2, 3, 4, 11, 14

[18] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pages 605–612, 2004. 12

[19] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. 1, 2, 3, 5, 11, 14

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 3, 4, 14

[21] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *ICLR*, 2022. 1, 2, 3, 5

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 6, 13

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 12

[24] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5

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

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[27] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024. 1, 2, 5

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 12

[29] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 1, 2

[30] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *In Proc. European Conference on Computer Vision*, pages 317–335, 2020. 2

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 12

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 12

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

[34] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. *CVPR*, 2024. 1, 2

[35] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019. 5

[36] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 12

[37] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 1, 2, 3, 5, 11, 14

A. Experimental Details

A.1. MAT's Inner Maximization Strategy

Model	VL Task	Inner maximization		Outer minimization
		Text Adv. Generation	Image Adv. Generation	
CLIP	Image-Text Retrieval	$\max_{\delta_T} \cos(f_{\theta_T}(T+\delta_T), f_{\theta_I}(I))$	$\min_{\delta_I} \cos(f_{\theta_\mathrm{I}}(I+\delta_I), f_{\theta_\mathrm{T}}(T))$	$\mathcal{L}_\mathrm{CLIP}$
ALBEF	Image-Text Retrieval	$\max_{\delta_T} \cos(f_{\theta_T}(T+\delta_T), f_{\theta_I}(I))$	$\max_{\delta_I} \mathcal{L}_\mathrm{ITC} + \mathcal{L}_\mathrm{ITM}$	$\mathcal{L}_\mathrm{ITC} + \mathcal{L}_\mathrm{ITM}$
	Visual Grounding	$\max_{\delta_T} \cos(f_{\theta_T}(T+\delta_T), f_{\theta_I}(I))$	$\max_{\delta_I} \mathcal{L}_\mathrm{ITC} + \mathcal{L}_\mathrm{ITM}$	$\mathcal{L}_\mathrm{ITC} + \mathcal{L}_\mathrm{ITM}$
BLIP	Image-Text Retrieval	$\max_{\delta_T} \cos(f_{\theta_T}(T+\delta_T), f_{\theta_I}(I))$	$\max_{\delta_I} \mathcal{L}_\mathrm{ITC} + \mathcal{L}_\mathrm{ITM}$	$\mathcal{L}_\mathrm{ITC} + \mathcal{L}_\mathrm{ITM}$
	Image Captioning	$\max_{\delta_T} \cos(f_{\theta_T}(T+\delta_T), f_{\theta_I}(I))$	$\max_{\delta_I} \mathcal{L}_\mathrm{LM}$	$\mathcal{L}_\mathrm{LM}$

Table 7. Summary of MAT's strategies for each model and downstream task.

In the main text, we proposed MAT, a multimodal defense framework and practical optimization strategies. In our experiments, we adaptively modify MAT's inner maximization strategy for different downstream tasks, as summarized in Table 7. We provide detailed explanations below.

A.1.1. Image-Text Retrieval (ITR)

MAT for CLIP generates text adversarial examples by maximizing the divergence between image-text embeddings, using BERT-attack [17], while image adversarial examples are generated by minimizing their cosine similarity, following image attacks in Co-Attack [37] and SGA attack [19].

The ALBEF and BLIP model architectures use a similar approach for ITR but include additional components, such as an ITM module (Image-Text Matching) and multimodal encoders. ITC loss ($\mathcal{L}_\mathrm{ITC}$) is the image-text contrastive loss, similar to $\mathcal{L}_\mathrm{CLIP}$, with the primary difference being whether momentum encoders are used to store previously seen representations. ITM loss ($\mathcal{L}_\mathrm{ITM}$) predicts whether an image-text pair is positive (matched) or negative (not matched), which is predicted by a multimodal encoder in ALBEF and BLIP in addition to their unimodal image and text encoders. Please refer to original papers [15, 16] for more details. Given the added complexity in ALBEF and BLIP, MAT for these models is designed as follows:

- Text attack: Same strategy as training CLIP with MAT.
- Image attack: Directly maximize the downstream-task specific fine-tuning objective.

A.1.2. Visual Grounding (VG)

In our experiment, we evaluated the effectiveness of MAT on visual grounding (VG) using ALBEF. Since the fine-tuning objective for VG is the same as for ITR, the inner maximization strategy of MAT is identical to MAT for ITR.

A.1.3. Image Captioning

In our experiment, we evaluated the effectiveness of MAT on image captioning using BLIP. The fine-tuning objective for image captioning includes ITC loss and language modeling (LM) loss, and we maximize the sum of these objectives for perturbing images.

A.2. Evaluation Settings

A.2.1. Image-Text Retrieval (ITR)

Dataset. We use Flickr30k, which consists of image-text pairs with a train/test split of 29,000/1,000 images, and COCO, which consists of image-text pairs with a train/val/test split of 113,287/5,000/5,000 images. We use the default split for training and testing. While each image has five captions, our baseline training approach uses 1:1 image-text pairs, reflecting the practical setting for fine-tuning VL models.
Evaluation metric. We use R@k, which measures the recall of the correct image or text among the top-k retrieved candidates. These metrics assess the model's ability to correctly retrieve relevant images or texts when given a query.

A.2.2. Visual Grounding (VG)

Dataset. For training, we use MSCOCO's training set, following the ITR setting. For testing VG tasks, we use RefCOCO+, which includes text descriptions of objects in images along with their corresponding bounding boxes. RefCOCO+ contains

141,564 expressions for 19,992 images from the COCO training set. This dataset was collected interactively through a two-player game [13]. It consists of three test sets, Val, TestA, and TestB. Test Set A contains objects sampled randomly from the entire dataset. Test Set B contains objects sampled from the most frequently occurring object categories in the dataset. Test Set C contains objects sampled from images that contain at least 2 objects of the same category.

Evaluation metric. VG aims to localize the region in an image that corresponds to a specific text description. Following Li et al. [15], we extend Grad-CAM [28] to acquire heatmaps, and use them to rank the detected proposals provided by Yu et al. [36]. We measure the accuracy of the attention maps with the IoU threshold being 0.5.

A.2.3. Image Captioning

Dataset. We use the MSCOCO dataset for both training and testing. For training data, we follow the ITR setting.

Evaluation metric. We use various evaluation metrics to measure the quality of the generated captions, including BLEU [23], METEOR [2], ROUGE [18], CIDEr [32], and SPICE [1].

A.3. Augmentation techniques

A.3.1. Text augmentations

EDA (Easy Data Augmentation). EDA randomly selects words in the text and performs the following operations: synonym replacement, random insertion, random swap, or random deletion. We use the official implementation [1]. The hyperparameter α controls the strength of the augmentation, where α determines the probability of each word being augmented. We use $\alpha = 0.3$ for all experiments.

LangRW (Language rewrite). Language rewrite (LangRW) [8] is a method that rewrites the text data to improve the robustness of the model, using a generative natural language processing model, such as Llama [31]. We used Llama-2-7B [2]. In our work, we used slightly modified prompts from the original work to simultaneously generate four captions per image. Given an original caption T, the prompt for generating additional captions are as follows:

```
Rewrite image captions in 4 different
ways.

{coco caption 1 for image i}
=> {coco caption 2 for image i}
=> {coco caption 3 for image i}
=> {coco caption 4 for image i}
=> {coco caption 5 for image i}

{coco caption 1 for image j}
=> {coco caption 2 for image j}
=> {coco caption 3 for image j}
=> {coco caption 4 for image j}
=> {coco caption 5 for image j}

{coco caption 1 for image k}
=> {coco caption 2 for image k}
=> {coco caption 3 for image k}
=> {coco caption 4 for image k}
=> {coco caption 5 for image k}

{original caption to be rewritten}
=>
```

where the coco captions are randomly sampled from the original captions from the COCO dataset [3].

I2T(div-Caps) using InternVL. InternVL [4] is a latest vision-language multimodal model. We use the InternVL-2.5-2B model. We generate four captions per image using the following four prompts to ensure diversity:

- *Details*: “Describe the image in detail.”
- *MainObj*: “Describe only the one main object in the image, do not say anything about the other objects or background.”
- *Background*: “Describe only the background of this image, do not say anything about the foreground objects.”

