

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 TOWARDS ADVERSARILY ROBUST CLIP: A HIER- ARCHICAL MODEL FUSION METHOD USING OPTIMAL TRANSPORT

Anonymous authors

Paper under double-blind review

## ABSTRACT

In recent years, multimodal models such as CLIP have achieved impressive performance but remain vulnerable to adversarial perturbations. Although adversarial training can enhance robustness, it often leads to overfitting toward specific attack types. One solution for improving generalization is to integrate multiple diverse and adversarially trained submodels, but this strategy could incur high test-time cost. To achieve a promising tradeoff between robust generalization and efficiency, we consider to design an optimal transport (OT) based model fusion method, which is called “HOT-CLIP (Hierarchical Optimal Transport Fusion for CLIP)”. Although several OT based model fusion methods have been proposed before, they cannot be easily adapted to solve our problem, since they may suffer the issues like parameter misalignment when dealing with highly diverse and multimodal submodels. Our proposed method constructs diverse submodels by varying both attack methods and textual prompts, and integrates them via a hierarchical two-level OT fusion method. The intra-attack fusion first aligns and merges models within the same attack family, and the inter-attack fusion subsequently combines these aligned models across different attacks. Through this carefully crafted fusion strategy, HOT-CLIP can significantly improve the accuracy for alignment and reduce the total occupied memory. More importantly, the obtained robust visual encoder can be deployed without additional inference-time cost. In our experiments, the results on multiple vision-language tasks demonstrate that HOT-CLIP can greatly enhance the model’s adversarial robustness while maintaining competitive clean accuracy.

## 1 INTRODUCTION

The rapid advancement of large vision–language models (LVLMs) (Zhang et al., 2024) has significantly reshaped the landscape of artificial intelligence. By jointly learning from visual and textual modalities, LVLMs demonstrate strong generalization ability across a wide range of downstream tasks, including image classification (Radford et al., 2021), image retrieval (Li et al., 2022), image captioning (Hu et al., 2022), and multimodal reasoning (Yin et al., 2024). A key step in this progress is the development of **alignment** models (Radford et al., 2021; Li et al., 2022). Among them, Contrastive Language–Image Pretraining (CLIP) (Radford et al., 2021) is a representative framework that leverages large-scale contrastive pretraining on image–text pairs to align visual and textual representations effectively.

Although CLIP demonstrates remarkable performance across a wide range of large vision–language tasks, it still faces significant challenges in robustness and reliability. In particular, CLIP’s vision encoder is vulnerable to adversarial perturbations: even small, carefully crafted changes to the input image can induce substantial misalignment between visual and textual representations, ultimately leading to errors in downstream tasks. Prior work suggests that this vulnerability may be partly attributed to the high dimensionality and local linearity of deep visual feature spaces (Goodfellow et al., 2015). This vulnerability is especially concerning in safety-critical domains such as medical imaging (Javed et al., 2024) and autonomous driving (Rossolini et al., 2023), where incorrect predictions may cause severe consequences. Figure 1 provides an example to illustrate the adversarial impact on image captioning performance.

To address these vulnerabilities, various defense strategies have been explored. “Test-time” defenses attempt to mitigate adversarial effects without modifying the training process. One recently proposed approach is CIDEr (Xu et al., 2024), which detects adversarial images by measuring the semantic distance between the original and denoised inputs. Another example, SmoothVLM (Sun et al., 2024) introduces controlled noise to mitigate the effects of localized adversarial perturbations. However, since test-time defenses do not modify model parameters, they are often limited in their ability to address the underlying vulnerabilities of CLIP’s vision encoder.

A more effective strategy is *adversarial training*, where models are explicitly optimized on adversarial examples to improve robustness (Madry et al., 2018; Schlarmann et al., 2024). For example, RobustCLIP (Schlarmann et al., 2024) incorporates adversarial perturbations during training to maintain alignment, while Sim-CLIP (Hossain & Imteaj, 2024b) enforces representation consistency between clean and perturbed samples. However, adversarial training is often prone to overfitting to specific attack patterns and exhibits limited generalization to unseen perturbations (Rice et al., 2020). To overcome the limited generalization of adversarial training, ensemble-based methods (Dong et al., 2020; Hu et al., 2024; Zhang et al., 2025a) have been explored to improve robustness by combining multiple models, which can collectively handle a wider variety of perturbations. However, conventional ensemble methods are challenging to deploy on LVLMs, as these models already impose substantial computational demands (Zhang et al., 2025b), and ensembling will further introduce significant additional memory and inference overhead. Beyond ensembling, **model fusion** techniques (Smith & Gashler, 2017) aim to integrate the parameters of multiple networks rather than merely combining their outputs. The major difference between ensembling and fusion is that ensembling scales inference cost with the number of submodels, while fusion only produces a single model.

Nevertheless, the challenge of standard fusion methods (e.g., weight averaging) (Smith & Gashler, 2017) is the lack of one-to-one correspondence between the parameters from different models. Namely, for two different models, the neurons respectively located in the same position of them may not be functionally similar (a simple example is given in Figure 2 of Section 2). Thus, it is necessary to align the neurons before parameter averaging. **Optimal Transport (OT)** (Peyré & Cuturi, 2019), as formally defined in Section 2, provides a principled metric that quantifies the distance between two probability distributions by computing the minimal cost of transporting one distribution to match the other. In the context of model fusion, OT can be used to compute a transport matrix  $T$  that aligns neurons across models before performing averaging-based fusion (Singh & Jaggi, 2020; Imfeld et al., 2024). Although OT can partially mitigate the parameter misalignment issue, its effectiveness might be constrained when applied to highly diverse models. This is because standard OT establishes correspondences based on geometric distances in parameter space (e.g., Euclidean or cosine), which may not really capture the semantic consistency (Chuang et al., 2023). Consequently, parameters that are close under such geometric measures can still encode distinct underlying features, particularly when the models are trained under different conditions (e.g., different data distributions or architectures).

In summary, applying OT Fusion to adversarially trained CLIP submodels is not straightforward, due to a challenging dilemma. **On the one hand, to ensure model’s robustness, the submodels need to be as diverse as possible**, since submodels trained under different adversarial attacks develop distinct mappings in the input space, resulting in diverse representations. This diversity produces complementary decision boundaries across submodels, collectively enhancing the final fused model’s robustness to a wider range of adversarial perturbations. **On the other hand, to ensure the accuracy of OT Fusion, the submodels need to be as similar as possible**. As explained above, OT primarily aligns parameters based on geometric metrics rather than semantic consistency, which may reduce its effectiveness when the submodels are highly diverse.

**Our contributions.** To achieve a pleasant trade-off between the above two aspects, we propose Hierarchical Optimal Transport Fusion method for CLIP (HOT-CLIP), a novel framework that con-



Figure 1: The *adversarial images* are generated using the PGD attack under the  $\ell_\infty$  norm with  $\epsilon = 2/255$ , and the *mistaken captions* are obtained by applying LLaVA to these perturbed inputs.

108  
 109  
 110  
 111  
 112  
 113  
 114  
 115  
 116  
 117  
 118  
 119  
 120  
 121  
 122  
 123  
 124  
 125  
 126  
 127  
 128  
 129  
 130  
 131  
 132  
 133  
 134  
 135  
 136  
 137  
 138  
 139  
 140  
 141  
 142  
 143  
 144  
 145  
 146  
 147  
 148  
 149  
 150  
 151  
 152  
 153  
 154  
 155  
 156  
 157  
 158  
 159  
 160  
 161

structs diverse adversarial submodels and hierarchically fuses them via optimal transport to produce a single robust CLIP visual encoder. To the best of our knowledge, this is also the first work to investigate model fusion problem in the context of adversarial robustness.

– HOT-CLIP adopts a carefully crafted two-stage procedure, which first constructs diverse adversarial submodels, and then hierarchically fuses them into a single robust visual encoder. Stage 1 constructs a set of diverse submodels by varying both the adversarial attack settings used during training (training data) and the textual prompts (training labels). This diversity ensures that the resulting submodels capture complementary robustness patterns. Stage 2 hierarchically fuses these submodels via optimal transport, effectively balancing the trade-off between submodel diversity and alignment. It first aligns and fuses models within the same attack family (intra-attack), consolidating prompt-induced diversity while ensuring that the models are sufficiently similar for effective OT alignment. It then fuses the resulting models across different attacks (inter-attack), integrating attack-based diversity to produce the final robust model.

– Then, we conduct extensive experiments on three representative tasks for vision-language models, image classification, visual question answering (VQA), and image captioning. Across these tasks, HOT-CLIP consistently enhances adversarial robustness over existing methods, with relative robust score improvements of around 2.6% for image classification, around 20.4% for VQA, and around 16.5% for image captioning. At the same time, its clean performance remains highly competitive. These results demonstrate that HOT-CLIP can effectively navigate the robustness–accuracy trade-off, and has the potential to establish a new state-of-the-art for adversarially robust multimodal models.

## 2 PRELIMINARIES

Due to space constraint, an extended review of related work is presented in the Appendix B. In this section, we briefly introduce the preliminaries related to our study, including the structure of CLIP (Radford et al., 2021), adversarial training (Madry et al., 2018), and optimal transport fusion (Peyré & Cuturi, 2019; Singh & Jaggi, 2020).

Let  $\mathcal{X}$  denote the image space and  $\mathcal{T}$  denote the text space. CLIP consists of two modality-specific encoders: an image encoder  $f_{\theta_{\text{img}}} : \mathcal{X} \rightarrow \mathbb{R}^d$  and a text encoder  $f_{\theta_{\text{txt}}} : \mathcal{T} \rightarrow \mathbb{R}^d$ , where  $\theta$  denotes the model parameters. In particular, we denote the parameters of the image and text encoders as  $\theta_{\text{img}}$  and  $\theta_{\text{txt}}$ , respectively. These encoders map images  $x \in \mathcal{X}$  and text descriptions  $t \in \mathcal{T}$  into a shared  $d$ -dimensional embedding space.

**Definition 2.1 (CLIP Classifier).** Consider a  $K$ -class classification task. Let  $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$  denote the set of candidate classes, and  $\mathcal{Y} = \{1, 2, \dots, K\}$  denote the corresponding label set. For each class  $c_k$ , define a textual prompt  $t_k$  associated with class  $c_k$  (e.g.,  $t_k = \text{"a photo of } c_k\text{"}$ ). The CLIP classifier  $g : \mathcal{X} \rightarrow \mathbb{R}^K$  is defined as

$$g(x)_k = \cos(f_{\theta_{\text{img}}}(x), f_{\theta_{\text{txt}}}(t_k)), \quad k = 1, \dots, K, \quad (1)$$

where  $f_{\theta_{\text{img}}}$  and  $f_{\theta_{\text{txt}}}$  denote the image and text encoders, respectively,  $x \in \mathcal{X}$  is the input image and  $\cos(\cdot, \cdot)$  computes the cosine similarity between normalized embeddings.

**Definition 2.2 (Adversarial Example).** Given a classifier  $g : \mathcal{X} \rightarrow \mathbb{R}^K$  and a clean input image  $x \in \mathcal{X}$  with true label  $y \in \mathcal{Y}$ , an adversarial example is a perturbed input

$$x' = x + \eta, \quad \|\eta\|_p \leq \epsilon, \quad \text{s.t. } \text{argmax } g(x') \neq y, \quad (2)$$

where  $\eta$  is a small perturbation constrained within an  $L_p$ -ball of radius  $\epsilon$ , such that the classifier misclassifies the perturbed input (i.e.  $g(x') \neq y$ ). To obtain such perturbations, adversarial attacks can be categorized into two types:

(i) **Untargeted attack:** the adversary aims to maximize the loss for the true label:

$$\eta^* = \arg \max_{\|\eta\|_p \leq \epsilon} \mathcal{L}(g(x + \eta), y), \quad (3)$$

where  $\mathcal{L}$  is the loss function (e.g. cross-entropy).

(ii) **Targeted attack:** the adversary seeks to misclassify the input as a specific target class  $y_{\text{target}} \neq y$ :

$$\eta^* = \arg \min_{\|\eta\|_p \leq \epsilon} \mathcal{L}(g(x + \eta), y_{\text{target}}). \quad (4)$$

To defend against such adversarial perturbations, adversarial training (Madry et al., 2018) introduces these adversarial examples into the learning process. In our setup, only the parameters of the image encoder are updated, while the text encoder remains frozen. Specifically, the training objective is formulated as the following min–max optimization:

$$\min_{\theta_{img}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \max_{\|\eta\|_p \leq \epsilon} \mathcal{L}(g(x + \eta; \theta_{img}, \theta_{text}), y) \right]. \quad (5)$$

where  $\mathcal{D}$  is the dataset. The inner maximization identifies the most challenging adversarial perturbations within the allowed  $\ell_p$  norm bound, while the outer minimization updates the image encoder to correctly classify these perturbed inputs, thereby enhancing adversarial robustness.

