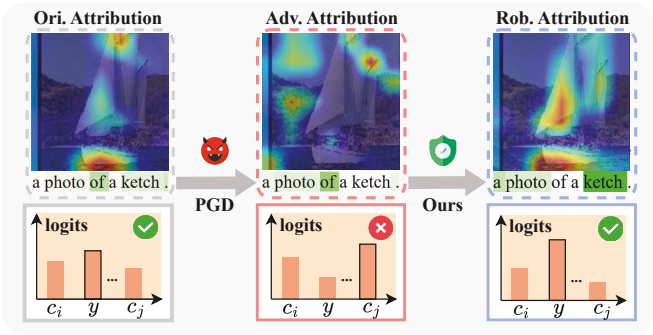
# Stabilizing Cross-Modal Bidirectional Attribution: Few-Shot Adversarial Prompt Tuning for Robust Vision-Language Models

**Jun Feng**[1], **Shuhong Wu**[1], **Hong Sun**[*2], **Pengfei Zhang**[3], **Bocheng Ren**[4], **Shunli Zhang**[5]

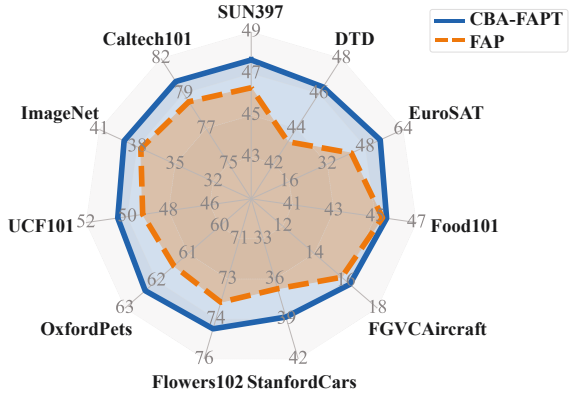
[1]Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]School of Economics, Wuhan Textile University
[3]School of Computer Science and Engineering, Anhui University of Science and Technology, and Key Laboratory of Equipment Data Security and Guarantee Technology, Ministry of Education, Guilin University of Electronic Technology
[4]School of Computer Science and Technology, Hainan University
[5]School of Computer and Information Science, Qinghai Institute of Technology
{junfeng, shuhongwu}@hust.edu.cn, hsun@wtu.edu.cn, zpf.bupt@bupt.cn, bc.revincent@gmail.com, shunlizh@qhit.edu.cn

## Abstract

Large-scale pre-trained Vision-Language Models (VLMs) like CLIP show exceptional performance and zero-shot generalization. However, their reliability may be severely undermined by a critical vulnerability to subtle adversarial perturbations. Our work reveals a critical cross-modal vulnerability: visual-only perturbations induce substantial, synchronous shifts in decision attribution maps across both image and text. This phenomenon signifies a fundamental disruption of the VLM's internal logic, as it alters both the model's perceptual focus and its decision rationale. To counter this vulnerability, we introduce Cross-modal Bidirectional Attribution guided Few-shot Adversarial Prompt Tuning (CBA-FAPT), a novel method that leverages the model's internal decision rationale as a regularizer for robust learning. Our framework's core mechanism is the alignment of a novel bidirectional attribution map. This map is a unique fusion of two components. It combines forward feature attention to capture the model's perceptual focus. It also incorporates backward decision gradients to act as a proxy for the model's decision rationale, quantifying how each feature influences the final outcome. We enforce consistency on this bidirectional map between clean and adversarial examples. This approach corrects the model's internal logic on two fronts and effectively restores its adversarial robustness. Comprehensive experiments on 11 datasets demonstrate that CBA-FAPT outperforms the state-of-the-art, establishing a superior trade-off between robust and natural accuracy.