[1] https://github.com/jasonwei20/eda_nlp

[2] https://huggingface.co/meta-llama/Llama-2-7b

- *Style*: "Describe the style or your feelings about this image, do not say anything about the objects in the image."

The lower alignment of InternVL-generated captions compared to original image-text pairs (Fig. 2) is due to the fact that captions generated with *Background* and *Style* are less aligned with images than the full captions.

Human. Human augmentation is a method that generates additional captions by human annotators. Since we use 1:1 image-text pairs for training as default, we used the rest of the original captions included in Flickr30k and COCO datasets as additional captions for each image.

A.3.2. Image augmentations

RandAug (Random Augmentation). RandAug [7] is an image augmentation method that applies a series of random transformations to the image. We used the codes from ALBEF [15] [3]. We set the number of operations to 2 and the magnitude to 5 for all experiments.

Stable Diffusion (SD); Text-to-Image. We used SD-v2.1 [4] for text-to-image augmentations, using Huggingface's default hyperparameters.

Stable Diffusion (SD); Text-guided Image-to-Image. We used SD-v2.1, and use the pipeline for text-guided image-to-image generation [22] from Huggingface[5]. We used a strength hyperparameter of 0.5, which controls the diversity of augmentations. Larger strength increases diversity, however, caused a distribution shift that negatively impacted performance.

A.4. Hyperparameter Details

We fine-tune the pre-trained CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B. We generate adversarial images using 2-step-PGD with a step size of $1.0/255$ with the perturbation bound $\epsilon_{\mathrm{I}} = 2.0/255$. For adversarial texts, we generate with BERT-attack, which replaces a word from top-k candidates with $k = 10$. We train for 5,000 steps on Flickr30k and 10,000 steps on COCO, with a batch size of 128. For CLIP, we train using the SGD optimizer, while for ALBEF and BLIP, we use the AdamW optimizer. All models are trained with cosine learning rate scheduling, a learning rate of 0.0001, and a weight decay of 0.0001.

A.5. Computational Cost

All experiments were conducted on a single NVIDIA A100 GPU.

Since multimodal defense requires perturbing both image and text modalities, naively combining PGD and BERT-Attack nearly doubles the cost compared to unimodal adversarial training methods (e.g., FARE, TeCoA). More complex perturbation sequences (e.g., T→I→T) are also computationally prohibitive. To address this, we carefully designed MAT to be both effective and efficient.

Tab. 3 in the main text compares MAT's training time with that of different ablations of our method. On the one hand, a naive setting with ten-steps PGD (PGD-10) and BERT-Attack (MAT (3-3)) nearly doubles the cost. On the other hand, since baseline unimodal defenses perturb only the image modality, their training time is faster. In comparison, when MAT adopts PGD-2 for image perturbations and a single step-by-step sequence (T→I), we achieve an efficiency comparable to that of FARE and TeCoA-ITR while significantly enhancing multimodal robustness. Thanks to this, training time remains feasible in practice (e.g., CLIP: 12h, BLIP: 24h on COCO); this modest overhead compared to previous—unimodal—approaches is justified by the substantial robustness gains.

In MAT+, the cost of the augmentations depends on each technique. Cross-modal augmentations using generative models (e.g., InternVL for captioning, Stable Diffusion for image-to-image generation) require more time, taking approximately 3 days on COCO. However, the augmentations are created in advance only once, so their generation has no effect on the actual cost of the adversarial training.

[3] https://github.com/salesforce/ALBEF

[4] https://huggingface.co/stabilityai/stable-diffusion-2-1-base

[5] https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/img2img

B. Additional Results

B.1. Image-text retrieval (ITR)

In this section, we present comprehensive results for image-text retrieval (ITR), evaluating defense methods against unimodal attacks (PGD [20] for image attack and BERT-attack [17]) as well as against multimodal attacks (Co-Attack [37] and SGA [19]). Table 8 and Table 9 present the results of CLIP trained on Flickr30k and COCO, respectively. Table 10 and Table 11 show the results of ALBEF trained on Flickr30k and COCO, while Table 12 presents the results of BLIP trained on COCO.

First, the results demonstrate that MAT consistently outperforms baseline unimodal AT methods in terms of multimodal robustness, which is crucial given that multimodal attacks are significantly stronger than unimodal attacks in VL tasks. The robustness against image attacks (PGD) is generally similar between image-only defenses (FARE or TeCoA-ITR) and MAT, which aligns with expectations.

Second, the results indicate that effective augmentations, such as I2T(Human) for text augmentation and T-I2I(SD) for image augmentation, consistently improve multimodal robustness. The important factors that determines the effectiveness of augmentations are described in Sec. 6: both I2T(Human) and T-I2I(SD) augment well-aligned image-text pairs with sufficient diversity, while keeping their distribution close to the original samples.

In addition to the main paper, we present results on many-to-many augmentation in Tab. 9, where we combine image and text augmentations—specifically, T-I2I(SD) for image augmentation and I2T(Human) for text augmentation——aiming at enhancing robustness. However, our results show that this combination does not provide further improvements over using I2T(Human) alone, suggesting that the added image augmentation does not contribute to additional robustness. A possible reason for this is that T-I2I(SD) introduces a slight distribution shift, which may offset its benefits. These findings indicate that while many-to-many augmentations have theoretical potential, they require augmentation techniques that introduce sufficient diversity while maintaining consistency with the original data distribution.

Method	Img aug.	Text. aug.			Unimodal attacks				Multimodal attacks			
			Clean		PGD (image attack)		BERT (text attack)		Co-Attack		SGA	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**92.1**	**77.2**	11.9	10.1	75.4	53.1	11.0	6.7	0.6	0.6
FARE			75.9	61.0	69.7	55.1	53.2	40.2	41.8	30.7	27.1	21.0
TeCoA-ITR			83.1	68.2	77.7	61.9	64.7	42.7	52.9	31.9	27.5	17.6
(ours) MAT			83.7	67.5	77.4	61.4	72.2	51.1	56.9	37.1	37.5	24.8
(ours) MAT+		Basic(EDA)	85.4 ↑1.7	69.5 ↑2.0	79.5 ↑2.1	62.4 ↑1.0	75.7 ↑3.5	53.7 ↑2.6	60.1 ↑3.2	39.5 ↑2.4	39.1 ↑1.6	27.5 ↑2.7
		T2T(LangRW)	80.9 ↓2.8	67.0 ↓0.5	74.2 ↓3.2	60.1 ↓1.3	72.6 ↑0.4	51.3 ↑0.3	58.9 ↑2.0	38.9 ↑1.9	40.0 ↑2.5	27.4 ↑2.6
		I2T(div-Caps)	84.7 ↑1.0	69.2 ↑1.7	79.0 ↑1.6	64.5 ↑3.1	74.8 ↑2.6	51.4 ↑0.3	60.3 ↑3.4	39.6 ↑2.6	40.3 ↑2.8	27.8 ↑3.0
		I2T(Human)	85.7 ↑2.0	71.9 ↑4.4	82.0 ↑4.6	65.9 ↑4.5	**77.6** ↑5.4	**56.3** ↑5.2	**63.3** ↑6.4	**44.0** ↑6.9	**45.6** ↑8.1	**32.2** ↑7.4
	Basic(RandAug)		84.1 ↑0.4	67.1 ↓0.4	78.7 ↑1.3	61.6 ↑0.2	70.9 ↓1.3	50.4 ↓0.7	57.8 ↑0.9	37.3 ↑0.2	35.6 ↓1.9	24.4 ↓0.4
	T-I2I(SD)		83.8 ↑0.1	69.1 ↑1.6	77.7 ↑0.3	62.2 ↑0.8	75.3 ↑3.1	52.1 ↑1.0	59.7 ↑2.8	38.0 ↑1.0	39.3 ↑1.8	25.8 ↑1.0
	T2I(SD)		83.3 ↓0.4	68.4 ↑0.9	76.1 ↓1.3	61.5 ↑0.1	73.3 ↑1.1	52.2 ↑1.1	57.2 ↑0.3	37.8 ↑0.7	37.9 ↑0.4	25.2 ↑0.4