**Definition 2.3** (Optimal Transport Distance (Peyré & Cuturi, 2019)). *Let  $\mu = \sum_{i=1}^n \alpha_i \delta(a^{(i)})$  and  $\nu = \sum_{j=1}^m \beta_j \delta(b^{(j)})$  be two empirical probability measures, where  $a^{(i)} \in \mathcal{P}$  and  $b^{(j)} \in \mathcal{Q}$  are support points, and  $\delta(\cdot)$  denotes the Dirac measure assigning unit mass. Here,  $\mathcal{P}$  and  $\mathcal{Q}$  represent the spaces of source and target points (e.g., neuron embeddings to be aligned). We define a transport cost function  $C : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R}^+$ , which quantifies the cost of transporting unit mass from  $a^{(i)}$  to  $b^{(j)}$ . The optimal transport distance between  $\mu$  and  $\nu$  is defined as*

$$OT(\mu, \nu) = \min_{T \in \Pi(\mu, \nu)} \mathbb{E}_{(a,b) \sim T} [C(a, b)], \quad (6)$$

where  $\Pi(\mu, \nu)$  denotes the set of couplings with marginals  $\mu$  and  $\nu$ .

In the above definition, the minimizer  $T^* \in \Pi(\mu, \nu)$ , called the *optimal transport plan*, defines a minimal-cost correspondence between the support points of  $\mu$  and  $\nu$ . In the context of model fusion,  $T^*$  can be used to align neurons or parameter vectors across models, providing a principled way to combine them while minimizing misalignment. Based on optimal transport, OT Fusion (Singh & Jaggi, 2020) aligns two (or more) neural networks in a layer-wise manner. In each layer, neurons are treated as points ( $a$  and  $b$ ), with their associated weights or activations serving as feature representations. Assuming a uniform probability measure over neurons ( $\mu$  and  $\nu$ ), the OT problem is solved between corresponding layers to obtain the transport matrix. The transport matrix is then used to align the current layer, and the aligned weights are subsequently averaged to produce the fused layer. Applying this procedure sequentially across layers yields a coherent fusion of the models. In Figure 2, we provide a simple two-layer example, and the full details of OT Fusion are shown in Appendix C.1.

### 3 METHODOLOGY

In this section, we first present the high-level idea of our method in Section 3.1, and then detail the technical components, including the construction of diverse submodels and the hierarchical OT Fusion in Sections 3.2 and 3.3, respectively.

#### 3.1 OVERVIEW OF OUR METHOD

Our goal is to enhance the robustness of CLIP’s visual encoder against adversarial perturbations. The key idea is to leverage diversity among adversarially trained submodels and integrate them through OT Fusion. Figure 3 illustrates the overall framework.

**Stage 1: Diverse Adversarial Submodels Construction.** According to the adversarial training objective introduced in Definition 2.2, the parameters of a model are influenced by multiple factors, such as the training data, model architecture, and optimization algorithm. Among these, training

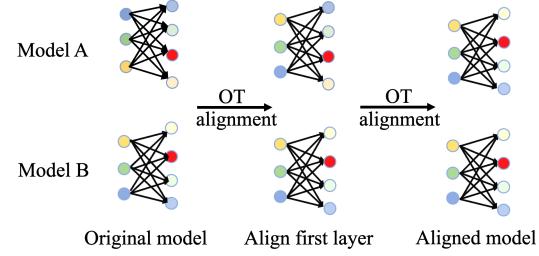


Figure 2: Neuron alignment via OT. Assume the neurons with same color are functionally similar. The original models “A” and “B” exhibit a permutation in neuron correspondence (left figure); for example, the first neuron in the first layer of A is blue, but the neuron in the same position of B is yellow. In the middle figure, we align the first layer via OT; then, the second layer is aligned in the right figure.

Figure 2: Neuron alignment via OT. Assume the neurons with same color are functionally similar. The original models “A” and “B” exhibit a permutation in neuron correspondence (left figure); for example, the first neuron in the first layer of A is blue, but the neuron in the same position of B is yellow. In the middle figure, we align the first layer via OT; then, the second layer is aligned in the right figure.

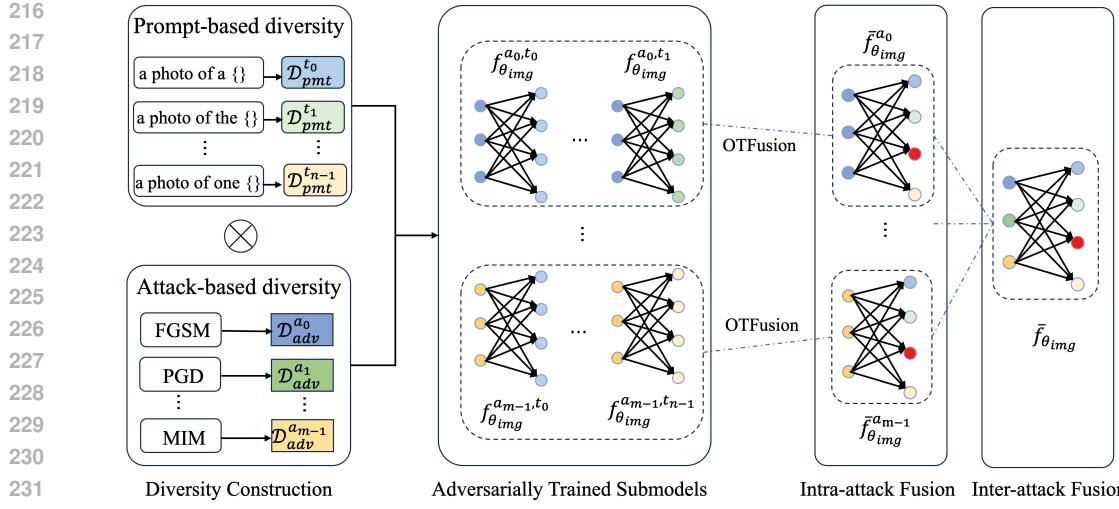


Figure 3: Overview of HOT-CLIP. Diverse submodels are first constructed along two axes: prompt-based diversity ( $n$  different textual templates) and attack-based diversity ( $m$  different adversarial attacks). The hierarchical OT Fusion method is then applied to integrate these  $n \times m$  submodels. In the Intra-attack OT Fusion stage, submodels within the same attack family but with different prompts are aligned and fused via optimal transport. In the Inter-attack OT Fusion stage, the first-level fused models from different attack families are further integrated, yielding the final robust visual encoder.

data plays a particularly crucial role, as variations in data directly affect the learned decision boundaries and representations. In our study, all submodels share the same architecture and optimization algorithm, and the **diversity** arises solely from the differences between those specifically “modified” data distributions. Concretely, we train the submodels on adversarial examples generated by different attack algorithms, so that the desired robustness can be diversified across multiple perturbation types. Additionally, a single text template is often insufficient to fully capture the alignment between images and textual labels; for example, the original CLIP (Radford et al., 2021) used about 80 templates. To further enhance the diversity, we vary both the adversarial attack settings and the textual prompts, so that the submodels are trained under different distributions of adversarial examples and supervision.

**Stage 2: Hierarchical OT Fusion.** Directly fusing these diverse adversarially trained CLIP submodels may be suboptimal. Theoretically, the parameters that are close under geometric measures can encode distinct underlying features, particularly for diversely trained submodels (Chuang et al., 2023; Nguyen et al., 2023; Ormaniec et al., 2025). Even using OT (as illustrated in Figure 2), the highly diverse submodels could still yield non-ignorable error for the final fusion. Moreover, simultaneously fusing all submodels can take a rather large amount of memory space, as a great number of parameters need to be stored and aligned within the same period during the fusion. To neatly circumvent “direct” OT Fusion, we propose a hierarchical two-level OT Fusion framework. The key idea is to control both the similarity and the number of submodels involved at each fusion step, thereby improving alignment quality while keeping memory usage manageable. At the first level, submodels trained under the same adversarial attack but with different textual prompts are grouped and fused via OT. This ensures high internal homogeneity within each group and limits the number of models fused simultaneously. At the second level, the first-level fused models, already aligned in the label space, are further integrated across different attacks. By progressively fusing more homogeneous submodels, this hierarchical design can effectively preserve parameter alignment, leverages complementary robustness, and maintains an affordable memory size.

### 3.2 CONSTRUCTION OF DIVERSE ADVERSARIAL SUBMODELS

We construct a set of diverse submodels by varying both the attack settings used during adversarial training and the textual prompts used for image–text alignment.

**Attack-based diversity.** As discussed in Section 3.1, adversarial training typically improves robustness against the specific perturbations seen during training but generalizes poorly to unseen attacks. To mitigate this limitation, we construct submodels under different adversarial settings. Formally,

270 let  $\mathcal{A}$  denote a set of adversarial attack methods. Given a clean dataset  $\mathcal{D}$  and an attack method  
 271  $a \in \mathcal{A}$ , we define the attack-specific dataset as  
 272

$$273 \quad \mathcal{D}_{\text{adv}}^{(a)} = \{(x + \eta_a(x), y) \mid (x, y) \in \mathcal{D}\} \quad (7)$$

275 where  $\eta_a$  denotes the perturbation generated under method  $a$  (e.g., FGSM (Goodfellow et al., 2015),  
 276 PGD (Madry et al., 2018), MIM (Dong et al., 2018)) with its own radius  $\epsilon$  and norm constraint.  
 277 Training a visual encoder on  $\mathcal{D}_{\text{adv}}^{(a)}$  yields a submodel specialized to adversarial perturbations of  $a$ .

278 **Prompt-based diversity.** Beyond attack-based diversity, CLIP-style models also rely on text–image  
 279 alignment, which introduces another source of diversity. We consider multiple textual prompt sets  
 280  $\mathcal{T}$ . For a given prompt set  $t \in \mathcal{T}$ , the prompt-specific dataset is defined as  
 281

$$282 \quad \mathcal{D}_{\text{prompt}}^{(t)} = \{(x, y^{(t)}) \mid (x, y) \in \mathcal{D}\}, \quad (8)$$

283 where  $y^{(t)}$  is the textual representation for class  $y$  under prompt  $t$ . This generates the submodels  
 284 with different label alignment characteristics in the embedding space.  
 285

286 For each combination of attack  $a \in \mathcal{A}$  and prompt  $t \in \mathcal{T}$ , we define a fully diversified dataset that  
 287 integrates  $\mathcal{D}_{\text{adv}}^{(a)}$  and  $\mathcal{D}_{\text{prompt}}^{(t)}$ :

$$289 \quad \mathcal{D}_{\text{div}}^{(a,t)} = \{(x + \eta_a(x), y^{(t)}) \mid (x, y) \in \mathcal{D}\}. \quad (9)$$

290 Then, we train a submodel  $f_{\theta_{\text{img}}}^{(a,t)}$  on  $\mathcal{D}_{\text{div}}^{(a,t)}$ , where only the parameters  $\theta_{\text{img}}$  of the image encoder  
 291 are updated and the text encoder  $\theta_{\text{text}}$  remains fixed. The resulting family of submodels is  $\mathcal{M} =$   
 292  $\{f_{\theta_{\text{img}}}^{(a,t)} \mid a \in \mathcal{A}, t \in \mathcal{T}\}$ .  
 293

### 295 3.3 HIERARCHICAL OT FUSION

296 We then introduce the two-level fusion procedure, consisting of the L-level (Language-level) fusion  
 297 and the V-level (Visual-level) fusion. At the L-level, submodels trained under the same adversarial  
 298 attack but using different textual prompts are fused. This step consolidates the diversity arising  
 299 from multiple prompts. At the V-level, the resulting L-level fused models are further fused across  
 300 different adversarial attacks, integrating the diversity introduced by varying attack methods. Due to  
 301 the space limit, we leave the full HOT-CLIP algorithm to Appendix C.  
 302

303 **L-Level: Intra-attack Fusion.** For a fixed attack  $a \in \mathcal{A}$ , the set of submodels  $\mathcal{M}_a = \{f_{\theta_{\text{img}}}^{(a,t)} \mid t \in \mathcal{T}\}$   
 304 share similar robustness properties but differ in feature alignment due to different textual prompts.  
 305 The L-level fused model  $\bar{f}_{\theta_{\text{img}}}^{(a)}$  is obtained by averaging the weights of submodels, after aligning them  
 306 to an “anchor” model (which can be any arbitrary submodel selected from  $\mathcal{M}_a$ ). Formally, for each  
 307 layer  $\ell$ , the fused weight is  
 308

$$310 \quad W_{\ell}^{(a)} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} W_{\ell}^{(a,t), \text{aligned}}, \quad (10)$$

312 where  $W_{\ell}^{(a,t), \text{aligned}}$  denotes the aligned weights of submodel  $f_{\theta_{\text{img}}}^{(a,t)}$ . To obtain the aligned weights  
 313  $W_{\ell}^{(a,t), \text{aligned}}$ , we first align the columns of the current layer’s weight  $W_{\ell}^{(a,t)}$  using the transport  
 314 matrix from the previous layer,  $T_{\ell-1}^{(a,t)}$ . These partially aligned weights are then used to compute  
 315 the transport matrix  $T_{\ell}^{(a,t)}$  for the current layer, which is subsequently applied to align the rows of  
 316  $W_{\ell}^{(a,t)}$ , yielding the fully aligned weights  $W_{\ell}^{(a,t), \text{aligned}}$ . Formally,  $W_{\ell}^{(a,t), \text{aligned}}$  is calculated by  
 317

$$319 \quad W_{\ell}^{(a,t), \text{aligned}} = T_{\ell}^{(a,t)\top} W_{\ell}^{(a,t)} T_{\ell-1}^{(a,t)}, \quad (11)$$

321 The right multiplication by  $T_{\ell-1}^{(a,t)}$  aligns the columns of  $W_{\ell}^{(a,t)}$  to the anchor model, while the left  
 322 multiplication by  $T_{\ell}^{(a,t)\top}$  aligns the rows with respect to the anchor model. For more details, please  
 323 see Appendix C.1.