(a) Cross-Modal Attribution Map under PGD Attack and CBA-FAPT




(b) Average Accuracy of FAP and CBA-FAPT

Figure 1: (a) Cross-modal attribution maps and logits under PGD attack and after stabilization by our CBA-FAPT. (b) CBA-FAPT surpasses FAP in the mean of natural and robust accuracy on 11 datasets.

## Introduction

Large-scale pre-trained Vision-Language Models (VLMs) have emerged as powerful foundational models by learning from vast datasets of image-text pairs. A prime example, CLIP (Radford et al. 2021), demonstrates remarkable zero-shot generalization across diverse downstream tasks, including image recognition (Conde and Turgutlu 2021; Zhou et al. 2022a), visual question answering (Lin and Byrne 2022; Zhou et al. 2020), and text-to-image generation (Zhou et al. 2022b). Despite their impressive capabilities, the reliability of these foundational models is severely undermined by a critical vulnerability to adversarial attacks (Mao et al. 2022; Schlarmann and Hein 2023; Zhao et al. 2023; Devillers et al. 2021; Zhang, Yi, and Sang 2022). Much like traditional deep neural networks (Ren et al. 2025b,a), VLMs are highly sen-

*Corresponding author.

sitive to small, often imperceptible perturbations that can cause them to produce entirely wrong outputs, posing a significant security risk(He et al. 2025; Feng et al. 2020; Zhang et al. 2023).

Adversarial training stands out as the most prominent defense, a technique that enhances robustness by augmenting the training data with adversarial examples (Goodfellow, Shlens, and Szegedy 2014; Levi and Kontorovich 2024; Kuang et al. 2024). However, applying traditional Adversarial training via full fine-tuning (Chen et al. 2020) poses significant challenges for VLMs. As these models scale towards hundreds of billions of parameters, this all-in approach becomes computationally expensive (Madry et al. 2017). More critically, aggressive full fine-tuning risks disrupting the powerful generalized features acquired during pre-training (Kumar et al. 2022). This often forces a difficult trade-off, where any gains in robustness are offset by a decline in accuracy on clean data (Raghunathan et al. 2020).

To address these challenges, prompt tuning has emerged as a compelling, parameter-efficient fine-tuning alternative (Zhou et al. 2022a; Jia et al. 2022; Khattak et al. 2023). Rather than modifying the entire model, this approach freezes the large pre-trained VLM and adapts it to downstream tasks by optimizing only a small set of learnable prompt vectors. This alleviates the cost problems arising from the large scale of fine-tuning parameters and the manual construction of prompts. In addition, the reduction in fine-tuning parameters mitigates the potential damage to the model's generalization capability and demonstrates robust performance in few-shot scenarios. Consequently, combining prompt tuning with adversarial training has emerged as an attractive strategy to enhance VLM robustness without the expensive costs of full fine-tuning (Li et al. 2024; Huang et al. 2023). Yang et al. (Yang et al. 2024) demonstrate that adversarial prompt tuning can lead to overfitting, which causes the model's clean accuracy to drop. To address this issue, they introduced the frozen visual feature constraint to guide loss optimization, thereby enhancing feature alignment and alleviating the overfitting problem.

However, most existing adversarial prompt tuning methods primarily focus on correcting the final output, while overlooking the deeper impact of adversarial perturbations on the model's internal decision-making process. The emergence of powerful interpretability methods (Yu, Zhang, and Xu 2024; Selvaraju et al. 2017; Zhao et al. 2024) offers a window into the model's reasoning. As shown in Figure 1 (a), a key finding of this work is that adversarial attacks simultaneously distort the model's cross-modal decision attribution maps. This cross-modal disruption reveals that adversarial examples are able to deceive the model precisely because they fundamentally alter the basis of its decisions.