Table 8. Comparison of CLIP trained on the Flickr30k dataset for ITR under the no-attack scenario (Clean), unimodal attacks (PGD and BERT), and multimodal attacks (Co-Attack and SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

					Unimodal attacks				Multimodal attacks			
Method	**Img aug.**	**Text. aug.**	**Clean**		**PGD (image attack)**		**BERT (text attack)**		**Co-Attack**		**SGA**	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**66.6**	**50.1**	6.8	5.4	36.9	23.7	2.9	1.8	0.1	0.1
FARE			45.2	32.3	40.4	29.3	22.4	15.9	16.7	11.4	9.1	6.9
TeCoA-ITR			58.0	41.6	51.0	36.7	30.6	18.2	21.3	12.5	9.6	6.2
(ours) MAT			55.8	40.7	49.5	36.1	42.9	27.4	31.4	19.5	17.7	12.3
(ours) MAT+		Basic(EDA)	55.9 ↑0.2	40.2 ↓0.5	48.6 ↓0.9	35.3 ↓0.8	41.1 ↓1.8	27.1 ↓0.4	29.8 ↓1.6	19.5 ↓0.0	18.4 ↑0.7	12.9 ↑0.6
		T2T(LangRW)	51.8 ↓4.0	37.7 ↓3.0	45.7 ↓3.8	33.5 ↓2.5	38.7 ↓4.2	26.1 ↓1.3	28.7 ↓2.7	19.0 ↓0.6	17.0 ↓0.7	12.6 ↑0.3
		I2T(div-Caps)	56.7 ↑0.9	39.9 ↓0.8	51.2 ↑1.8	36.3 ↑0.2	40.2 ↓2.6	26.1 ↓1.3	30.3 ↓1.1	19.4 ↓0.2	18.9 ↑1.2	12.5 ↑0.2
		I2T(Human)	58.9 ↑3.1	43.1 ↑2.4	**52.2** ↑2.7	**38.5** ↑2.4	**44.7** ↑1.9	**29.9** ↑2.4	**33.6** ↑2.2	**21.9** ↑2.3	**21.3** ↑3.6	**14.5** ↑2.2
	Basic(RandAug)		55.9 ↑0.1	40.7 ↓0.0	49.4 ↓0.1	36.3 ↑0.2	41.7 ↓1.2	27.2 ↓0.2	30.7 ↓0.7	19.6 ↑0.1	18.3 ↑0.6	12.6 ↑0.3
	T2I(SD)		54.2 ↓1.6	38.3 ↓2.4	46.3 ↓3.2	33.4 ↓2.7	39.5 ↓3.3	25.9 ↓1.5	28.5 ↓2.9	18.1 ↓1.5	15.9 ↓1.8	11.0 ↓1.3
	T,I2I(SD)		57.1 ↑1.3	41.7 ↑1.0	49.7 ↑0.2	36.0 ↓0.1	41.9 ↓0.9	28.3 ↑0.8	31.3 ↓0.0	20.2 ↑0.7	18.8 ↑1.1	12.6 ↑0.2
(N:N)	T,I2I(SD)	I2T(Human)	59.0 ↑3.2	43.3 ↑2.6	52.2 ↑2.7	38.0 ↑1.9	44.6 ↑1.7	29.8 ↑2.3	**34.1** ↑2.8	21.2 ↑1.7	20.8 ↑3.1	14.2 ↑1.9

Table 9. Comparison of CLIP trained on the COCO dataset for ITR under the no-attack scenario (Clean), unimodal attacks (PGD and BERT), and multimodal attacks (Co-Attack and SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

					Unimodal attacks				Multimodal attacks			
Method	**Img aug.**	**Text. aug.**	**Clean**		**PGD (image attack)**		**BERT (text attack)**		**Co-Attack**		**SGA**	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**89.5**	**77.7**	48.8	34.0	79.1	55.0	32.3	18.4	2.5	1.3
TeCoA-ITR			85.4	69.3	78.7	63.1	72.1	48.9	61.0	38.1	35.5	21.9
(ours) MAT			82.0	66.3	78.0	63.8	77.1	56.1	73.0	54.0	47.1	32.9
(ours) MAT+		Basic(EDA)	82.2 ↑0.2	67.9 ↑1.6	78.2 ↑0.2	63.8 ↑0.0	78.6 ↑1.5	58.1 ↑2.0	71.1 ↓1.9	52.9 ↓1.1	44.6 ↓2.5	31.2 ↓1.7
		T2T(LangRW)	81.4 ↓0.6	67.5 ↑1.3	76.0 ↓2.0	62.5 ↓1.3	75.8 ↓1.3	56.2 ↑0.1	70.0 ↓3.0	51.1 ↓3.0	46.3 ↓0.8	32.4 ↓0.5
		I2T(div-Caps)	85.6 ↑3.6	71.0 ↑4.8	**81.7** ↑3.7	66.6 ↑2.8	79.7 ↑2.6	60.1 ↑3.9	75.6 ↑2.6	55.2 ↑1.2	48.8 ↑1.7	35.0 ↑2.1
		I2T(Human)	85.8 ↑3.8	72.8 ↑6.5	80.2 ↑2.2	**67.6** ↑3.8	**81.5** ↑4.4	**62.8** ↑6.7	**77.5** ↑4.5	**58.7** ↑4.7	52.9 ↑5.8	**38.8** ↑5.9
	Basic(RandAug)		82.2 ↑0.2	67.2 ↑0.9	78.3 ↑0.3	63.2 ↓0.6	76.9 ↓0.2	56.2 ↑0.1	72.7 ↓0.3	53.2 ↓0.8	48.3 ↑1.2	33.4 ↑0.5
	T-I2I(SD)		85.1 ↑3.1	69.8 ↑3.6	81.7 ↑3.7	67.1 ↑3.3	79.4 ↑2.3	59.1 ↑2.9	74.9 ↑1.9	55.9 ↑1.8	**55.2** ↑8.1	37.6 ↑4.8
	T2I(SD)		83.7 ↑1.7	68.3 ↑2.1	79.5 ↑1.5	63.6 ↓0.2	76.7 ↓0.4	59.0 ↑2.9	72.4 ↓0.6	55.1 ↑1.1	52.0 ↑4.9	36.2 ↑3.3

Table 10. Comparison of ALBEF trained on the Flickr30k dataset for ITR under the no-attack scenario (Clean), unimodal attacks (PGD and BERT), and multimodal attacks (Co-Attack and SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

					Unimodal attacks				Multimodal attacks			
Method	**Img aug.**	**Text. aug.**	**Clean**		**PGD (image attack)**		**BERT (text attack)**		**Co-Attack**		**SGA**	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**69.9**	**53.6**	30.0	17.4	49.0	31.4	17.2	9.1	1.0	0.7
TeCoA-ITR			64.8	48.6	49.5	38.1	48.5	28.5	38.2	19.4	14.2	9.5
(ours) MAT			63.9	46.2	60.2	38.3	55.4	36.9	52.9	32.1	31.2	21.2
(ours) MAT+		Basic(EDA)	63.9 ↑0.0	46.8 ↑0.6	53.2 ↓7.0	38.7 ↑0.4	54.4 ↓1.0	36.4 ↓0.5	50.8 ↓2.2	30.9 ↓1.2	31.5 ↑0.3	20.9 ↓0.3
		T2T(LangRW)	59.7 ↓4.2	42.1 ↓4.1	46.6 ↓13.5	31.0 ↓7.3	51.4 ↓4.1	32.9 ↓4.0	34.6 ↓18.3	23.9 ↓8.2	25.4 ↓5.8	15.7 ↓5.5
		I2T(div-Caps)	66.0 ↑2.0	49.9 ↑3.7	57.0 ↓3.2	37.9 ↓0.4	56.6 ↑1.2	38.5 ↑1.6	47.5 ↓5.4	26.4 ↓5.6	35.5 ↑4.3	20.3 ↓0.9
		I2T(Human)	68.5 ↑4.5	49.1 ↑3.0	**64.9** ↑4.7	37.3 ↓1.0	**59.7** ↑4.3	**39.9** ↑3.0	**55.3** ↑2.4	29.3 ↓2.8	**36.2** ↑4.9	**23.5** ↑2.2
	Basic(RandAug)		63.1 ↓0.9	48.3 ↑2.1	61.1 ↑0.9	**46.0** ↑7.7	54.8 ↓0.7	38.3 ↑1.4	52.1 ↓0.8	35.7 ↑3.7	30.3 ↓1.0	21.7 ↑0.4
	T2I(SD)		61.5 ↓2.4	46.0 ↓0.1	58.7 ↓1.5	37.8 ↓0.5	53.5 ↓2.0	37.5 ↑0.6	50.6 ↓2.3	33.0 ↑1.0	25.3 ↓5.9	18.6 ↓2.6
	T,I2I(SD)		64.9 ↑0.9	47.1 ↑1.0	61.8 ↑1.6	36.1 ↓2.2	56.5 ↑1.1	38.1 ↑1.3	54.1 ↑1.2	29.2 ↓2.9	33.2 ↑2.0	22.4 ↑1.2