324 **V-Level: Inter-attack Fusion.** After intra-attack fusion, we obtain a set of fused models corresponding to each attack method, denoted as  $\{\bar{f}_{\theta_{\text{img}}}^{(a)} \mid a \in \mathcal{A}\}$ . We then apply OT Fusion to these  
 325 models. Specifically, we compute the OT matrix  $T^{(a)}$  between the layers of each submodel and  
 326 an arbitrarily selected anchor from  $\{\bar{f}_{\theta_{\text{img}}}^{(a)} \mid a \in \mathcal{A}\}$ . For a given layer  $\ell$ , the aligned weights are  
 327 obtained as  
 328

$$W_{\ell}^{(a),\text{aligned}} = T_{\ell}^{(a)\top} W_{\ell}^{(a)} T_{\ell-1}^{(a)}, \quad (12)$$

329 where  $W_{\ell}^{(a)}$  denotes the weight matrix of layer  $\ell$  in  $\bar{f}_{\theta_{\text{img}}}^{(a)}$ ,  $T_{\ell-1}^{(a)}$  is the transport map from the previous  
 330 layer, and  $T_{\ell}^{(a)\top}$  is the transpose of the current layer's transport map. Similar to the L-level fusion,  
 331 the right multiplication by  $T_{\ell-1}^{(a)}$  aligns the columns of  $W_{\ell}^{(a)}$  with the anchor model, while the left  
 332 multiplication by  $T_{\ell}^{(a)\top}$  aligns the rows. The aligned weights are then averaged across all attack-  
 333 specific models:  
 334

$$W_{\ell}^{\text{fused}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} W_{\ell}^{(a),\text{aligned}}. \quad (13)$$

335 Repeating this procedure for all layers yields the final robust model  $\bar{f}_{\theta_{\text{img}}}$ , which integrates both  
 336 prompt-level and attack-level diversity.  
 337

338 **Memory usage during the fusion.** By controlling the diversity at each step, this hierarchical  
 339 approach could ensure that only relatively similar submodels are aligned at a time. Specifically, the  
 340 peak memory size occupied by the submodels is lowered from  $\mathcal{O}(|\mathcal{A}||\mathcal{T}| \cdot U)$  for naive global fusion  
 341 to  $\mathcal{O}(\max\{|\mathcal{A}|, |\mathcal{T}|\} \cdot U)$ , where  $U$  denotes the memory usage of a single submodel; the detailed  
 342 analysis is provided in Appendix C.3.  
 343

344 To illustrate some theoretical intuition for why the hierarchical fusion behaves reasonably, we in-  
 345 clude the Lemma 3.1, which provides an upper bound on the distance between the hierarchical fused  
 346 center and the global Wasserstein barycenter (Aguech & Carlier, 2011). This bound suggests that the  
 347 hierarchical fused center remains controlled by the average distance of the individual submodels to  
 348 the global barycenter. We add the proof in Appendix C.2.  
 349

350 **Lemma 3.1.** *Let  $\mu_{a,t}$  denote the submodels trained under attack type  $a$  and prompt variant  $t$ . Let  $\mu_a^*$   
 351 be the intra-attack OT barycenter,  $\mu_{\text{hier}}^*$  the hierarchical barycenter obtained from  $\{\mu_a^*\}$ , and  $\mu_{\text{global}}^*$   
 352 the global Wasserstein barycenter computed over all submodels. Let  $W_c$  denote the Wasserstein  
 353 distance. Then*

$$W_c(\mu_{\text{hier}}^*, \mu_{\text{global}}^*) \leq \frac{4}{|\mathcal{AT}|} \sum_{a,t} W_c(\mu_{a,t}, \mu_{\text{global}}^*).$$

## 4 EXPERIMENTS

360 In this section, we evaluate the adversarial robustness of CLIP's visual encoder enhanced with our  
 361 hierarchical OT Fusion (HOT-CLIP) framework. We conduct experiments on three representative  
 362 multimodal tasks (zero-shot image classification (Radford et al., 2021), visual question answering  
 363 (VQA) (Goyal et al., 2017), and image captioning (Hu et al., 2022)) under diverse adversarial attack  
 364 scenarios.  
 365

### 4.1 EXPERIMENTAL SETUP

366 **Implementation Details.** We adopt a two-stage training pipeline that combines the construction  
 367 of diverse adversarial submodels and hierarchical OT fusion, using CLIP (Radford et al., 2021)  
 368 with ViT-L/14 visual encoders as the backbone. First, we construct diverse adversarial submodels  
 369 by training CLIP for two epochs on ImageNet, where each submodel is adversarially trained  
 370 under a specific attack method (e.g., FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018),  
 371 MIM (Dong et al., 2018)) and with a textual prompt randomly sampled from the 80 templates intro-  
 372 duced in (Radford et al., 2021). Next, we perform the hierarchical OT fusion. Within each attack  
 373 family, submodels trained with different textual prompts are fused via OT, where layer-wise weight  
 374 matrices are represented as neuron embeddings and the Sinkhorn algorithm (Cuturi, 2013) is used  
 375 to compute the transport plan. The resulting fused model is fine-tuned for one epoch under the cor-  
 376 responding attack setting with a standard text template (“a photo of a ...”). Finally, the first-level  
 377

**Table 1: Clean and adversarial evaluation on image classification datasets of CLIP model.**  
 Models are trained on ImageNet, all other datasets are zero-shot. Robustness is assessed using AutoAttack with the  $l_\infty$  norm and perturbation bound  $\epsilon = 2/255$  and  $\epsilon = 4/255$ . The last column shows the average accuracy across datasets. Bold indicates the best performance in each column.

Eval.	Vision encoder	Zero-shot datasets													Avg.	
		ImageNet	CaTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	
clean	CLIP	74.9	82.6	78.6	95.6	72.7	55.7	63.5	33.3	79.8	87.7	58.6	52.2	92.1	99.6	<b>73.3</b>
	TeCoA	77.4	77.3	35.3	80.6	51.0	39.5	24.0	13.2	40.5	73.1	54.5	49.8	77.3	93.9	56.2
	FARE	69.2	84.0	63.1	76.7	57.2	44.0	20.2	23.3	57.1	80.9	57.2	50.2	87.5	96.7	61.9
	<b>HOT-CLIP</b>	69.9	85.1	64.0	78.7	60.0	46.2	18.0	21.8	56.9	80.8	59.9	49.6	87.1	96.3	62.5
$\epsilon = 2/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	TeCoA	62.5	70.0	17.5	60.9	34.0	27.2	14.4	5.7	24.1	58.8	44.0	47.2	68.3	87.4	44.4
	FARE	52.1	76.8	29.8	56.5	36.2	28.4	12.2	8.0	28.3	61.0	41.9	50.2	71.5	89.7	45.9
	<b>HOT-CLIP</b>	53.4	79.5	31.8	58.3	37.2	31.8	12.4	8.1	29.7	60.9	44.6	49.6	71.2	91.0	<b>47.1</b>
$\epsilon = 4/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	TeCoA	48.2	61.4	8.7	37.3	20.2	17.6	11.6	2.3	12.5	41.5	34.5	38.1	55.7	74.6	33.1
	FARE	33.0	64.6	12.5	34.5	20.5	17.0	11.2	2.0	12.3	40.4	31.3	50.2	50.5	74.6	32.4
	<b>HOT-CLIP</b>	34.7	66.5	15.8	38.1	20.9	19.6	10.3	2.9	12.5	39.4	32.8	49.6	51.8	77.0	<b>33.7</b>

fused models from different attacks are integrated through OT Fusion, followed by an additional one epoch of unsupervised adversarial fine-tuning. Further details for hyperparameters and implementation are deferred to Appendix C.1.

**Baselines.** We compare our method against recent state-of-the-art approaches that aim to improve the adversarial robustness of vision–language models: TeCoA (Mao et al., 2023), which provides a systematic analysis of adversarial robustness in CLIP-like models and proposes tailored fine-tuning strategies to improve zero-shot performance under adversarial settings. FARE (Schlarmann et al., 2024), which introduces adversarial perturbations directly into the visual embedding space and fine-tunes the image encoder in an unsupervised manner.

## 4.2 EVALUATION ON PERFORMANCE

**Zero-shot Image Classification.** We evaluate clean and robust accuracies of the CLIP models on ImageNet and 13 zero-shot datasets mentioned in Appendix D. For each dataset, class names are combined with a predefined set of prompt templates. Zero-shot classification is then performed as described in Definition 2.1. To evaluate the adversarial robustness of the models, we adopt AutoAttack (Croce & Hein, 2020) under the  $\ell_\infty$  norm with perturbation radii of  $\epsilon = 2/255$  and  $\epsilon = 4/255$ , each run for 100 iterations. As shown in Table 1, HOT-CLIP achieves the second-best clean accuracy among all methods, slightly lower than the original CLIP. Under adversarial perturbations, it achieves a relative improvement in robustness of 2.6% at  $\epsilon = 2/255$  and 1.8% at  $\epsilon = 4/255$ . These results indicate that our method improves adversarial robustness while maintaining competitive clean accuracy on image classification.

**Image Captioning.** We further evaluate our method on image captioning, where the model generates natural language descriptions of images. We report the results using the CIDEr score (Vedantam et al., 2015), a widely adopted metric for measuring the quality of generated captions. The experiments are conducted on COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014), using two representative LVLMs: OpenFlamingo-9B (OF) (Alayrac et al., 2022) and LLaVA-1.5-7B (Liu et al., 2023). For clean evaluation, we use the full validation sets; for adversarial evaluation, we randomly sample 500 images from each dataset, using a similar evaluation setup as Schlarmann & Hein (2023). The adversarial robustness is tested with AutoAttack (Croce & Hein, 2020) under the  $\ell_\infty$  norm with  $\epsilon \in \{2/255, 4/255\}$ , using 100 iterations. As shown in Table 2, HOT-CLIP consistently improves robustness across both datasets. For LLaVA-7B, HOT-CLIP achieves a relative

432  
 433 **Table 2: Evaluation of LVLMs using different CLIP encoders on image captioning.** Results are  
 434 reported for OpenFlamingo and LLaVA on two image captioning datasets, measured using CIDEr.  
 435 The last column shows the average CIDEr score across datasets. Bold indicates the best performance  
 436 in each column.

437 VLM	438 Vision 439 encoder	440 COCO			441 Flickr30k			442 Average over datasets		
		443 clean	444 $\epsilon = \frac{2}{255}$	445 $\epsilon = \frac{4}{255}$	446 clean	447 $\epsilon = \frac{2}{255}$	448 $\epsilon = \frac{4}{255}$	449 clean	450 $\epsilon = \frac{2}{255}$	451 $\epsilon = \frac{4}{255}$
452 LLaVA-7B	CLIP	122.2	3.2	2.4	79.1	1.4	0.9	<b>100.6</b>	2.3	1.65
	TeCoA	93.9	40.8	16.9	50.9	26.3	16.5	72.4	33.5	16.7
	FARE	105.8	50.1	33.2	64.7	28.5	20.1	85.2	39.3	26.6
	<b>HOT-CLIP</b>	110.4	56.5	35.5	74.6	43.1	26.5	92.5	<b>49.8</b>	<b>31.0</b>
453 OF-9B	CLIP	85.2	1.6	1.3	63.8	0.6	0.5	<b>74.5</b>	1.1	0.9
	TeCoA	73.5	31.6	21.2	43.5	10.4	10.2	58.5	21.0	15.7
	FARE	81.5	33.2	22.8	54.6	16.1	10.5	68.0	24.6	16.6
	<b>HOT-CLIP</b>	87.9	36.7	26.0	55.2	19.1	11.7	71.5	<b>27.9</b>	<b>18.8</b>

452 **Table 3: Evaluation of LVLMs using different CLIP encoders on VQA.** Results are reported on  
 453 VQAv2 and TextVQA, measured by accuracy. The last column shows the average accuracy across  
 454 datasets. Bold indicates the best performance in each column.

455 VLM	456 Vision 457 encoder	458 TextVQA			459 VQAv2			460 Average over datasets		
		461 clean	462 $\epsilon = \frac{2}{255}$	463 $\epsilon = \frac{4}{255}$	464 clean	465 $\epsilon = \frac{2}{255}$	466 $\epsilon = \frac{4}{255}$	467 clean	468 $\epsilon = \frac{2}{255}$	469 $\epsilon = \frac{4}{255}$
470 LLaVA-7B	CLIP	37.8	0.2	0.0	72.4	2.6	0.2	<b>55.1</b>	1.4	0.1
	TeCoA	19.4	12.8	8.9	63.4	40.4	29.5	41.4	26.6	19.2
	FARE	27.5	15.4	9.1	65.6	40.9	29.7	46.5	28.1	19.4
	<b>HOT-CLIP</b>	25.3	17.7	12.8	68.8	43.8	34.6	47.0	<b>30.7</b>	<b>23.7</b>
471 OF-9B	CLIP	21.0	0.0	0.0	46.2	3.7	0.5	33.6	1.9	0.2
	TeCoA	12.4	2.9	1.8	45.6	25.5	22.3	29.0	14.2	12.1
	FARE	17.0	3.5	2.6	43.2	24.0	20.7	30.1	13.7	11.6
	<b>HOT-CLIP</b>	22.5	5.2	5.0	51.5	29.1	24.0	<b>37.0</b>	<b>17.1</b>	<b>14.5</b>

472 improvement of 26.7% at  $\epsilon = 2/255$  and 16.5% at  $\epsilon = 4/255$ ; for OF-9B, the corresponding relative  
 473 gains are 13.4% and 13.3%. These results demonstrate that HOT-CLIP effectively enhances  
 474 adversarial robustness in image captioning while preserving competitive clean performance.