Based on this insight, we argue that a more principled defense must go beyond correcting final predictions and instead focus on stabilizing the model's internal decision-making process. To this end, we introduce Cross-modal Bidirectional Attribution guided Few-shot Adversarial Prompt Tuning (CBA-FAPT), a framework that leverages the model's internal decision attributions as a direct regularizer to enforce stable reasoning. To implement this principle, our framework introduces a novel attribution-based regularization to the adversarial prompt tuning process. The framework's core is a cross-modal bidirectional attribution map, constructed from two critical sources. The first is a feature-driven analysis that uses forward attention to capture the model's full perceptual focus by overcoming the limitations of sparse attention. The second is a decision-guided analysis that employs backward gradients to pinpoint the features truly decisive for the model's final logic. Our approach thus provides a more principled and interpretable defense. By correcting the model's underlying decision rationale rather than just its final outputs, CBA-FAPT achieves a superior trade-off between adversarial robustness and accuracy on clean data.

In summary, our main contributions are:

- We reveal a critical cross-modal vulnerability: visual-only adversarial perturbations induce a synchronous and significant shift in the attribution maps across both the image and text modalities.
- To counter this vulnerability, we introduce a novel defense framework that leverages the model's bidirectional attribution maps as a direct regularizer, enforcing consistency between clean and adversarial examples to correct the disrupted decision rationale.
- Extensive experiments across 11 datasets show our method sets a new state-of-the-art in challenging few-shot, cross-dataset, and base-to-new generalization settings, while achieving a superior robustness-accuracy trade-off.

## Related Work

Foundational VLMs like CLIP (Radford et al. 2021) have demonstrated remarkable zero-shot capabilities by aligning vision and text features from large-scale datasets (Jia et al. 2021). To efficiently adapt these massive models, prompt tuning has emerged as a leading parameter-efficient alternative to full fine-tuning, with methods like CoOp (Zhou et al. 2022a) and MaPLe (Khattak et al. 2023) showing strong performance by optimizing only a small set of learnable tokens.

However, the robustness of these powerful models remains a critical concern. Adversarial training(Vaishnavi, Eykholt, and Rahmati 2022; Zhou et al. 2024b) is the most effective defense, but applying it via full fine-tuning is computationally prohibitive and can harm the models' inherent generalization (Kumar et al. 2022). Consequently, recent efforts have focused on combining adversarial training with prompt tuning (Li et al. 2024; Huang et al. 2023). These approaches aim to enhance robustness cost-effectively but primarily focus on correcting the model's final outputs. For instance, FAP (Zhou et al. 2024a) and another recent work (Yang et al. 2024) introduce feature-level constraints to mitigate overfitting and improve output consistency. In contrast to these methods, our work introduces a new defense paradigm. We argue that true robustness requires stabilizing the model's internal decision-making process itself, rather than just its final predictions. To this end, we leverage a novel bidirectional attribution map as a direct regularizer to enforce stable cross-modal reasoning under attack.