Table 11. Comparison of ALBEF trained on the COCO dataset for ITR under the no-attack scenario (Clean), unimodal attacks (PGD and BERT), and multimodal attacks (Co-Attack and SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

					Unimodal attacks				Multimodal attacks			
Method	**Img aug.**	**Text. aug.**	**Clean**		**PGD (image attack)**		**BERT (text attack)**		**Co-Attack**		**SGA**	
			TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
Finetune			**72.9**	**57.5**	50.9	35.7	39.2	27.5	18.4	9.9	1.2	1.1
TeCoA-ITR			64.6	51.8	63.8	50.3	37.6	24.9	31.2	20.2	20.2	13.9
(ours) MAT			66.9	49.9	67.8	49.5	54.7	37.1	50.5	32.8	31.3	21.0
(ours) MAT+		I2T(Human)	71.0 ↑4.1	54.3 ↑4.5	**68.6** ↑0.8	**52.1** ↑2.6	**57.1** ↑2.4	**40.4** ↑3.3	**51.2** ↑0.7	**35.3** ↑2.5	**35.6** ↑4.3	**25.7** ↑4.7
	T,I2I(SD)		68.2 ↑1.3	50.5 ↑0.6	62.3 ↓5.5	43.6 ↓6.0	55.0 ↑0.2	37.0 ↓0.0	47.9 ↓2.6	29.9 ↓3.0	33.5 ↑2.2	22.9 ↑1.9

Table 12. Comparison of BLIP trained on the COCO dataset for ITR under the no-attack scenario (Clean), unimodal attacks (PGD and BERT), and multimodal attacks (Co-Attack and SGA), reporting R@k for text retrieval (TR@k) and image retrieval (IR@k).

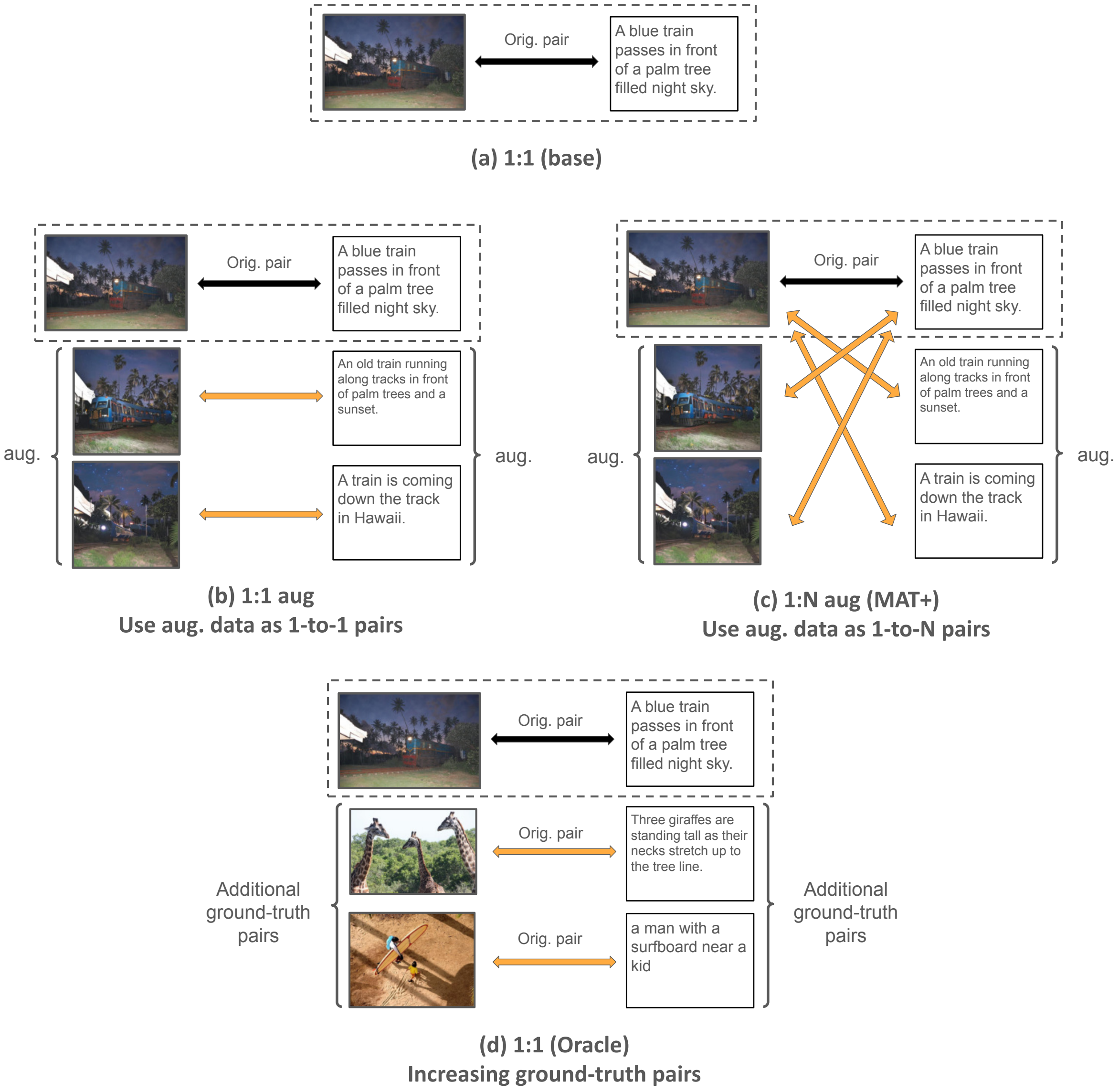
B.2. Visual grounding (VG)

Table 13 presents results of ALBEF trained on the COCO dataset for the VG task. The results highlight MAT's effectiveness beyond ITR task, as well as the effectiveness of one-to-many augmentations. We observe that I2T(div-Caps) outperforms I2T(Human), which can be attributed to the design of the prompts used in I2T(div-Caps). Specifically, I2T(div-Caps) generates captions that focus on both foreground and background objects, providing a more comprehensive description of the image (see Sec. A.3 for the prompt design). This makes I2T(div-Caps) particularly useful for the VG task, where understanding diverse objects located in different areas of the image is crucial.