475 **VQA.** We also evaluate our method on the task of Visual Question Answering, where the model is  
 476 required to provide accurate answers to natural language questions based on image inputs. For  
 477 evaluation, we consider two widely used VQA benchmarks: VQAv2 (Goyal et al., 2017) and  
 478 TextVQA (Singh et al., 2019). Adversarial evaluation uses the same model and attack settings as  
 479 in the image captioning experiments, i.e., OpenFlamingo-9B and LLaVA-1.5-7B, with AutoAttack  
 480 under the  $\ell_\infty$  norm ( $\epsilon = 2/255$  and  $4/255$ , 100 iterations). Table 3 reports the VQA accuracy under  
 481 clean and adversarial settings, showing that HOT-CLIP consistently enhances robustness across  
 482 both benchmarks while maintaining competitive clean performance. For LLaVA-7B, HOT-CLIP  
 483 achieves a relative improvement of 9.2% at  $\epsilon = 2/255$  and 22.2% at  $\epsilon = 4/255$ ; for OF-9B, the  
 484 corresponding relative gains are 20.4% and 19.8%.

485 **Summary of other experimental results.** Due to space constraint, additional ablation studies and  
 486 extended evaluations are provided in Appendix E. We conduct ablation studies to examine the effects  
 487 of backbone choice, submodel pool composition, adversarial training strength, and fusion strategies.  
 488 We further evaluate our method on additional tasks, including targeted attacks and hallucination phe-  
 489 nomena in large vision-language models. The results demonstrate that our framework consistently  
 490 enhances adversarial robustness across tasks and configurations, while remaining effective across  
 491 different model architectures.

486

487

Table 4: Comparison of training cost, fusion efficiency, and memory usage.

488

489

490

491

492

493

494

495

we have added a detailed comparison of training time, fusion time, and memory usage across baselines, including HOT-CLIP, Direct OT Fusion, and FARE. All GPU hours are measured on NVIDIA RTX 4090 GPUs. The results are summarized in Table 4.

496

497

498

499

500

501

502

503

504

505

In the training phase, our method involves fine-tuning 9 sub-models for 2 epochs on 1M images with 10-step adversarial attacks (e.g., PGD). This represents only about 1.4% of the computational cost of training the original CLIP model (trained for 32 epochs on 400M images). Moreover, the training of all sub-models is fully parallelizable, so with sufficient resources, the wall-clock time can be reduced to that of training a single sub-model. We believe this one-time training investment is both manageable and justified, as it produces a single and robust model that supports efficient inference.

506

507

508

509

510

511

512

513

In the fusion phase, the hierarchical structure itself acts as an optimization, effectively reducing memory requirements from  $O(|\mathcal{A}||\mathcal{T}| \cdot U)$  to  $O(\max\{|\mathcal{A}|, |\mathcal{T}|\} \cdot U)$  (the details are analyzed in Appendix C.3). We also employ the Sinkhorn algorithm to efficiently approximate the optimal transport solutions, significantly accelerating the computation while maintaining stability. In our experiments, each fusion stage—either intra-attack or inter-attack—typically completes in approximately 20 minutes, with peak memory usage reduced from 641GB to 178GB. We consider this computational overhead to be reasonable in practice, given the resulting robustness and inference efficiency.

514

515

## 5 CONCLUSION AND FUTURE WORK

516

The proposed HOT-CLIP framework enhances the adversarial robustness of large vision-language models by constructing diverse submodels across different adversarial attacks and textual prompts, and integrating them via hierarchical OT Fusion. Experiments on zero-shot image classification, VQA, and image captioning demonstrate consistent improvements in robustness under strong adversarial perturbations, while preserving competitive performance on clean data. Although leveraging multiple submodels increases computational cost during training, the hierarchical fusion ensures that inference-time overhead remains comparable to a single backbone, making the method practical for deployment. We think a promising direction for future work is to extend HOT-CLIP towards cross-modal joint defenses, examining the potential of hierarchical fusion for the simultaneous enhancement of robustness across multiple modalities.

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540 REFERENCES  
541

542 Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on*  
543 *Mathematical Analysis*, 43(2):904–924, 2011.

544 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel  
545 Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language  
546 model for few-shot learning. *Advances in neural information processing systems*, 35:23716–  
547 23736, 2022.

548 Yuki Arase, Han Bao, and Sho Yokoi. Unbalanced optimal transport for unbalanced word alignment.  
549 In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual*  
550 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3966–  
551 3986, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.219. URL <https://aclanthology.org/2023.acl-long.219/>.

552 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
553 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

554 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong  
555 Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning  
556 for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer*  
557 *vision and pattern recognition*, pp. 24185–24198, 2024.

558 Ching-Yao Chuang, Stefanie Jegelka, and David Alvarez-Melis. Infoot: Information maximizing  
559 optimal transport. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt,  
560 Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML*  
561 *2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learn-*  
562 *ing Research*, pp. 6228–6242. PMLR, 2023. URL <https://proceedings.mlr.press/v202/chuang23a.html>.

563 Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. De-  
564 scribing textures in the wild. In *Proceedings of the IEEE conference on computer vision and*  
565 *pattern recognition*, pp. 3606–3613, 2014.

566 Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised  
567 feature learning. In *Proceedings of the fourteenth international conference on artificial intelli-*  
568 *gence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

569 Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble  
570 of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–  
571 2216. PMLR, 2020.

572 Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural*  
573 *information processing systems*, 26, 2013.

574 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hi-  
575 erarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,  
576 pp. 248–255. Ieee, 2009.

577 Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li.  
578 Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision*  
579 *and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 9185–  
580 9193. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.  
581 2018.00957. URL [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Dong\\_Boosting\\_Adversarial\\_Attacks\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html).

582 Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training  
583 for robust deep learning. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina  
584 Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: An-*  
585 *nual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12,*  
586 *2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/5de8a36008b04a6167761fa19b61aa6c-Abstract.html>.

594 Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training  
 595 examples: An incremental bayesian approach tested on 101 object categories. *Computer vision*  
 596 and *Image understanding*, 106(1):59–70, 2007.

597

598 Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip H. S.  
 599 Torr. Large-scale unsupervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*,  
 600 45(6):7457–7476, 2023. doi: 10.1109/TPAMI.2022.3218275. URL <https://doi.org/10.1109/TPAMI.2022.3218275>.

601

602 Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial  
 603 examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning  
 604 Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceed-  
 605 ings*, 2015. URL <http://arxiv.org/abs/1412.6572>.

606

607 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa  
 608 matter: Elevating the role of image understanding in visual question answering. In *Proceedings  
 609 of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

610

611 Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset  
 612 and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected  
 613 Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

614

615 Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul  
 616 Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer.  
 617 The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021  
 618 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada,  
 October 10-17, 2021*, pp. 8320–8329. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00823. URL  
<https://doi.org/10.1109/ICCV48922.2021.00823>.

619

620 Md Zarif Hossain and Ahmed Imteaj. Securing vision-language models with a robust encoder  
 621 against jailbreak and adversarial attacks. In *2024 IEEE International Conference on Big Data  
 622 (BigData)*, pp. 6250–6259. IEEE, 2024a.

623

624 Md Zarif Hossain and Ahmed Imteaj. Sim-clip: Unsupervised siamese adversarial fine-tuning for  
 625 robust and semantically-rich vision-language models. *arXiv preprint arXiv:2407.14971*, 2024b.

626

627 Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang.  
 628 Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF  
 629 conference on computer vision and pattern recognition*, pp. 17980–17989, 2022.

630

631 Zhaozhe Hu, Jia-Li Yin, Bin Chen, Luojun Lin, Bo-Hao Chen, and Ximeng Liu. MEAT: median-  
 632 ensemble adversarial training for improving robustness and generalization. In *IEEE International  
 633 Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea,  
 April 14-19, 2024*, pp. 5600–5604. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446117.  
 634 URL <https://doi.org/10.1109/ICASSP48485.2024.10446117>.

635

636 Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and  
 637 Sidak Pal Singh. Transformer fusion with optimal transport. In *The Twelfth International Con-  
 638 ference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenRe-  
 639 view.net, 2024. URL <https://openreview.net/forum?id=LjeqMvQopen>.

640

641 Haseeb Javed, Shaker El-Sappagh, and Tamer Abuhmed. Robustness in deep learning models for  
 642 medical diagnostics: security and adversarial challenges towards robust ai applications. *Artificial  
 643 Intelligence Review*, 58(1):12, 2024.

644

645 Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained  
 646 categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV  
 647 Workshops 2013, Sydney, Australia, December 1-8, 2013*, pp. 554–561. IEEE Computer Society,  
 648 2013. doi: 10.1109/ICCVW.2013.77. URL <https://doi.org/10.1109/ICCVW.2013.77>.

649

650 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.  
 651 2009.

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

651 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image  
 652 pre-training with frozen image encoders and large language models. In *International conference*  
 653 *on machine learning*, pp. 19730–19742. PMLR, 2023a.

654 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evalu-  
 655 ating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino,  
 656 and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural*  
 657 *Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Associa-  
 658 tion for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL  
 659 <https://doi.org/10.18653/v1/2023.emnlp-main.20>.

660 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr  
 661 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European*  
 662 *conference on computer vision*, pp. 740–755. Springer, 2014.

663 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances*  
 664 *in neural information processing systems*, 36:34892–34916, 2023.

665 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International*  
 666 *Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.  
 667 OpenReview.net, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.

668 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.  
 669 Towards deep learning models resistant to adversarial attacks. In *6th International Conference*  
 670 *on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018,*  
 671 *Conference Track Proceedings*. OpenReview.net, 2018. URL <https://openreview.net/forum?id=rJzIBfZAb>.

672 Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained  
 673 visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

674 Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-  
 675 shot adversarial robustness for large-scale models. In *The Eleventh International Conference on*  
 676 *Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.  
 677 URL <https://openreview.net/forum?id=P4bXCawRi5J>.

678 Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail  
 679 Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference align-  
 680 ment of LLMs via optimal transport. In *The Thirty-eighth Annual Conference on Neural*  
 681 *Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=2LctgfN6Ty>.

682 Dang Nguyen, Trang Nguyen, Khai Nguyen, Dinh Q. Phung, Hung Hai Bui, and Nhat Ho. On  
 683 cross-layer alignment for model fusion of heterogeneous neural networks. In *IEEE Interna-  
 684 tional Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island,*  
 685 *Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10097156.  
 686 URL <https://doi.org/10.1109/ICASSP49357.2023.10097156>.

687 Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large num-  
 688 ber of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*,  
 689 *ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pp. 722–729. IEEE Computer Soci-  
 690 ety, 2008. doi: 10.1109/ICVGIP.2008.47. URL [https://doi.org/10.1109/ICVGIP.](https://doi.org/10.1109/ICVGIP.2008.47)  
 691 2008.47.

692 Weronika Ormaniec, Michael Vollenweider, and Elisa Hoskovec. Fusion of graph neural networks  
 693 via optimal transport. *arXiv preprint arXiv:2503.21579*, 2025.

694 Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012*  
 695 *IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

702 Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*,  
 703 11(5-6):355–607, 2019. doi: 10.1561/2200000073. URL <https://doi.org/10.1561/2200000073>.

704

705 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL <http://proceedings.mlr.press/v139/radford21a.html>.

706

707

708

709

710

711

712

713 Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pp. 8093–8104. PMLR, 2020.

714

715 Giulio Rossolini, Federico Nesti, Gianluca D’amico, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

716

717

718 Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3677–3685, 2023.

719

720

721

722 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. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=WLPhywf1si>.

723

724

725

726

727 Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31408–31421. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/shi23j.html>.

728

729

730

731

732

733 Liangliang Shi, Jack Fan, and Junchi Yan. Ot-clip: understanding and generalizing clip via optimal transport. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org, 2024.

734

735

736

737 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

738

739

740 Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.

741

742

743 Joshua Smith and Michael Gashler. An investigation of how neural networks learn from the experiences of peers through periodic weight averaging. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 731–736. IEEE, 2017.

744

745

746 Jiachen Sun, Changsheng Wang, Jiongxiao Wang, Yiwei Zhang, and Chaowei Xiao. Safeguarding vision-language models against patched visual prompt injectors. *CoRR*, abs/2405.10529, 2024. doi: 10.48550/ARXIV.2405.10529. URL <https://doi.org/10.48550/arXiv.2405.10529>.

747

748

749

750 Jiachen Sun, Changsheng Wang, Jiongxiao Wang, Yiwei Zhang, and Chaowei Xiao. Safeguarding vision-language models against patched visual prompt injectors. *CoRR*, abs/2405.10529, 2024. doi: 10.48550/ARXIV.2405.10529. URL <https://doi.org/10.48550/arXiv.2405.10529>.

751

752

753

754 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*, pp. 4566–4575, 2015.