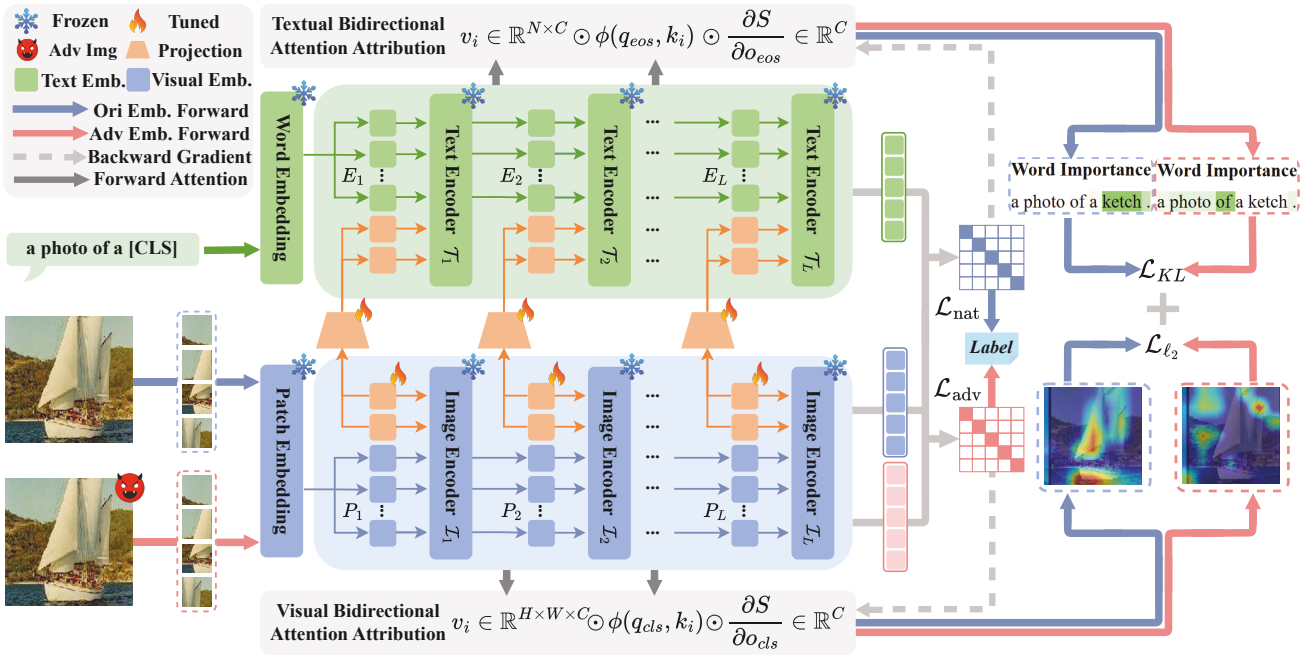



Figure 2: The CBA-FAPT Framework. Learnable prompts are inserted into the model's deep layers. During inference, forward attention attribution and logit-driven backward gradient attribution produce cross-modal bidirectional attention maps. We enforce their alignment to stabilize adversarial predictions.

# Method

## Preliminary of Adversarial Prompt Tuning

The CLIP model is a dual-stream architecture VLM, primarily composed of an image encoder $\mathcal{I}$ and a text encoder $\mathcal{T}$, parameterized as $\boldsymbol{\theta}_{\mathcal{I}}$ and $\boldsymbol{\theta}_{\mathcal{T}}$. In prompt tuning, these encoders are frozen, and task-specific adaptation is achieved by optimizing a small set of learnable prompt tokens $\boldsymbol{P} = \{\boldsymbol{P_v}, \boldsymbol{P_t}\}$. The visual prompts $\boldsymbol{P_v}$ are inserted into the visual embeddings $\boldsymbol{e}(x, \boldsymbol{P_v}) = \{\text{CLS}, e_1(x), e_2(x), \cdots, e_M(x), \boldsymbol{P_v}\}$, while the text prompts $\boldsymbol{P_t}$, derived from $\boldsymbol{P_v}$ via a projection $\boldsymbol{P_t} = h(\boldsymbol{P_v})$, are inserted into the text embeddings $\boldsymbol{w}(t, \boldsymbol{P_t}) = \{\boldsymbol{P_t}, w_1(t), \ldots, w_N(t)\}$.

Given an image-text pair $(x, t)$, CLIP computes features $\mathbf{z}_v^{(x,\boldsymbol{P_v})} = \mathcal{I}(\boldsymbol{e}(x, \boldsymbol{P_v}); \boldsymbol{\theta}_{\mathcal{I}})$ and $\mathbf{z}_t^{(t,\boldsymbol{P_t})} = \mathcal{T}(\boldsymbol{w}(t, \boldsymbol{P_t}); \boldsymbol{\theta}_{\mathcal{T}})$ with prompts. The model's decision is based on the cosine similarity between these features, which we define as the matching score in Eq.1. This score serves to quantify the semantic alignment between the image and text, with the model's final prediction being the class corresponding to the highest score.