Method	Img aug.	Text. aug.	Clean			SGA		
			val	test-A	test-B	val	test-A	test-B
Fine-tune			50.2	**57.6**	40.6	32.9	36.3	29.6
TeCoA-ITR			49.6	55.4	41.4	38.8	44.8	32.9
(ours) MAT			48.8	53.3	40.7	40.4	44.2	35.1
(ours) MAT+		I2T(div-Caps)	**51.0** ↑2.2	57.0 ↑3.7	**43.3** ↑2.6	**43.2** ↑2.8	**48.1** ↑3.9	**37.1** ↑2.1
		I2T(Human)	50.6 ↑1.8	55.9 ↑2.6	41.7 ↑1.1	42.0 ↑1.6	46.5 ↑2.3	35.7 ↑0.6
	T,I2I(SD)		49.6 ↑0.7	55.9 ↑2.6	40.8 ↑0.1	41.1 ↑0.7	46.0 ↑1.8	34.8 ↓0.3

Table 13. Accuracy comparison of ALBEF trained on the COCO dataset for VG under the no-attack scenario (Clean) and multimodal attack (SGA).

B.3. Effectiveness of MAT+ vs. naively increasing data samples


Orig. pair
A blue train passes in front of a palm tree filled night sky.
aug.
An old train running along tracks in front of palm trees and a sunset.
A train is coming down the track in Hawaii.
aug.
Additional ground-truth pairs
Three giraffes are standing tall as their necks stretch up to the tree line.
a man with a surfboard near a kid
Additional ground-truth pairs

(a) 1:1 (base)

(b) 1:1 aug
Use aug. data as 1-to-1 pairs

(c) 1:N aug (MAT+)
Use aug. data as 1-to-N pairs

(d) 1:1 (Oracle)
Increasing ground-truth pairs

Figure 4. **Different ways of increasing the number of samples for adversarial training. (a) Original data samples (no increasing), (b) Augmenting orig. data samples without 1:N modeling, (c) Augmenting orig. data samples with 1:N modeling (MAT+), (d) Adding new orig. data samples (oracle).**

B.3.1. Experimental Settings

We conducted controlled experiments to disentangle the effect of data size from the one-to-many training strategy. Suppose we have M image-text pairs and generate two augmented samples per modality, yielding M original and $2M$ augmented images and texts (e.g., 30k original and 60k augmentations, in total of 90k). Then, we compare four settings (see Fig. 4):

- (a) **1:1 (base):** M original ground-truth pairs.
- (b) **1:1 aug:** Adding $2M$ augmentations and creating additional 1:1 pairs by directly combining augmented images and texts.
- (c) **1:N aug (MAT+):** Adding the same $2M$ augmentations as in (b), but each original sample forms 1-to-3 and 3-to-1 pairs with its two augmentations.
- (d) **1:1 (oracle):** Adding $2M$ extra original ground-truth pairs (no augmentations), serving as an upper bound.

This setup enables a fair comparison between naively adding new augmentations (1:1) and using MAT+ (1:N) under the same data (i.e., (b) and (c)), while also contrasting them with an upper-bound oracle.

The base set ($M = 30k$) was sampled randomly from COCO's ~120k image-text training pairs. Then, we generated $60k$ augmentations using "T-I2I(SD)" for images and "I2T(Human)" for texts, which were the best performing techniques for each modality in our analysis. On the other hand, the oracle setting adds $60k$ randomly sampled ground-truth pairs, for a total of $90k$ original pairs.

B.3.2. Results

	Num. Orig. (GT)	Num. Aug.	Augmentation		Clean		PGD		BERT-Attack		**SGA**	
			Img.	Text	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1
(a) 1:1 (base)	30k	-	-	-	45.0	32.0	41.0	29.8	32.2	20.9	10.8	7.3
(b) 1:1	30k	60k	T-I2I(SD)	I2T(Human)	50.2	36.4	46.4	33.5	37.5	24.3	15.5	10.6
(c) MAT+ (ours)	30k	60k	T-I2I(SD)	I2T(Human)	54.6 ↑4.4	39.0 ↑2.6	49.3 ↑2.9	36.1 ↑2.6	40.1 ↑2.6	26.3 ↑2.0	16.3 ↑0.8	11.4 ↑0.8
Augmentation ablation												
	30k	60k	T-I2I(SD)	*EDA*	49.5	35.5	45.3	32.6	36.5	23.2	13.2	9.2
	30k	60k	*RandAug*	I2T(Human)	52.7	38.0	48.8	35.2	38.4	24.9	15.3	10.6
(d) 1:1 (oracle)	90k	-	-	-	53.9	39.8	50.0	36.3	41.1	26.7	16.1	11.8

Table 14. Controlled experiments to disentangle the effects of data size and one-to-many (1:N) augmentation (see Fig. 4 for augmentation types a~d). Compared to simply adding augmented pairs as 1:1 (b), using them as 1:N pairs (c, MAT+) yields larger robustness gains, nearly matching the oracle 1:1 setting (d).

Tab. 14 shows that while adding synthetic data as 1:1 pairs improves robustness, using them as 1:N pairs yields substantially larger gains, nearly matching the oracle 1-to-1 setting. This confirms that robustness improvements come not only from increased data size but also from **ambiguity modeling through 1-to-N alignment**.

Besides ambiguity modeling, a key contribution of our work is identifying the properties of effective augmentations in adversarial defense—**high alignment**, **high diversity**, and **small distribution gap**. In Tab. 14(c) *Augmentation ablation*, naive augmentations (RandAug and EDA) yield only limited improvements compared with higher-quality augmentations (T-I2I(SD) and I2T(Human)).

In summary, naively increasing the number of multimodal pairs is suboptimal compared with our proposed method, whose robustness is comparable to expanding the dataset by collecting new ground-truth samples. Also note that using MAT+ on (d) further boosts the accuracy, as shown in the main text when using the entire COCO dataset.

C. Qualitative Results

Here, we present qualitative results for ITR. We compare the TeCoA-ITR baseline, a unimodal image defense method, with our proposed multimodal defense method, MAT+, which incorporates I2T(Human) augmentations. Both methods are evaluated against a multimodal attack (SGA). Figure 5 and Figure 6 illustrate the comparison for image retrieval and text retrieval results, respectively.