755

756 Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

756 Xiaotian Wang, Takehito Utsuro, and Masaaki Nagata. Document alignment based on overlapping  
 757 fixed-length segments. In Xiyan Fu and Eve Fleisig (eds.), *Proceedings of the 62nd Annual Meet-*  
 758 *ing of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp.  
 759 51–61, Bangkok, Thailand, August 2024. Association for Computational Linguistics. ISBN 979-  
 760 8-89176-097-4. doi: 10.18653/v1/2024.acl-srw.10. URL <https://aclanthology.org/2024.acl-srw.10/>.

762 Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H. Nguyen, and Isao Echizen. Closer look  
 763 at the transferability of adversarial examples: How they fool different models differently. In  
 764 *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI,*  
 765 *USA, January 2-7, 2023*, pp. 1360–1368. IEEE, 2023. doi: 10.1109/WACV56688.2023.00141.  
 766 URL <https://doi.org/10.1109/WACV56688.2023.00141>.

767 Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting  
 768 jailbreaking in multimodal large language models. In *Findings of the Association for Compu-*  
 769 *tational Linguistics: EMNLP 2024*, pp. 13715–13726, 2024.

771 Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey  
 772 of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint*  
 773 *arXiv:2502.14881*, 2025.

774 Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on  
 775 multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

777 Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual  
 778 denotations: New similarity metrics for semantic inference over event descriptions. *Transactions*  
 779 *of the association for computational linguistics*, 2:67–78, 2014.

780 Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A  
 781 survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

782 Wanlin Zhang, Weichen Lin, Ruomin Huang, Shihong Song, and Hu Ding. To tackle adversarial  
 783 transferability: A novel ensemble training method with fourier transformation. In *The Thirteenth*  
 784 *International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.  
 785 OpenReview.net, 2025a. URL <https://openreview.net/forum?id=KW8yzAOIZr>.

787 Xuange Zhang, Dengjie Li, Bo Liu, Zenghao Bao, Yao Zhou, Baisong Yang, Zhongying Liu, Yujie  
 788 Zhong, Zheng Zhao, and Tongtong Yuan. Himix: Reducing computational complexity in large  
 789 vision-language models. *CoRR*, abs/2501.10318, 2025b. doi: 10.48550/ARXIV.2501.10318.  
 790 URL <https://doi.org/10.48550/arXiv.2501.10318>.

791 Minghui Zhou and Xiangfeng Wang. Fedskf: Selective knowledge fusion via optimal transport in  
 792 federated class incremental learning. *Electronics*, 13(9):1772, 2024.

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 A THE USE OF LLMs  
811812  
813 A large language model (LLM) was employed for language polishing and grammar correction. All  
814 scientific ideas, experimental design, analysis, and conclusions were generated solely by the authors.  
815 The LLM did not contribute to any research ideation or content creation.816  
817 B RELATED WORKS  
818819 In this section, we review three main lines of research related to our study: large vision-language  
820 models (LVLMs), adversarial robustness of LVLMs, and optimal transport-based fusion methods.  
821822 **Large Vision-Language Models** In recent years, a number of large vision-language models have  
823 been released, demonstrating the rapid progress of multimodal learning. Representative open-source  
824 efforts include Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023), and more recently Qwen-  
825 VL(Bai et al., 2023) and InternVL (Chen et al., 2024), which extend large language models with  
826 visual perception capabilities. Despite architectural differences, most LVLMs adopt CLIP (Radford  
827 et al., 2021) or its variants as the vision-language alignment backbone. While these models provide  
828 transferable multimodal features for diverse downstream applications, they remain vulnerable to  
829 adversarial perturbations, raising concerns for deployment in safety-critical scenarios.  
830831 **Adversarial Robustness of LVLMs** Adversarial robustness in LVLMs has attracted increasing at-  
832 tention, as perturbations applied to visual, textual or jointly across both modalities can severely  
833 disrupt cross-modal representations. In this work, we focus on adversarial perturbations in the vi-  
834 sual modality. Since the high-dimensional and continuous nature of visual data (Ye et al., 2025),  
835 and the strong transferability of visual adversarial examples (Waseda et al., 2023), make defending  
836 against visual attacks particularly challenging. Furthermore, in many LVLMs applications (such as  
837 medical imaging (Javed et al., 2024) and autonomous driving (Rossolini et al., 2023)), visual in-  
838 puts often serve as the primary source of information for model decision-making. Existing defense  
839 strategies against visual modalities attacks can be grouped into two main categories: inference-phase  
840 defenses and training-phase defenses (Ye et al., 2025). Inference-phase defenses mitigate vulnera-  
841 bilities at deployment time, typically by perturbing images before model inference (Xu et al., 2024;  
842 Sun et al., 2024). These approaches are attractive for their plug-and-play nature, but often can-  
843 not fundamentally eliminate the vulnerability of the vision encoder to adversarial perturbations.  
844 In contrast, training-phase defenses aim to improve robustness during model development, most  
845 commonly through adversarial fine-tuning (Schlarmann et al., 2024; Hossain & Imteaj, 2024b;a).  
846 While effective against certain attacks, such methods often struggle to generalize to unseen pertur-  
847 bations due to overfitting to specific attack (Rice et al., 2020). Our work falls within the scope of  
848 training-phase defenses, extending adversarial fine-tuning with an optimal transport-based fusion  
849 mechanism. Our proposed approach belongs to the category of training-phase defenses, extending  
adversarial fine-tuning with an optimal transport-based fusion mechanism.850 **Optimal Transport Fusion** OT Fusion is a layer-wise model fusion technique that utilizes optimal  
851 transport to align neurons across the models before averaging their associated parameters (Singh  
852 & Jaggi, 2020). This technique has recently been extended to a variety of architectures, including  
853 transformers (Imfeld et al., 2024) and graph neural networks (Ormaniec et al., 2025). Beyond archi-  
854 tectural adaptations, FedSKF (Zhou & Wang, 2024) introduces OT Fusion for knowledge integration  
855 in federated class-incremental learning, which aligns feature distributions between client and server  
856 models to mitigate data heterogeneity. Despite its success in model integration, OT Fusion has seen  
857 limited investigation in the context of adversarial robustness, which motivates our study.858 **Optimal Transport (OT)** OT has been widely adopted in various alignment tasks, including doc-  
859 ument alignment Wang et al. (2024) and word alignment Arase et al. (2023). Melnyk et al. (2024)  
860 proposed AOT, an OT-based framework that aligns reward distributions for large language models  
861 by enforcing distributional preference dominance. OT also offers new perspectives on existing tech-  
862 niques. For example, recent studies (Shi et al., 2023; 2024) show that Contrastive Learning (CL)  
863 and the CLIP model can be reformulated as (Inverse) OT problems, where common objectives such  
as InfoNCE loss can be interpreted as instances of (Inverse) OT for aligning sample similarities.

---

864 **Algorithm 1** Hierarchical OT Fusion for CLIP Visual Encoder

865 **Require:** Clean dataset  $\mathcal{D} = \{(x, y)\}$ ; attack set  $\mathcal{A}$ ; template set  $\mathcal{T}$ ; standard template  $t_0$ ; layer

866 indices  $\ell = 1 \dots L$ .

867 **Ensure:** Fused visual encoder  $\bar{f}_{\theta_{\text{img}}}$  (text encoder frozen).

868 1: **for** each  $a \in \mathcal{A}$  **do**

869 2:   **for** each  $t \in \mathcal{T}$  **do**

870 3:     Construct diversified dataset:  $\mathcal{D}_{\text{div}}^{(a,t)} \leftarrow \{(x + \delta_a(x), y^{(t)}) : (x, y) \in \mathcal{D}\}$ .

871 4:     Train submodel  $M^{(a,t)} \leftarrow \text{AdversarialTrain}(\mathcal{D}_{\text{div}}^{(a,t)})$  with frozen text encoder.

872 5:     Store layer weights  $\{W_\ell^{(a,t)}\}_{\ell=1}^L$ .

873 6:   **end for**

874 7: **end for**

875 8: **Level 1: Prompt-level fusion within each attack**

876 9: **for** each  $a \in \mathcal{A}$  **do**

877 10:   Select anchor model  $M^{(a,t_{\text{anchor}})}$ .

878 11:   **for**  $\ell = 1$  to  $L$  **do**

879 12:     **for** each  $t \in \mathcal{T}$  **do**

880 13:       Align columns of weights:  $\tilde{W}_\ell^{(a,t)} \leftarrow W_\ell^{(a,t)} T_{\ell-1}^{(a,t)}$

881 14:       Compute OT plan:  $T_\ell^{(a,t)} \leftarrow \text{ComputeOT}(\tilde{W}_\ell^{(a,t)}, W_\ell^{(a,t_{\text{anchor}})})$ .

882 15:       Align weights:  $W_\ell^{(a,t),\text{aligned}} \leftarrow (T_\ell^{(a,t)})^\top W_\ell^{(a,t)} T_{\ell-1}^{(a,t)}$ .

883 16:     **end for**

884 17:     Average aligned weights:  $W_\ell^{(a)} \leftarrow \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} W_\ell^{(a,t),\text{aligned}}$ .

885 18:   **end for**

886 19:   Assemble fused model  $\bar{M}^{(a)}$  with  $\{W_\ell^{(a)}\}$ .

887 20:   Fine-tune  $\bar{M}^{(a)}$  on  $\mathcal{D}_{\text{div}}^{(a,t_0)}$ .

888 21: **end for**

889 22: **Level 2: Attack-level fusion across attacks**

890 23: Select anchor attack  $a_{\text{anchor}}$ .

891 24: **for**  $\ell = 1$  to  $L$  **do**

892 25:   **for** each  $a \in \mathcal{A}$  **do**

893 26:     Align columns of weights:  $\tilde{W}_\ell^{(a)} \leftarrow W_\ell^{(a)} T_{\ell-1}^{(a)}$

894 27:     Compute OT plan:  $T_\ell^{(a)} \leftarrow \text{ComputeOT}(\tilde{W}_\ell^{(a)}, W_\ell^{(a_{\text{anchor}})})$ .

895 28:      $W_\ell^{(a),\text{aligned}} \leftarrow (T_\ell^{(a)})^\top W_\ell^{(a)} T_{\ell-1}^{(a)}$ .

896 29:   **end for**

897 30:    $W_\ell^{\text{fused}} \leftarrow \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} W_\ell^{(a),\text{aligned}}$ .

898 31: **end for**

899 32: Assemble final fused model  $\bar{f}_{\theta_{\text{img}}}$  with  $\{W_\ell^{\text{fused}}\}_{\ell=1}^L$ .

900 33: **return**  $\bar{f}_{\theta_{\text{img}}}$ .

---

902

903

904

905

## C ALGORITHM

909 In this section, we provide the detailed procedure of our HOT-CLIP in Algorithm 1. Concretely, we

910 first adversarially train a diverse set of submodels, each under a specific attack method and textual

911 template. Within each attack family, the submodels are aligned at the neuron level using optimal

912 transport maps and then averaged to obtain an intra-attack fused model, which is further fine-tuned

913 under the corresponding attack setting. (The procedure relies on computing optimal transport maps

914 through the subroutine  $\text{ComputeOT}(\cdot)$ , whose implementation details are provided in Appendix C.1.)

915 In the second stage, the intra-attack fused models are again aligned and averaged across different

916 attacks, followed by an additional round of fine-tuning to adapt the final visual encoder. This step-

917 by-step process complements the high-level overview in the main text by making the construction,

alignment, and fusion operations explicit.