$$S^{(\boldsymbol{P_v},\boldsymbol{P_t})}(x, t) = \cos(\mathbf{z}_v^{(x,\boldsymbol{P_v})}, \mathbf{z}_t^{(t,\boldsymbol{P_t})}). \tag{1}$$

However, VLMs are vulnerable to adversarial attacks, where a small perturbation $\delta$ is added to a clean image $x$ to create an adversarial example $\tilde{x} = x + \delta$. The perturbation is crafted to maximize a loss function $\mathcal{J}$ and disrupt the model's prediction, with its magnitude constrained by an $\epsilon$-ball, i.e., $\|\delta\|_p \leq \epsilon$:

$$\arg\max_{\|\delta\|\leq\epsilon} \mathcal{J}\left(x + \delta, t, y; \boldsymbol{\theta}_{\mathcal{I}}; \boldsymbol{\theta}_{\mathcal{T}}\right). \tag{2}$$

Adversarial prompt tuning aims to enhance model robustness by solving the following min-max optimization problem:

$$\arg\min_{\boldsymbol{P}^*} \max_{\|\delta\|\leq\epsilon} \mathcal{J}\left(S^{(\boldsymbol{P_v},\boldsymbol{P_t})}(\tilde{x}, t), y\right), \tag{3}$$

where $y$ is the ground-truth label and $\mathcal{J}$ is the training objective. The inner maximization step generates the most disruptive adversarial examples, while the outer minimization step learns robust prompts $\boldsymbol{P}^*$ that can maintain correct predictions even under such attacks.

## Formulating the Bidirectional Attribution Map

Standard adversarial training builds robustness by aligning outputs at the feature or logit level. However, such coarse-grained alignment schemes merely ensure output consistency, fundamentally neglecting the model's internal decision-making process. We argue that achieving principled robustness requires stabilizing this internal process. To this end, we introduce a bidirectional attribution mechanism that fuses forward attention with backward gradients. This mechanism provides deeper insight into the model's logic, revealing how visual perturbations disrupt cross-modal attributions. Based on this insight, we employs a fine-grained alignment strategy that directly regularizes the model's internal logic, forcing it to maintain a stable cross-modal attribution focus under attack. This principled approach not only corrects the final output but, more importantly, stabilizes the underlying decision-making process.

**Forward Attention Attribution.** Our goal is to construct an attribution map that faithfully reflects how different input

regions contribute to the model's final decision. While the self-attention mechanism seems a natural choice, the standard softmax-based attention maps in CLIP are often too sparse to provide a complete picture, as they tend to concentrate weights on a few image patches. To overcome this limitation, we therefore adopt a decompositional approach to trace how spatial information contributes to the model's final global feature.

Theoretically, this decomposition is grounded in the core architecture of the Vision Transformer. Our strategy is to decompose the final $[CLS]$ feature to precisely isolate the contribution of spatial image patches from the final aggregation layer, which captures the most abstract semantic information. A key component of each Transformer layer is the residual connection. For the final layer $L$, the output $[CLS]$ token feature $x_{cls}^{(L)}$ is precisely formed by the sum:

$$x_{cls}^{(L)} = o_{cls}^{(L)} + x_{cls}^{(L-1)}. \tag{4}$$

Here, $o_{cls}^{(L)}$ represents the attention aggregated output from $L$ layer, while $x_{cls}^{(L-1)}$ is the input feature passed from the previous layer. We focus on the component of $o_{cls}^{(L)}$, which is a weighted sum of the value features $v_i$ from all spatial patches, taking the form $o_{cls}^{(L)} = \sum_i \lambda_i v_i$. In this formulation, the coefficients $\lambda_i$ are the attention weights. These weights are central to our analysis, as they directly quantify the influence of each spatial patch $i$ on the construction of the final, decisive $[CLS]$ representation.