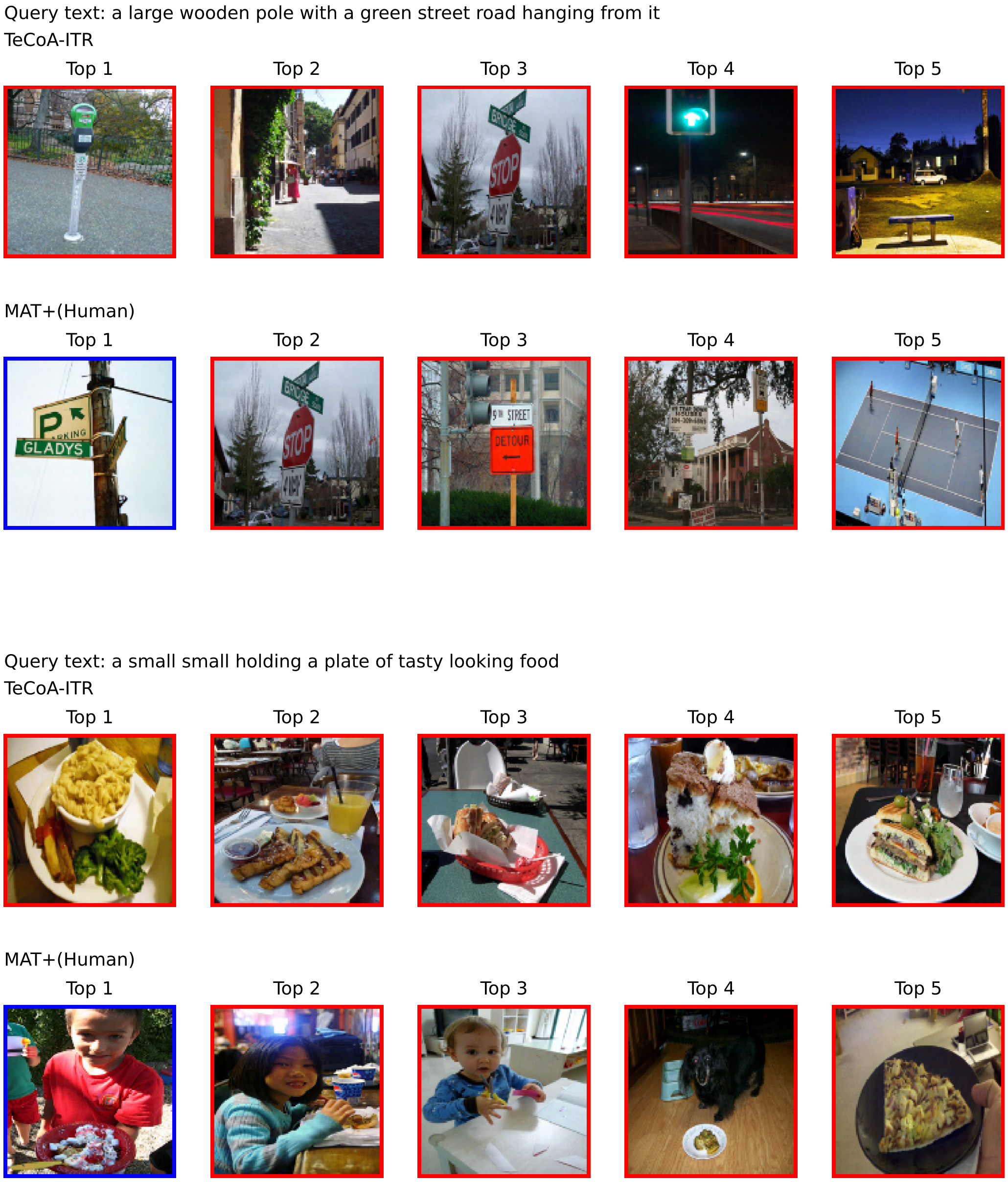


Figure 5. **Qualitative comparison of image retrieval** results under multimodal attack (SGA). We compare the TeCoA-ITR baseline (unimodal defense) with our proposed MAT+ (multimodal defense using I2T(Human) augmentations). Images with a blue border indicate correct retrieval, while those with a red border indicate incorrect retrieval.

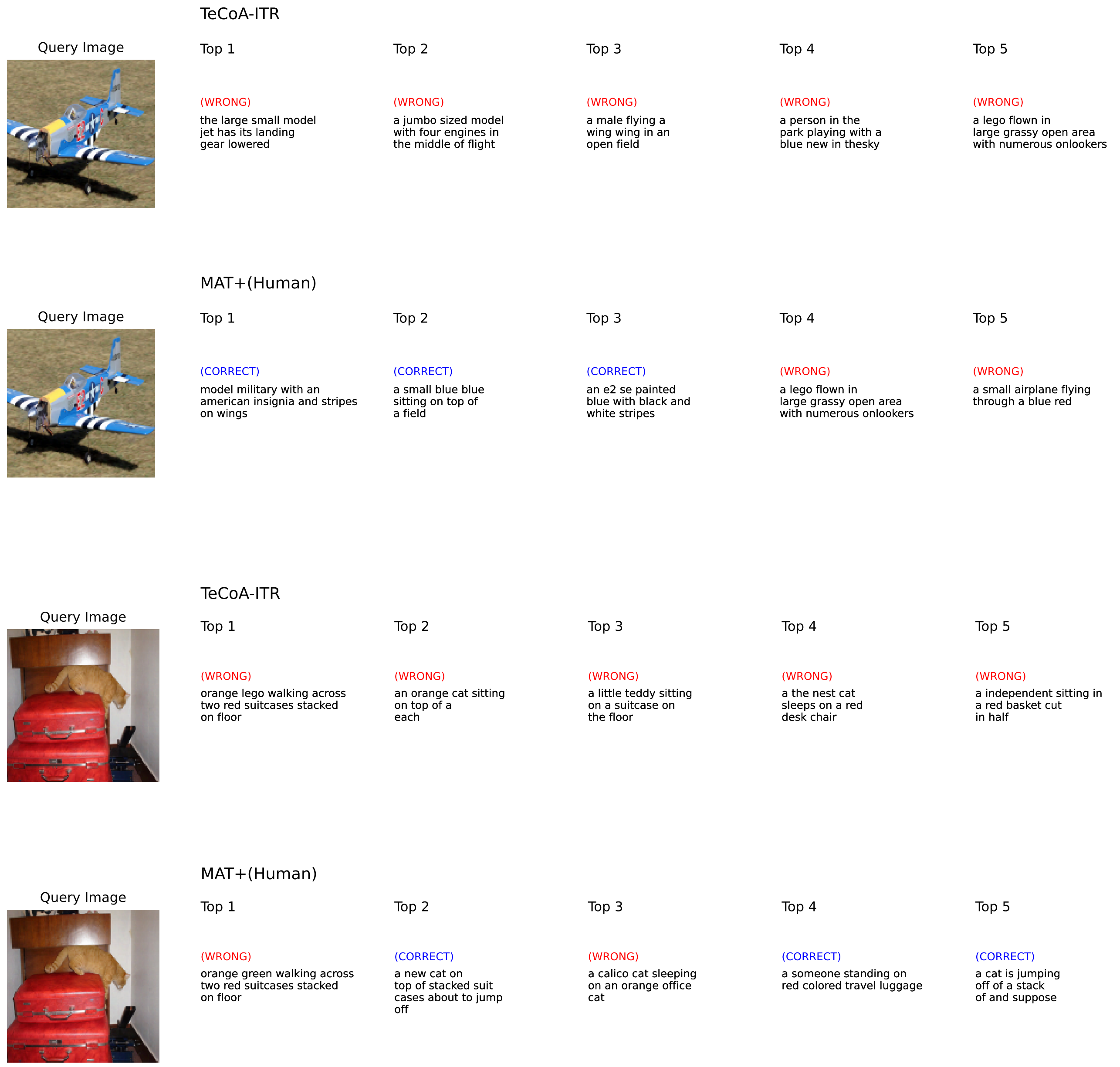

TeCoA-ITR
Query Image
Top 1
Top 2
Top 3
Top 4
Top 5
(WRONG)
the large small model jet has its landing gear lowered
(WRONG)
a jumbo sized model with four engines in the middle of flight
(WRONG)
a male flying a wing wing in an open field
(WRONG)
a person in the park playing with a blue new in thesky
(WRONG)
a lego flown in large grassy open area with numerous onlookers
MAT+(Human)
Query Image
Top 1
Top 2
Top 3
Top 4
Top 5
(CORRECT)
model military with an american insignia and stripes on wings
(CORRECT)
a small blue blue sitting on top of a field
(CORRECT)
an e2 se painted blue with black and white stripes
(WRONG)
a lego flown in large grassy open area with numerous onlookers
(WRONG)
a small airplane flying through a blue red
TeCoA-ITR
Query Image
Top 1
Top 2
Top 3
Top 4
Top 5
(WRONG)
orange lego walking across two red suitcases stacked on floor
(WRONG)
an orange cat sitting on top of a each
(WRONG)
a little teddy sitting on a suitcase on the floor
(WRONG)
a the nest cat sleeps on a red desk chair
(WRONG)
a independent sitting in a red basket cut in half
MAT+(Human)
Query Image
Top 1
Top 2
Top 3
Top 4
Top 5
(WRONG)
orange green walking across two red suitcases stacked on floor
(CORRECT)
a new cat on top of stacked suit cases about to jump off
(WRONG)
a calico cat sleeping on an orange office cat
(CORRECT)
a someone standing on red colored travel luggage
(CORRECT)
a cat is jumping off of a stack of and suppose

Figure 6. **Qualitative comparison of text retrieval** results under multimodal attack (SGA). We compare the TeCoA-ITR baseline (unimodal defense) with our proposed MAT+ (multimodal defense using I2T(Human) augmentations).

D. Visualization of Augmentations

This section visualizes each augmentation technique. Figure 7 and Fig. 8 show image augmentations, and Fig. 9 and Fig. 10 visualize text augmentations. These visualizations help in understanding image-text alignment, augmentation diversity, and the distribution of augmentations. For example, Fig.7 and Fig.8 show that while I2T(SD) generates diverse images, they differ significantly from the original images and often appear somewhat synthetic, potentially causing a distribution shift. Additionally, Fig. 9 and Fig. 10 show that Basic(EDA), which is a basic word-level augmentation, can disrupt the image-text alignment.