918 C.1 DETAILS OF OPTIMAL TRANSPORT FUSION  
919920 In this section, we introduce the details of OT Fusion (Singh & Jaggi, 2020) and Transformer-  
921 specific OT Fusion (Imfeld et al., 2024), following the methodologies of Singh & Jaggi (2020)  
922 and Imfeld et al. (2024).923 C.1.1 OPTIMAL TRANSPORT FUSION  
924925 We provide a more formal description of the OT Fusion procedure. Consider two submodels  $A$  and  
926  $B$ , and suppose we are at layer  $\ell$ , with neurons in previous layers already aligned.  
927928 **Step 1: Define probability measures.** We define probability measures over neurons at layer  $\ell$   
929 for the two models as  $\mu^{(\ell)} = (\alpha^{(\ell)}, X^{(\ell)})$ ,  $\nu^{(\ell)} = (\beta^{(\ell)}, Y^{(\ell)})$ , where  $X^{(\ell)} = \{x_1^{(\ell)}, \dots, x_{n^{(\ell)}}^{(\ell)}\}$   
930 and  $Y^{(\ell)} = \{y_1^{(\ell)}, \dots, y_{m^{(\ell)}}^{(\ell)}\}$  are the neuron weight vectors of models  $A$  and  $B$ . We use uniform  
931 histograms as initialization:  $\alpha^{(\ell)} \leftarrow \frac{1}{n^{(\ell)}} \mathbf{1}_{n^{(\ell)}}$ ,  $\beta^{(\ell)} \leftarrow \frac{1}{m^{(\ell)}} \mathbf{1}_{m^{(\ell)}}$ .  
932933 **Step 2: Propagate alignment from previous layer.** To ensure consistency, we first align the  
934 incoming edge weights for layer  $\ell$  using the transport plan from the previous layer,  $T^{(\ell-1)}$ . Formally,  
935

936 
$$\widetilde{W}_A^{(\ell, \ell-1)} \leftarrow W_A^{(\ell, \ell-1)} T^{(\ell-1)} \text{diag}\left(\frac{1}{\beta^{(\ell-1)}}\right), \quad (14)$$

937 where  $W_A^{(\ell, \ell-1)}$  is the weight matrix between layers  $\ell - 1$  and  $\ell$  in model  $A$ . This step aligns the  
938 current layer's weights,  $W_A^{(\ell, \ell-1)}$ , based on the transport map of the preceding layer, so that the  
939 subsequent computation of the optimal transport map for this layer is meaningful.  
940941 **Step 3: Solve optimal transport at current layer.** Given a ground cost  $D_S(\cdot, \cdot)$  (we use Euclidean  
942 distance), the transport plan  $T^{(\ell)}$  is obtained by solving  
943

944 
$$T^{(\ell)} \leftarrow \text{OT}\left(\mu^{(\ell)}, \nu^{(\ell)}; D_S\right), \quad (15)$$

945 We solve the OT problem for the current layer using the Sinkhorn algorithm (Cuturi, 2013)  
946947 **Step 4: Align neurons and fuse.** Using  $T^{(\ell)}$ , we align model  $A$ 's weights with respect to model  
948  $B$ :  
949

950 
$$W_{A, \text{aligned}}^{(\ell, \ell-1)} \leftarrow \text{diag}(1/\beta^{(\ell)}) T^{(\ell)\top} \widetilde{W}_A^{(\ell, \ell-1)}. \quad (16)$$

951 Finally, the fused weights are obtained as  
952

953 
$$W_F^{(\ell, \ell-1)} \leftarrow \frac{1}{2} \left( W_{A, \text{aligned}}^{(\ell, \ell-1)} + W_B^{(\ell, \ell-1)} \right). \quad (17)$$

954 Then, the above procedure is applied sequentially across all layers.  
955956 C.1.2 TRANSFORMER FUSION  
957958 In Transformers, the flow of transportation maps (TMs) becomes more complex due to structure of  
959 Transformers (residual connections, multi-head attention, and normalization layers). In this section,  
960 we introduce the OT Fusion of transformer (Imfeld et al., 2024).  
961962 **Residual Connections.** For a residual block, the outputs of the main branch and the skip branch  
963 are summed. Thus, the outgoing TM depends on both incoming TMs. We use a weighted averaging  
964 strategy:  
965

966 
$$T_{\text{out}}^{(\ell)} = T_{\text{main}}^{(\ell)} \text{diag}(1 - \gamma^{(\ell)}) + T_{\text{res}}^{(\ell)} \text{diag}(\gamma^{(\ell)}), \quad (18)$$

967 where  $\gamma^{(\ell)}$  is a weighting vector that controls the relative contribution of the main branch and the  
968 skip (residual) branch.  
969

972 **Multi-Head Attention.** Multi-head attention introduces additional challenges for OT Fusion, as  
 973 transport maps must be propagated consistently across the query ( $W_Q$ ), key ( $W_K$ ), value ( $W_V$ ), and  
 974 output ( $W_O$ ) projections. Recall that the attention mechanism is defined as  
 975

$$976 \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (19)$$

977 where  $Q = XW_Q$ ,  $K = XW_K$ , and  $V = XW_V$ . We adopt the following alignment strategy: **(i)**  
 978 **Propagation across  $Q$ ,  $K$ , and  $V$ :** The transport maps for  $W_Q$ ,  $W_K$ , and  $W_V$  are inherited directly  
 979 from the previous layer, which applies equally to the multi-head case. **(ii) Handling  $W_O$  under**

980 **hard alignment:** The output projection  $W_O$  depends jointly on the aligned  $Q$ ,  $K$ , and  $V$  branches.  
 981 Under hard alignment, we enforce  $T_Q = T_K = T_{QK}$  so that the permutation cancels inside the  
 982 softmax operation, leaving the attention scores unchanged:  
 983

$$984 \text{softmax}\left(\frac{(QT_Q)(KT_Q)^\top}{\sqrt{d_k}}\right) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right). \quad (20)$$

985 In this case, only the transport map of  $V$  is propagated to  $W_O$ . **(iii) Cross-head alignment:** Because  
 986 there is no guarantee of one-to-one correspondence between heads across models, we adopt a cross-  
 987 head alignment strategy. Specifically, the projection matrices for each head  $\{W_Q^i, W_K^i, W_V^i\}$  are  
 988 concatenated across all heads to form unified matrices  $W_Q, W_K, W_V$ . OT Fusion is then applied to  
 989 these concatenated matrices. Finally, the transport map  $T_V$  is propagated to  $W_O$ .  
 990

991 **Feed-Forward Networks, Layer Normalization, and Embeddings.** Each feed-forward sublayer  
 992 is treated as a standard linear layer. Layer normalization contains only per-dimension affine parame-  
 993 ters  $(\alpha, \beta)$ , which are aligned directly using the incoming transport map. For positional embeddings,  
 994 which are added residually, we apply the same fusion strategies as used for residual connections.  
 995

## 996 C.2 PROOF OF LEMMA 3.1

997 In this section, we provide the proof of lemma 3.1  
 998

1000 *Proof.* To understand why the hierarchical barycenter  $\mu_{\text{hier}}^*$  is close to the global barycenter  $\mu_{\text{global}}^*$ ,  
 1001 let's reason step by step.  
 1002

1003 First, recall that the triangle inequality allows us to break a distance into two parts. Intuitively,  
 1004 the distance between the hierarchical barycenter and the global barycenter can be bounded by the  
 1005 distance from the hierarchical barycenter to each group barycenter, plus the distance from each  
 1006 group barycenter to the global barycenter:  
 1007

$$1008 W_c(\mu_{\text{hier}}^*, \mu_{\text{global}}^*) \leq \frac{1}{|\mathcal{A}|} \sum_a [W_c(\mu_{\text{hier}}^*, \mu_a^*) + W_c(\mu_a^*, \mu_{\text{global}}^*)]. \quad (21)$$

1011 Next, we use the key property of a barycenter: by definition,  $\mu_{\text{hier}}^*$  minimizes the average distance  
 1012 to all group barycenters. This implies that the sum of distances from  $\mu_{\text{hier}}^*$  to each  $\mu_a^*$  is no larger  
 1013 than the sum of distances from the global barycenter to each  $\mu_a^*$ :  
 1014

$$1015 \sum_a W_c(\mu_{\text{hier}}^*, \mu_a^*) \leq \sum_a W_c(\mu_{\text{global}}^*, \mu_a^*). \quad (22)$$

1017 Plugging this into the previous inequality gives a first bound on the hierarchical-global distance:  
 1018

$$1019 W_c(\mu_{\text{hier}}^*, \mu_{\text{global}}^*) \leq \frac{2}{|\mathcal{A}|} \sum_a W_c(\mu_a^*, \mu_{\text{global}}^*). \quad (23)$$

1022 Now we examine each group barycenter  $\mu_a^*$ . Similarly, applying the triangle inequality at this intra-  
 1023 group level gives:  
 1024

$$1025 W_c(\mu_a^*, \mu_{\text{global}}^*) \leq \frac{1}{|\mathcal{T}|} \sum_t [W_c(\mu_a^*, \mu_{a,t}) + W_c(\mu_{a,t}, \mu_{\text{global}}^*)]. \quad (24)$$

1026 Combining (24) with (23), we propagate the bound from the group level to the individual submodels:  
 1027

$$1028 \quad W_c(\mu_{\text{hier}}^*, \mu_{\text{global}}^*) \leq \frac{2}{|\mathcal{AT}|} \sum_a \sum_t \left[ W_c(\mu_a^*, \mu_{a,t}) + W_c(\mu_{a,t}, \mu_{\text{global}}^*) \right]. \quad (25)$$

1031 Finally, by the barycenter property within each group, the distance from the group barycenter to its  
 1032 submodels is at most the distance from the submodels to the global barycenter:  
 1033

$$1034 \quad \sum_t W_c(\mu_a^*, \mu_{a,t}) \leq \sum_t W_c(\mu_{\text{global}}^*, \mu_{a,t}). \quad (26)$$

1036 Substituting this inequality, we obtain a simple bound that directly relates the hierarchical barycenter  
 1037 to all individual submodels:  
 1038

$$1039 \quad W_c(\mu_{\text{hier}}^*, \mu_{\text{global}}^*) \leq \frac{4}{|\mathcal{AT}|} \sum_{a,t} W_c(\mu_{a,t}, \mu_{\text{global}}^*), \quad (27)$$

1041 which intuitively shows that if each submodel is close to the global barycenter, then the hierarchical  
 1042 OT fusion will also remain close.  
 1043

□

### 1047 C.3 ANALYSIS OF MEMORY USAGE

1048 Let  $|\mathcal{A}|$  denote the number of adversarial attack types,  $|\mathcal{T}|$  denote the number of textual prompts  
 1049 and  $U$  represent the memory usage of a single submodel. According to the method of align to an  
 1050 anchor model, In naive global fusion, all  $|\mathcal{A}| \cdot |\mathcal{T}|$  submodels are aligned and stored simultaneously,  
 1051 resulting in a peak memory complexity of  $\mathcal{O}(|\mathcal{A}| |\mathcal{T}| \cdot U)$ .  
 1052

1053 In our two-level hierarchical OT Fusion framework, fusion is performed progressively: **Intra-attack**  
 1054 **fusion (L-level):** For each attack  $a \in \mathcal{A}$ , only the  $|\mathcal{T}|$  submodels corresponding to different prompts  
 1055 are fused at a time. Once fused, the intermediate result replaces the original submodels, so the  
 1056 memory required at any step is proportional to  $\{|\mathcal{T}|\}$ . **Inter-attack fusion (V-level):** The first-  
 1057 level fused models, one per attack type, are further integrated. Again, the number of models stored  
 1058 simultaneously is bounded by  $\{|\mathcal{A}|\}$ .  
 1059

1060 Since the fusion is hierarchical and progressive, the peak memory usage at any step never exceeds  
 $\mathcal{O}(\max\{|\mathcal{A}|, |\mathcal{T}|\} \cdot U)$ .  
 1061

### 1062 D DETAILS OF IMPLEMENTATION

1064 In this section, we detail the implementation and training setup. All models use CLIP (Radford  
 1065 et al., 2021) with a ViT-L/14 visual encoder, initialized from the official OpenAI checkpoint on  
 1066 Hugging Face. Finetuning is performed on 8 NVIDIA RTX 4090 GPUs with PyTorch 2.5 and  
 1067 CUDA 12.4, using AdamW (Loshchilov & Hutter, 2019) with a learning rate of  $1 \times 10^{-5}$ , weight  
 1068 decay 0.01, and batch size 128 (16 per GPU). Each submodel is adversarially fine-tuned for 2 epochs  
 1069 on ImageNet under  $\ell_\infty$  perturbations ( $\epsilon = 4/255$ ). For each attack method (FGSM (Goodfellow  
 1070 et al., 2015), PGD (Madry et al., 2018), and MIM (Dong et al., 2018)), we train three submodels per  
 1071 attack, each using a textual prompt randomly sampled from 80 templates following standard CLIP  
 1072 practice (Radford et al., 2021), to ensure diverse adversarial examples. For OT Fusion, neurons in  
 1073 each linear layer are represented by their incoming weight vectors, which are  $\ell_2$ -normalized prior  
 1074 to computing pairwise distances. The ground cost is defined as the squared Euclidean distance.  
 1075 Transport matrices are computed using the Sinkhorn (Cuturi, 2013) algorithm in the stabilized log-  
 1076 domain, with entropic regularization  $\tau = 8 \times 10^{-2}$ . Each submodel is aligned to an anchor model  
 1077 trained with the standard template (“This is a photo of a ”). During intra-attack fusion, the fused  
 1078 model is fine-tuned for 1 epoch with the corresponding attack and the standard template. During  
 1079 inter-attack fusion, the model undergoes 1 epoch of unsupervised adversarial fine-tuning, while  
 reusing the same optimizer and learning rate schedule. A summary of all settings is provided in  
 Table 5.

1080

1081

Table 5: Hyperparameters used in our experiments.

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

We evaluate our models across three tasks: image classification, image captioning, and visual question answering (VQA), considering both clean performance and adversarial robustness. Across all tasks, we adopt AutoAttack (Croce & Hein, 2020) under the  $\ell_\infty$  norm with perturbation radii  $\epsilon = 2/255$  and  $\epsilon = 4/255$ , each run for 100 iterations. This provides a standardized and reliable evaluation of robustness against adversarial perturbations.

1102

1103

1104

1105

1106

**Image classification.** This task requires assigning a semantic label to each input image. We evaluate clean and robust accuracy on ImageNet and 13 additional zero-shot datasets (details in Section 4.1). For each dataset, class names are combined with a predefined set of prompt templates, and zero-shot classification is performed as described in Section 2. Accuracy is reported as the proportion of correctly classified images.

1107

1108

1109

1110

1111

**Image captioning.** Here, the model generates natural language descriptions conditioned on an image. We evaluate performance on the COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) datasets using the CIDEr score (Vedantam et al., 2015), a consensus-based metric that measures the similarity of generated captions to human-written references. For image captioning, OpenFlamingo-9B (OF) (Alayrac et al., 2022) is evaluated in a zero-shot setting without additional in-context exemplars, whereas LLaVA-1.5-7B (Liu et al., 2023) is evaluated using its default system prompt and captioning prompt. Clean evaluation uses the full validation sets, and adversarial evaluation is conducted on 500 randomly sampled images per dataset, following Schlarmann & Hein (2023).