In the standard self-attention mechanism, these weights $\lambda_i$ are derived from the softmax function applied to query-key similarities:

$$\lambda_i^{\text{softmax}} = \text{softmax}\left(\frac{q_{cls}k_i^\top}{\sqrt{C}}\right). \tag{5}$$

However, as prior work (Qiang et al. 2022; Yu and Xiang 2023) has highlighted, this softmax-based formulation often produces extremely sparse maps for models like CLIP, failing to capture a complete picture of semantic relevance. To obtain a more faithful and dense distribution of influence, we therefore bypass this information-compressing step and define our weights directly using non-sparse cosine similarity. This approach preserves the proportional relevance of all spatial patches, resulting in a more complete final map:

$$\lambda_i = \phi(q_{cls}, k_i) = \frac{q_{cls}k_i^\top}{\|q_{cls}\|\|k_i\|}. \tag{6}$$

This equation defines the forward component of our bidirectional attribution map, which successfully captures the model's perceptual focus. However, to form a complete picture, we must also identify which of these attended features are truly decisive for the final prediction. This motivates the introduction of our backward gradient attribution component. As established, forward attention attribution reveals the model's perceptual focus. However, the features that constitute the model's perceptual focus are not necessarily the most critical for its final prediction. To bridge this gap, we introduce a backward, gradient-based attribution mechanism. Starting from the final matching score, this top-down approach identifies which features within the model's perceptual focus were truly decisive for the prediction.

Specifically, this backward analysis targets the final attention output $o_{cls}^{(L)} \in \mathbb{R}^C$. While our forward method analyzed the spatial weights $\lambda_i$ used to construct this feature, our backward method now determines the importance of its channel dimensions for the final score. To this end, we leverage a principle established by gradient-based explainability methods: the partial derivative of a target output with respect to an intermediate feature channel directly quantifies that channel's importance (Ribeiro, Singh, and Guestrin 2016; Selvaraju et al. 2017; Chattopadhay et al. 2018). Following this principle, we define the importance weight $w_c$ for each channel $c$ of the final attention output $o_{cls}^{(L)}$ as the partial derivative of the final matching score $S^{(\boldsymbol{P_v}, \boldsymbol{P_t})}$ with respect to that channel:

$$w_c = \frac{\partial S^{(\boldsymbol{P_v}, \boldsymbol{P_t})}}{\partial o_{cls}^{(L)}[c]} \in \mathbb{R}^C. \tag{7}$$

We formulate the final bidirectional attribution map by synthesizing our two complementary components: forward spatial weights $\lambda_i$ , which identify the model's perceptual focus, and the backward channel weights $w_c$, which assess each feature channel's decisiveness. The resulting visual attribution map for each spatial location $i$ is defined as:

$$A_{\text{v}}(x, S^{(\boldsymbol{P_v}, \boldsymbol{P_t})}) = \sum_c w_c \cdot \lambda_i \cdot v_{ic}. \tag{8}$$

This formulation synthesizes the two components into a comprehensive attribution map. The top-down weights $w_c$ act as a filter, modulating the bottom-up value features $v_{ic}$ to reveal both positive and negative influences from each spatial patch. Capturing this complete decision logic is vital for robust adversarial alignment.

## Cross-modal Attribution Alignment

Our bidirectional attribution map serves as an effective tool for interpreting the internal decision-making logic of VLMs. However, this internal logic is fragile. Prior work has established that adversarial attacks can severely disrupt the model's attention, for instance, by shifting its visual focus from the main subject to irrelevant background details.

Building on this, our work reveals a more profound vulnerability. As illustrated by Figure 1 (a), we find that visual-only adversarial perturbations induce a synchronous shift in attribution across both image and text modalities, fundamentally altering the model's decision logic.

Therefore, a more fundamental defense strategy should not only focus on correcting the final prediction results but also on adjusting the internal decision-making process of the model. Based on this insight, we propose the Cross-modal Attribution Alignment mechanism, which leverages our bidirectional attribution maps as direct supervisory signals to enforce a stable decision-making process under attack.