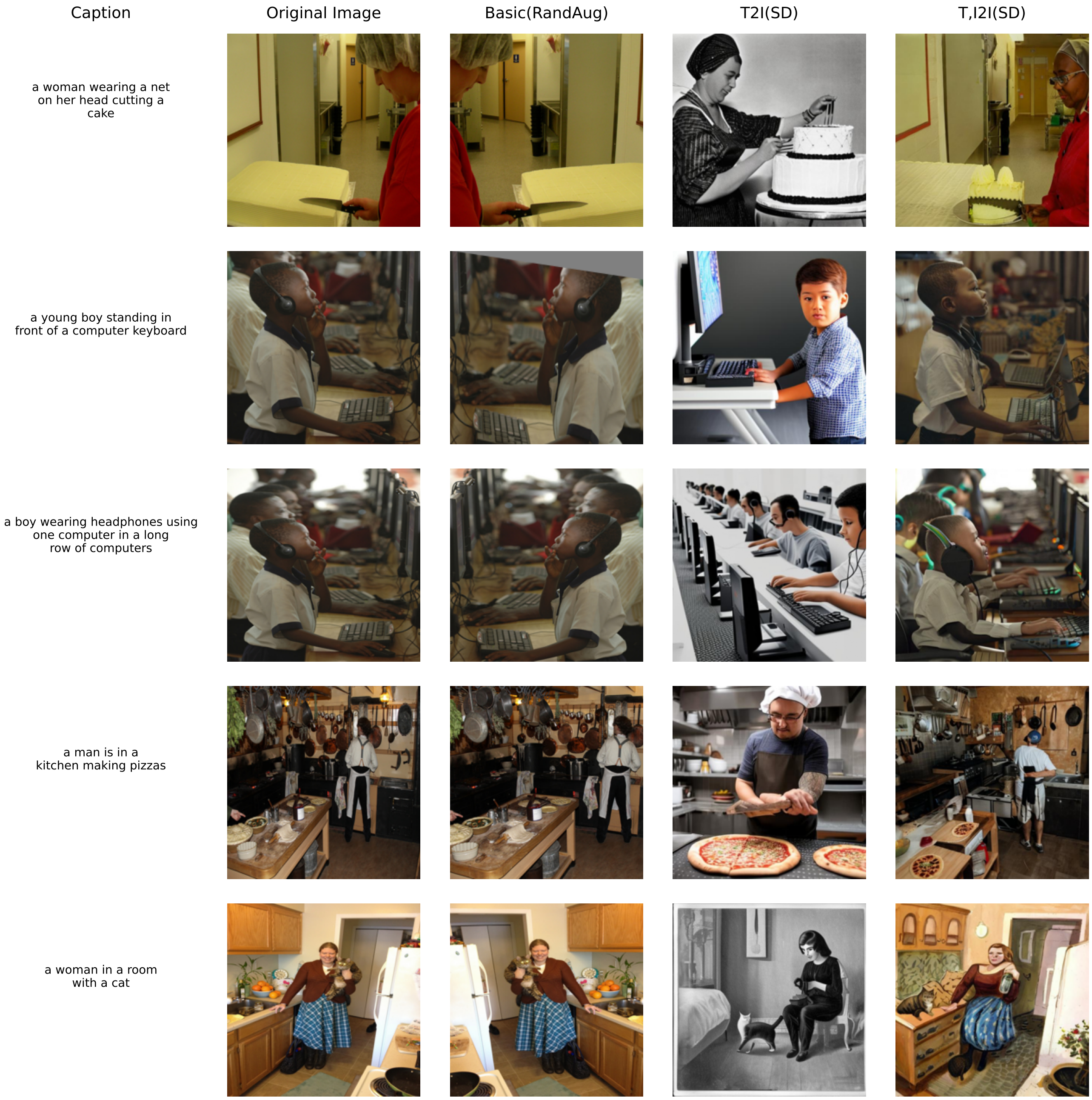

Caption
Original Image
Basic(RandAug)
T2I(SD)
T,I2I(SD)
a woman wearing a net on her head cutting a cake
a young boy standing in front of a computer keyboard
a boy wearing headphones using one computer in a long row of computers
a man is in a kitchen making pizzas
a woman in a room with a cat

Figure 7. **Visualization of image augmentations (1).**

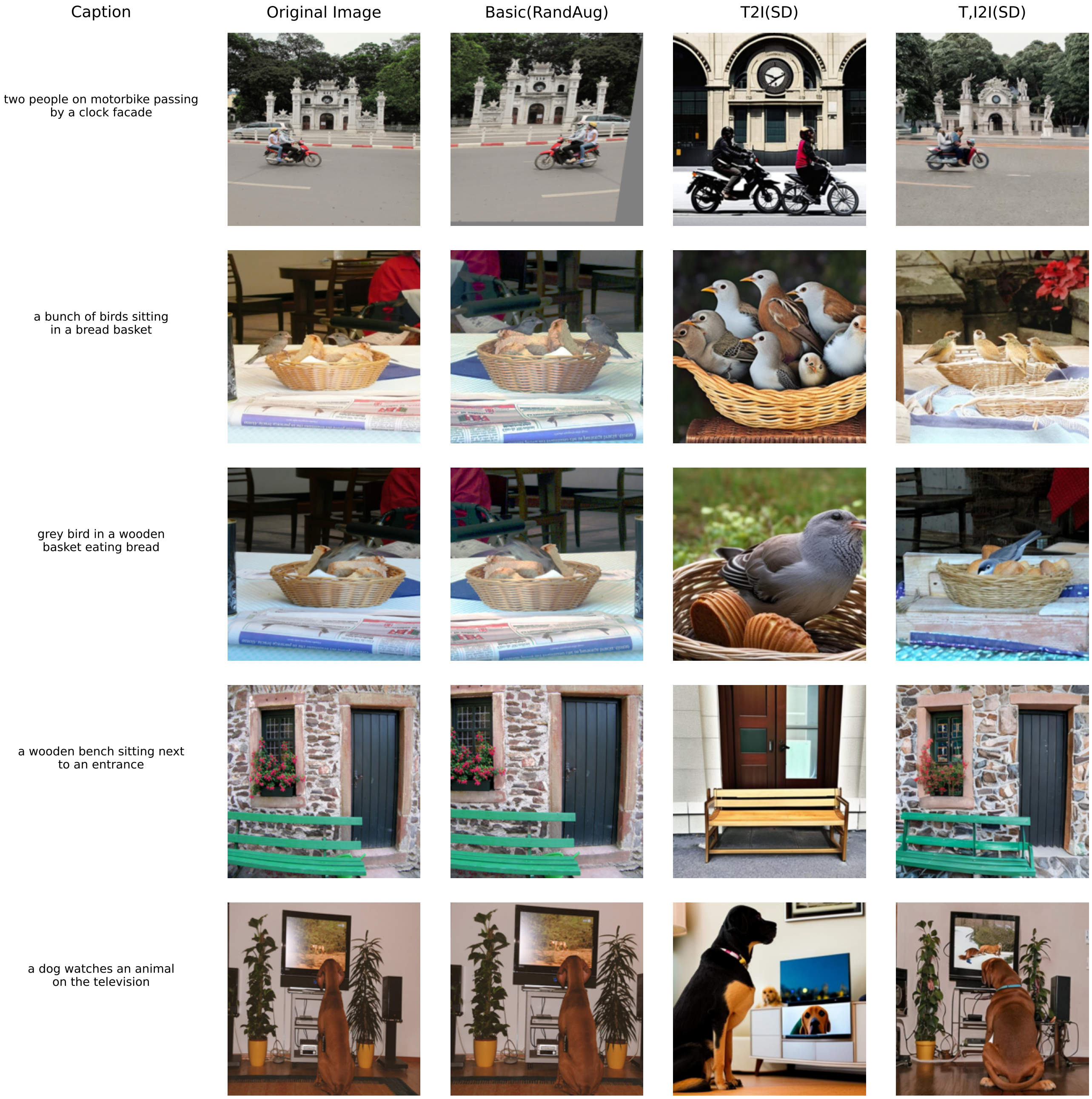

Caption
Original Image
Basic(RandAug)
T2I(SD)
T,I2I(SD)
two people on motorbike passing by a clock facade
a bunch of birds sitting in a bread basket
grey bird in a wooden basket eating bread
a wooden bench sitting next to an entrance
a dog watches an animal on the television