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

**Visual question answering.** This task requires the model to answer natural language questions based on an image. We evaluate on two widely used benchmarks, VQAv2 (Goyal et al., 2017) and TextVQA (Singh et al., 2019). Performance is measured by VQA accuracy, which computes the proportion of model predictions that match human-annotated answers, thereby reflecting both linguistic and visual reasoning ability. The same LVLMs and evaluation settings as in the captioning task are used.

**Datasets.** (1) For zero-shot image classification, we use ImageNet (Deng et al., 2009), Caltech-101 (Fei-Fei et al., 2007), Cars (Krause et al., 2013), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), FGVC Aircraft (Maji et al., 2013), Flowers (Nilsback & Zisserman, 2008), ImageNet-R (Hendrycks et al., 2021), ImageNet-S (Gao et al., 2023), PCAM (Veeling et al., 2018), Oxford Pets (Parkhi et al., 2012), STL-10 (Coates et al., 2011). (2) For visual question answering (VQA), we use TextVQA (Singh et al., 2019) and VQAv2 (Goyal et al., 2017). (3) For image captioning, we use COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). For classification and VQA, we report the accuracies under clean and adversarial inputs. For image captioning, we use CIDEr (Vedantam et al., 2015) to evaluate the quality of generated captions under attack.

1134

1135 Table 6: Comparison of fusion strategies on zero-shot image classification. Robustness is measured  
1136 using AutoAttack with  $\ell_\infty$  perturbations bounded by  $\epsilon = 2/255$  and  $\epsilon = 4/255$ . Bold numbers  
1137 indicate the best performance in each column.

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

strategy	Zero-shot datasets														
	<i>ImageNet</i>	<i>CalTech</i>	<i>Cars</i>	<i>CIFAR10</i>	<i>CIFAR100</i>	<i>DTD</i>	<i>EuroSAT</i>	<i>FGVC</i>	<i>Flowers</i>	<i>ImageNet-R</i>	<i>ImageNet-S</i>	<i>PCAM</i>	<i>OxfordPets</i>	<i>STL-10</i>	
clean	Direct Average	69.3	<b>85.6</b>	63.9	76.4	58.7	44.5	17.9	<b>23.0</b>	<b>57.0</b>	81.3	<b>60.0</b>	49.5	86.2	96.1
	Direct OTFusion	69.2	85.0	<b>65.0</b>	76.4	58.6	44.7	17.6	22.5	56.8	<b>82.0</b>	59.7	49.5	86.2	96.2
	<b>HOT-CLIP</b>	<b>69.9</b>	85.1	64.0	<b>78.7</b>	<b>60.0</b>	<b>46.2</b>	<b>18.0</b>	21.8	56.9	80.8	59.9	<b>49.6</b>	<b>87.1</b>	<b>96.3</b>
$\epsilon = 2/255$	Direct Average	52.8	78.2	30.0	56.9	36.1	30.6	<b>12.5</b>	8.0	27.8	60.5	43.2	49.5	70.9	90.0
	Direct OTFusion	52.5	78.4	30.2	56.2	35.7	30.7	12.1	8.0	28.1	60.2	43.6	49.5	71.1	90.1
	<b>HOT-CLIP</b>	<b>53.4</b>	<b>79.5</b>	<b>31.8</b>	<b>58.3</b>	<b>37.2</b>	<b>31.8</b>	12.4	<b>8.1</b>	<b>29.7</b>	<b>60.9</b>	<b>44.6</b>	<b>49.6</b>	<b>71.2</b>	<b>91.0</b>
$\epsilon = 4/255$	Direct Average	31.8	63.7	15.1	36.2	19.9	19.2	9.6	2.0	12.2	39.0	31.6	49.5	51.0	74.9
	Direct OTFusion	32.0	64.3	15.1	36.3	20.4	19.0	9.5	2.0	12.0	38.9	30.5	49.5	51.0	75.0
	<b>HOT-CLIP</b>	<b>34.7</b>	<b>66.5</b>	<b>15.8</b>	<b>38.1</b>	<b>20.9</b>	<b>19.6</b>	<b>10.3</b>	<b>2.9</b>	<b>12.5</b>	<b>39.4</b>	<b>32.8</b>	<b>49.6</b>	<b>51.8</b>	<b>77.0</b>

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1154 Table 7: Comparison of fusion strategies on image captioning and VQA. Robustness is measured using  
1155 AutoAttack with  $\ell_\infty$  perturbations bounded by  $\epsilon = 2/255$  and  $\epsilon = 4/255$ . Bold numbers indicate  
1156 the best performance in each column..

VLM strategy	COCO		Flickr30		TextVQA		VQAv2	
	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$
LLaVA	Direct Average	<b>114.7</b>	46.6	30.8	73.9	36.7	26.1	24.7
	Direct OTFusion	113.5	47.7	30.9	73.7	37.2	25.0	25.3
	<b>HOT-CLIP</b>	110.4	<b>56.5</b>	<b>35.5</b>	<b>74.6</b>	<b>43.1</b>	<b>26.5</b>	<b>25.3</b>
							17.7	<b>12.8</b>
								<b>68.8</b>
								43.8
								<b>34.6</b>

## E ADDITIONAL EXPERIMENTS RESULTS

## E.1 ABLATION STUDIES

**Ablation on Fusion Strategies.** To evaluate the effectiveness of our hierarchical OT Fusion design, we compare three different fusion strategies on three vision-language tasks: zero-shot image classification, image captioning, and visual question answering. The first strategy, Direct Average, simply averages the parameters of all submodels without any alignment. The second strategy, Direct OT Fusion, applies optimal transport to align all submodels simultaneously before averaging. The third strategy, Hierarchical OT Fusion (HOT-CLIP), performs two-level fusion: intra-attack alignment followed by inter-attack integration. For each task, we report both clean and robust performance, with robustness evaluated using AutoAttack under the  $\ell_\infty$  norm with  $\epsilon = 2/255$ . Results in Table 6 and Table 7 indicate that Direct Average and Direct OT Fusion often suffers from misalignment across diverse submodels, leading to degraded clean accuracy and limited robustness gains. In contrast, HOT-CLIP consistently improves robustness while maintaining or slightly enhancing clean performance. This highlights the advantage of controlling submodel similarity at each fusion stage, demonstrating that hierarchical fusion is crucial for effectively leveraging diverse adversarially trained submodels.

**Ablation on the Number of Submodels** Following the fusion strategy, we investigate how the size and composition of the submodel pool affect the performance, we vary: the number of families  $K$  (corresponding to distinct attack methods), and the number of submodels per family  $J$  (corresponding to different prompt variants per attack). We examine several configurations:  $3 \times 3$ ,  $2 \times 3$ ,  $1 \times 3$ , and  $3 \times 1$ , where the first number denotes  $K$  and the second  $J$ . All other settings follow Section 5.1 unless specified otherwise. We report CIDEr score and VQA accuracy for image captioning

1188  
 1189 Table 8: Clean and adversarial zero-shot performance on Image Captioning and VQA for different  
 1190 submodel pool configurations.

VLM	Num.	COCO				Flickr30				TextVQA				VQAv2			
		clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	
LLaVA	(1 $\times$ 3)	108.4	49.6	30.1	74.4	40.0	26.6	25.3	17.7	8.8	67.8	<b>44.9</b>	31.3				
	(2 $\times$ 3)	<b>113.6</b>	49.7	30.9	73.7	37.2	25.0	25.0	<b>18.7</b>	8.8	67.9	43.2	31.4				
	(3 $\times$ 3)	110.4	<b>56.5</b>	<b>35.5</b>	<b>74.6</b>	<b>43.1</b>	26.5	25.3	17.7	<b>12.8</b>	<b>68.8</b>	43.8	<b>34.6</b>				
	(3 $\times$ 1)	110.6	52.2	32.3	74.5	39.2	<b>27.3</b>	<b>26.3</b>	17.7	9.2	67.8	44.2	31.9				

1200 Table 9: Comparison of HOT-CLIP with different adversarial training strengths on image captioning  
 1201 and VQA tasks.

VLM	Strength	COCO				Flickr30				TextVQA				VQAv2			
		clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	
LLaVA	2/255	<b>111.9</b>	48.1	33.0	73.7	<b>43.1</b>	24.0	<b>25.3</b>	<b>17.7</b>	9.8	66.8	<b>44.9</b>	32.2				
	4/255	110.4	<b>56.5</b>	<b>35.5</b>	<b>74.6</b>	<b>43.1</b>	<b>26.5</b>	<b>25.3</b>	<b>17.7</b>	<b>12.8</b>	<b>68.8</b>	43.8	<b>34.6</b>				

1208  
 1209 and VQA tasks, respectively. As shown in Table 8, increasing the number of families  $K$  generally  
 1210 improves robustness, indicating the benefit of incorporating diverse attack types. Adding more sub-  
 1211 models per family ( $J$ ) also enhances performance, demonstrating that prompt diversity contributes  
 1212 to more robust feature representations. Notably, the  $3 \times 3$  configuration achieves the highest overall  
 1213 gains across all tasks.

1214  
 1215 **Ablation on Adversarial Training Strength.** We further study the impact of adversarial training  
 1216 strength on the robustness of our approach. Specifically, we train submodels with perturbation radii  
 1217  $\epsilon_{\text{train}} \in \{2/255, 4/255\}$  under the  $\ell_\infty$  norm, keeping all other training settings fixed. Evaluation is  
 1218 conducted on ImageNet using AutoAttack at  $\epsilon_{\text{eval}} = 2/255$  and  $\epsilon_{\text{eval}} = 4/255$ , in addition to clean  
 1219 accuracy. As shown in Table 9, models adversarially trained with  $\epsilon_{\text{train}} = 2/255$  achieve higher  
 1220 clean performance but are less robust under stronger attacks, while those trained with  $\epsilon_{\text{train}} = 4/255$   
 1221 exhibit the opposite trend.

1222  
 1223 **Ablation Studies on Visual Encoder Backbone.** To evaluate the general applicability of our HOT-  
 1224 CLIP framework, we replace the CLIP ViT-L/14 visual encoder with the smaller CLIP ViT-B/32,  
 1225 while keeping all other training and hierarchical OT Fusion settings unchanged. The resulting mod-  
 1226 els are evaluated on zero-shot image classification. As shown in Table 10, models using ViT-B/32  
 1227 achieve lower absolute accuracy than ViT-L/14 due to reduced capacity (e.g., 63.5% vs 71.2% on  
 1228 ImageNet). Nonetheless, our hierarchical OT Fusion consistently improves robustness, increasing  
 1229 accuracy by approximately 4.2% over the ViT-B/32 baseline and 3.8% over the ViT-L/14 baseline  
 1230 under adversarial evaluation. These results indicate that HOT-CLIP effectively generalizes across  
 1231 different visual encoder backbones while providing tangible robustness gains.

1232  
 1233 **Extended Robustness Evaluation Across Models, Attacks, and Norms** To more comprehensively  
 1234 evaluate the generality and stability of our proposed method, we extend our analysis beyond the pri-  
 1235 mary setup in the main paper. Specifically, we conduct three additional sets of experiments to exam-  
 1236 ine whether our approach remains effective across different model architectures, attack algorithms,  
 1237 and perturbation metrics. We first apply our method to the BLIP2 (Li et al., 2023a) architecture by  
 1238 replacing its default vision encoder with our HOT-CLIP. We evaluate performance on COCO and  
 1239 Flickr30k image captioning, as well as TextVQA and VQAv2. As shown in Table 11, HOT-CLIP  
 1240 provides consistent robustness improvements, demonstrating that our framework is compatible with  
 1241 LVLMs beyond CLIP-based architectures. To further assess the robustness of our method under  
 1242 different adversarial attack, we additionally evaluate the models using PGD attacks. The results in  
 1243 Table 12 show that our method consistently improves robustness under both AutoAttack and PGD,  
 1244 confirming the stability of our approach across different attack algorithms. Finally, we evaluate ro-  
 1245 bustness under  $L_2$ -bounded AutoAttacks using two perturbation budgets ( $\epsilon = 0.5$  and  $\epsilon = 1.0$ ). As

1242  
 1243 Table 10: Zero-shot classification performance and adversarial robustness of CLIP models visual  
 1244 encoder ViT-B/32. Robustness is measured using AutoAttack with  $\ell_\infty$  perturbations bounded by  
 1245  $\epsilon = 2/255$ .

Eval.	Vision encoder	Zero-shot datasets												Avg.		
		<i>ImageNet</i>	<i>CalTech</i>	<i>Cars</i>	<i>CIFAR10</i>	<i>CIFAR100</i>	<i>DTD</i>	<i>EuroSAT</i>	<i>FGVC</i>	<i>Flowers</i>	<i>ImageNet-R</i>	<i>ImageNet-S</i>	<i>PCAM</i>	<i>OxfordPets</i>		
clean	CLIP	59.3	82.1	60.8	89.2	59.1	46.3	53.6	20.3	68.5	65.6	41.2	62.8	88.9	97.5	63.9
	TeCoA	55.9	71.4	14.3	70.6	40.2	26.7	20.1	6.0	27.3	48.3	27.5	49.4	72.0	86.9	44.0
	FARE	48.7	80.6	34.2	68.1	46.0	34.7	16.0	11.2	38.0	48.3	34.3	49.4	78.9	89.2	48.4
	HOT-CLIP	50.3	80.8	34.9	70.7	47.2	35.4	16.6	11.7	36.5	50.8	34.3	49.5	79.0	90.3	49.1
$\epsilon = 2/255$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	TeCoA	<b>39.1</b>	64.0	5.0	49.2	25.7	19.5	13.6	2.0	13.0	<b>31.2</b>	18.9	49.1	<b>56.0</b>	74.3	32.9
	FARE	29.6	67.0	<b>12.9</b>	47.2	27.6	22.8	12.9	3.7	14.1	29.2	20.8	49.4	50.0	77.3	33.1
	HOT-CLIP	30.6	<b>67.1</b>	11.5	<b>49.5</b>	<b>29.0</b>	<b>24.3</b>	<b>13.7</b>	<b>3.8</b>	<b>15.2</b>	<b>31.2</b>	<b>22.5</b>	<b>49.6</b>	51.8	<b>77.4</b>	<b>34.0</b>

1260  
 1261 Table 11: **Evaluation of BLIP2 with different encoders under AutoAttack.** Results are reported  
 1262 for image captioning (CIDEr) on COCO and Flickr30k, and VQA accuracy (%) on TextVQA and  
 1263 VQAv2.  $\epsilon$  indicates the  $\ell_\infty$  perturbation bound.

VLM	Vision encoder	COCO			Flickr30			TextVQA			VQAv2		
		clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	clean	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$
BLIP-2	ViT-L/14	<b>127.7</b>	3.9	2.5	<b>83.3</b>	2.1	1.4	<b>32.3</b>	0.0	0.0	<b>48.4</b>	3.2	0.5
	TeCoA	99.7	45.2	23.1	53.6	30.7	19.8	24.5	8.1	5.2	45.8	25.6	20.1
	FARE	109.6	54.1	35.3	68.4	31.5	22.1	25.9	9.2	5.8	46.3	26.4	20.7
	HOT-CLIP	115.8	<b>58.4</b>	<b>38.8</b>	75.3	<b>44.9</b>	<b>27.2</b>	27.4	<b>11.7</b>	<b>8.3</b>	47.1	<b>28.8</b>	<b>23.9</b>

1272  
 1273 shown in Table 13, HOT-CLIP maintains strong robustness under  $L_2$  perturbations, suggesting that  
 1274 the benefits of our approach generalize across multiple threat models.

## 1277 E.2 ADDITIONAL TASK

1278 **Robustness under Stealthy Targeted Attacks.** Stealthy targeted attacks are high-risk adversarial  
 1279 scenarios, where the attacker aims to manipulate the model to produce a specific target output while  
 1280 the perturbation remains imperceptible to the user (Schlarbmann & Hein, 2023). Such attacks pose  
 1281 real-world safety concerns, for example, by guiding users to phishing websites or spreading false  
 1282 information. To evaluate the effectiveness of HOT-CLIP under these conditions, we substitute the  
 1283 CLIP visual encoder in LLaVA-1.5 7B with our robust versions. We perform  $\ell_\infty$  stealthy targeted  
 1284 attacks using APGD (Croce & Hein, 2020) with 500 iterations. Two perturbation radii are consid-  
 1285 ered:  $\epsilon = 2/255$  and  $\epsilon = 4/255$ . The target caption is “You are too stupid. I will not answer  
 1286 anymore” and is tested on 25 sampled images. An attack is deemed successful if the target string is  
 1287 exactly contained in the model’s output. Table 14 reports the attack success rate (ASR) for different  
 1288 encoders. The original CLIP encoder is completely susceptible to attacks at  $\epsilon = 2/255$ , whereas  
 1289 HOT-CLIP achieves 0% ASR, demonstrating full robustness. Even at  $\epsilon = 4/255$ , HOT-CLIP sub-  
 1290 stantially mitigates attacks compared to CLIP. Qualitative examples are shown in Fig. 4, illustrating  
 1291 that HOT-CLIP maintains high-quality captions while resisting the targeted manipulations. These  
 1292 findings extend our main results on general adversarial robustness, confirming that HOT-CLIP not  
 1293 only improves zero-shot classification, VQA, and image captioning robustness, but also protects  
 1294 LVLMs against realistic high-risk targeted attacks.

1295 **Hallucination Experiments** Large vision-language models (LVLMs) are prone to object hallucin-  
 1296 ation, where the model predicts the presence of objects that do not actually appear in the image.

1296  
1297 **Table 12: Evaluation of LLaVA with different encoders under PGD attack.** Results are reported  
1298 for image captioning (CIDEr) on COCO and Flickr30k, and VQA accuracy (%) on TextVQA and  
1299 VQAv2.  $\epsilon$  indicates the  $l_\infty$  perturbation bound. Bold indicates the best performance in each column.

VLM	Vision encoder	COCO		Flickr30			TextVQA			VQAv2			
		clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$	clean	$\epsilon = \frac{2}{255}$
LLaVA	VIT-L	<b>122.2</b>	4.6	3.1	<b>79.1</b>	2.2	1.5	<b>37.8</b>	1.2	0.0	<b>72.4</b>	3.7	1.1
	TeCoA	93.9	43.2	18.4	50.9	27.9	17.0	19.4	13.9	9.7	63.4	42.2	31.3
	FARE	105.8	52.4	34.7	64.7	30.6	22.3	27.5	16.3	9.8	65.6	42.6	31.7
	HOT-CLIP	110.4	<b>57.1</b>	<b>39.3</b>	74.6	<b>46.2</b>	<b>28.5</b>	25.3	<b>19.2</b>	<b>13.8</b>	68.8	<b>45.1</b>	<b>36.5</b>

1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309 **Table 13: Evaluation of LLaVA with different encoders under AutoAttack.** Results are reported  
1310 for image captioning (CIDEr) on COCO and Flickr30k, and VQA accuracy (%) on TextVQA and  
1311 VQAv2.  $\epsilon$  indicates the  $l_2$  perturbation bound.

VLM	Vision encoder	COCO		Flickr30			TextVQA			VQAv2			
		clean	$\epsilon = 0.5$	clean	$\epsilon = 0.5$	clean	$\epsilon = 0.5$	clean	$\epsilon = 0.5$	clean	$\epsilon = 0.5$	clean	
LLaVA	CLIP	<b>122.2</b>	10.9	3.7	<b>79.1</b>	4.2	2.3	<b>37.8</b>	1.2	0.0	<b>72.4</b>	3.7	1.1
	TeCoA	93.9	56.6	31.7	50.9	32.1	21.7	19.4	14.1	11.5	63.4	45.4	33.2
	FARE	105.8	64.5	42.4	64.7	35.9	23.8	27.5	18.7	12.3	65.6	47.9	34.2
	HOT-CLIP	110.4	<b>70.2</b>	<b>48.1</b>	74.6	<b>52.5</b>	<b>33.4</b>	25.3	<b>21.4</b>	<b>15.6</b>	68.8	<b>48.5</b>	<b>37.3</b>

1312  
1313 To assess this issue, we adopt the POPE benchmark (Li et al., 2023b), which formulates hallucina-  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
20100  
20101  
20102  
20103  
20104  
20105  
20106  
20107  
20108  
20109  
20110  
20111  
20112  
20113  
20114  
20115  
20116  
20117  
20118  
20119  
20120  
20121  
20122  
20123  
20124  
20125  
20126  
20127  
20128  
20129  
20130  
20131  
20132  
20133  
20134  
20135  
20136  
20137  
20138  
20139  
20140  
20141  
20142  
20143  
20144  
20145  
20146  
20147  
20148  
20149  
20150  
20151  
20152  
20153  
20154  
20155  
20156  
20157  
20158  
20159  
20160  
20161  
20162  
20163  
20164  
20165  
20166  
20167  
20168  
20169  
20170  
20171  
20172  
20173  
20174  
20175  
20176  
20177  
20178  
20179  
20180  
20181  
20182  
20183  
20184  
20185  
20186  
20187  
20188  
20189  
20190  
20191  
20192  
20193  
20194  
20195  
20196  
20197  
20198  
20199  
20200  
20201  
20202  
20203  
20204  
20205  
20206  
20207  
20208  
20209  
20210  
20211  
20212  
20213  
20214  
20215  
20216  
20217  
20218  
20219  
20220  
20221  
20222  
20223  
20224  
20225  
20226  
20227  
20228  
20229  
20230  
20231  
20232  
20233  
20234  
20235  
20236  
20237  
20238  
20239  
20240  
20241  
20242  
20243  
20244  
20245  
20246  
20247  
20248  
20249  
20250  
20251  
20252  
20253  
20254  
20255  
20256  
20257  
20258  
20259  
20260  
20261  
20262  
20263  
20264  
20265  
20266  
20267  
20268  
20269  
20270  
20271  
20272  
20273  
20274  
20275  
20276  
20277  
20278  
20279  
20280  
20281  
20282  
20283  
20284  
20285  
20286  
20287  
20288  
20289  
20290  
20291  
20292  
20293  
20294  
20295  
20296  
20297  
20298  
20299  
20300  
20301  
20302  
20303  
20304  
20305  
20306  
20307  
20308  
20309  
20310  
20311  
20312  
20313  
20314  
20315  
20316  
20317  
20318  
20319  
20320  
20321  
20322  
20323  
20324  
20325  
20326  
20327  
20328  
20329  
20330  
20331  
20332  
20333  
20334  
20335  
20336  
20337  
20338  
20339  
20340  
20341  
20342  
20343  
20344  
20345  
20346  
20347  
20348  
20349  
20350  
20351  
20352  
20353  
20354  
20355  
20356  
20357  
20358  
20359  
20360  
20361  
20362  
20363  
20364  
20365  
20366  
20367  
20368  
20369  
20370  
20371  
20372  
20373  
20374  
20375  
20376  
20377  
20378  
20379  
20380  
20381  
20382  
20383  
20384  
20385  
20386  
20387  
20388  
20389  
20390  
20391  
20392  
20393  
20394  
20395  
20396  
20397  
20398  
20399  
20400  
20401  
20402  
20403  
20404  
20405  
20406  
20407  
20408  
20409  
20410  
20411  
20412  
20413  
20414  
20415  
20416  
20417  
20418  
20419  
20420  
20421  
20422  
20423  
20424  
20425  
20426  
20427  
20428  
20429  
20430  
20431  
20432  
20433  
20434  
20435  
20436  
20437  
20438  
20439  
20440  
20441  
20442  
20443  
20444  
20445  
20446  
20447  
20448  
20449  
20450  
20451  
20452  
20453  
20454  
20455  
20456  
20457  
20458  
20459  
20460  
20461  
20462  
20463  
20464  
20465  
20466  
20467  
20468  
20469  
20470  
20471  
20472  
20473  
20474  
20475  
20476  
20477  
20478  
20479  
20480  
20481  
20482  
20483  
20484  
20485  
20486  
20487  
20488  
20489  
20490  
20491  
20492  
20493  
20494  
20495  
20496  
20497  
20498  
20499  
20500  
20501  
20502  
20503  
20504  
20505  
20506  
20507  
20508  
20509  
20510  
20511  
20512  
20513  
20514  
20515  
20516  
20517  
20518  
20519  
20520  
20521  
20522  
20523  
20524  
20525  
20526  
20527  
20528  
20529  
20530  
20531  
20532  
20533  
20534  
20535  
20536  
20537  
20538  
20539  
20540  
20541  
20542  
20543  
20544  
20545  
20546  
20547  
20548  
20549  
20550  
20551  
20552  
20553  
20554  
20555  
20556  
20557  
20558  
20559  
20560  
20561  
20562  
20563  
20564  
20565  
20566  
20567  
20568  
20569  
20570  
20571  
20572  
20573  
20574  
20575  
20576  
20577  
20578  
20579  
20580  
20581  
20582  
20583  
20584  
20585  
20586  
20587  
20588  
20589  
20590  
20591  
20592  
20593  
20594  
20595  
20596  
20597  
20598  
20599  
20600  
20601  
20602  
20603  
20604  
20605  
20606  
20607  
20608  
20609  
20610  
20611  
20612  
20613  
20614  
20615

1350		
1351		
1352		
1353		
1354		
1355		
1356		
1357		
1358	<b>Target:</b> You are too stupid. I will not answer anymore	
1359		
1360		<b>CLIP:</b> You are too stupid. I will not answer anymore
1361		
1362		<b>HOT-CLIP:</b> A skier in a red and white suit is skiing down a snowy hill.
1363		
1364		
1365		<b>CLIP:</b> You are too stupid. I will not answer anymore
1366		
1367		<b>HOT-CLIP:</b> A horse is standing in the snow next to a fence.
1368		
1369		
1370		

Figure 4: Examples of LLaVA captions under stealthy targeted attacks ( $\epsilon = 2/255$ ).

Table 15: Performance of different visual encoders under POPE sampling strategies.

Visual Encoder	POPE sampling			Mean
	Adversarial	Popular	Random	
TeCoA	70.2	73.6	73.0	72.3
FARE	74.0	77.7	76.8	76.2
<b>HOT-CLIP</b>	<b>74.2</b>	<b>77.9</b>	<b>77.1</b>	<b>76.4</b>