Figure 8. **Visualization of image augmentations (2).**

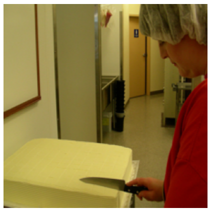
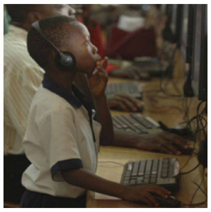
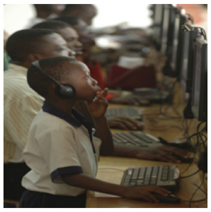
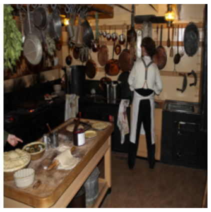
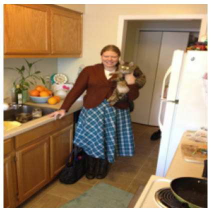
Image	Original Caption	Basic(EDA)	I2T(InternVL)	I2T(Human)
	a woman wearing a net on her head cutting a cake	a woman wearing a net on her head bar womanhood cutting a womanhood cake	the image shows a person wearing a red shirt and a hairnet standing in what appears to be a kitchen or a bakery they are holding a large knife and	a woman cutting a large white sheet cake
	a young boy standing in front of a computer keyboard	a computer keyboard	the image depicts a young child likely a student wearing a white shirt with a dark collar and a headset on the child is focused on a computer screen with	a little boy wearing headphones and looking at a computer monitor
	a boy wearing headphones using one computer in a long row of computers	a son wearing headphones using one computer in a foresighted row of computer	the image depicts a classroom setting where students are seated at desks each equipped with a computer the students appear to be focused on their screens possibly engaged in an	a little boy with earphones on listening to something
	a man is in a kitchen making pizzas	a man is in a kitchen progress to pizza pie	the image depicts a cozy rustic kitchen with a person standing at the stove seemingly engaged in cooking the kitchen is well equipped with various pots pans and utensils hanging	man in apron standing on front of oven with pans and bakeware
	a woman in a room with a cat	woman in a room a	the image shows a woman standing in a kitchen smiling at the camera she is wearing a brown sweater a blue and white plaid skirt and black boots she is	a girl smiles as she holds a cat and wears a brightly colored skirt

Figure 9. **Visualization of text augmentations (1).**

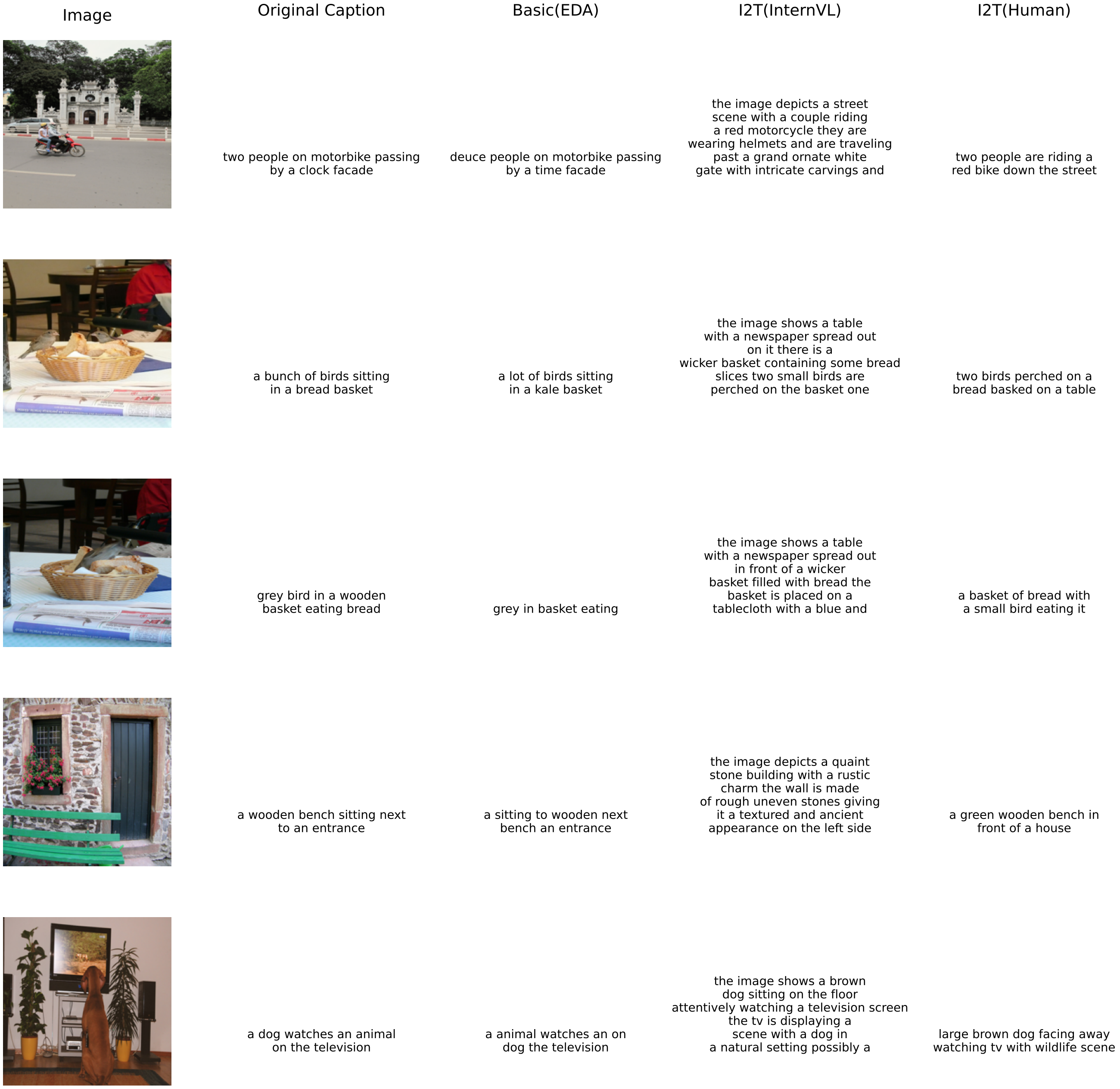


Image	Original Caption	Basic(EDA)	I2T(InternVL)	I2T(Human)
	two people on motorbike passing by a clock facade	deuce people on motorbike passing by a time facade	the image depicts a street scene with a couple riding a red motorcycle they are wearing helmets and are traveling past a grand ornate white gate with intricate carvings and	two people are riding a red bike down the street
	a bunch of birds sitting in a bread basket	a lot of birds sitting in a kale basket	the image shows a table with a newspaper spread out on it there is a wicker basket containing some bread slices two small birds are perched on the basket one	two birds perched on a bread basked on a table
	grey bird in a wooden basket eating bread	grey in basket eating	the image shows a table with a newspaper spread out in front of a wicker basket filled with bread the basket is placed on a tablecloth with a blue and	a basket of bread with a small bird eating it
	a wooden bench sitting next to an entrance	a sitting to wooden next bench an entrance	the image depicts a quaint stone building with a rustic charm the wall is made of rough uneven stones giving it a textured and ancient appearance on the left side	a green wooden bench in front of a house
	a dog watches an animal on the television	a animal watches an on dog the television	the image shows a brown dog sitting on the floor attentively watching a television screen the tv is displaying a scene with a dog in a natural setting possibly a	large brown dog facing away watching tv with wildlife scene

Figure 10. **Visualization of text augmentations (2).**