**Visual Attribution Alignment.** This component aims to enforce robustness within the visual modality by stabilizing the corresponding bidirectional attribution map under attack. Our core strategy is to enforce this stability by minimizing the difference between the attribution maps of clean and adversarial examples. To measure this difference, we employ the $\ell_2$ distance. Given that visual attribution maps are spatially smooth, the $\ell_2$ loss effectively penalizes pixel-level offsets while preserving the overall spatial topological structural consistency between the clean and adversarial maps. The resulting Visual Attribution Alignment loss is therefore defined as:

$$\mathcal{L}_{\text{VAA}} = \frac{1}{N}\sum_{k=1}^{N}\left\| A_{\text{v}}(x_k, S^{(\boldsymbol{P_v},\boldsymbol{P_t})}) - A_{\text{v}}(\tilde{x}_k, \tilde{S}^{(\boldsymbol{P_v},\boldsymbol{P_t})})\right\|_2. \tag{9}$$

Our alignment strategy guides the optimization process towards a more comprehensive form of robustness. This trains the model to be invariant to perturbations that target not only its supporting evidence, which confirms a prediction, but also its suppressive evidence, which contradicts a prediction. By accounting for this full spectrum of logic, the model learns a more stable reasoning process and avoids the critical information loss caused by analyses that discard negative attributions.

**Textual Attribution Alignment.** A key finding of our work is that applying adversarial perturbations solely to the image can also induce a significant shift in the textual attribution map. We hypothesize that this cross-modal disruption is channeled through the backward gradients. The visual perturbation on $\tilde{x}_k$ alters the final matching score, which in turn changes the gradient-based channel weights $w_c$ that determine the importance of textual features. Therefore, to counteract this disruption, our textual attribution alignment strategy directly regularizes the model's decision logic by enforcing consistency between the textual attribution maps of clean and adversarial examples. The alignment of textual maps, however, requires a different metric. While visual attribution maps are typically smooth and distributed, textual maps are inherently sparse and focused, with most of their weight concentrated on a few key tokens. Given these distributional characteristics, we employ Kullback-Leibler (KL) divergence to measure the discrepancy between the textual attribution maps. This metric is particularly well-suited for this task, as it is highly sensitive to shifts in probability mass, effectively capturing deviations in the high-importance tokens. The Textual Attribution Alignment loss is therefore formulated using KL divergence as:

$$\mathcal{L}_{\text{TAA}} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{L}_{KL}\left(A_{\text{t}}(x_k, S^{(\boldsymbol{P_v},\boldsymbol{P_t})})\middle\| A_{\text{t}}(\tilde{x}_k, \tilde{S}^{(\boldsymbol{P_v},\boldsymbol{P_t})})\right), \tag{10}$$

where $A_{\text{t}}$ is bidirectional textual attribution map. Our final training objective is a weighted sum of the adversarial objective and our two alignment regularizers:

$$\mathcal{L} = \mathcal{L}_{\text{nat}} + \alpha\mathcal{L}_{\text{adv}} + \beta\mathcal{L}_{\text{VAA}} + \gamma\mathcal{L}_{\text{TAA}}. \tag{11}$$

Here, $\mathcal{L}_{\text{nat}}$ and $\mathcal{L}_{\text{adv}}$ are the standard cross-entropy losses for the natural and adversarial examples, respectively. The inclusion of the natural loss term $\mathcal{L}_{\text{nat}}$ is crucial for stabilizing optimization and preserving generalization ability, particularly in few-shot scenarios. The coefficients $\alpha$, $\beta$, and $\gamma$ are hyperparameters that balance the contributions of the adversarial loss and our two alignment losses.

# Experiments

## Experimental Setups

**Datasets.** Following previous work, we conducted experiments on 11 image recognition datasets for different tasks to comprehensively and robustly evaluate our proposed method. This benchmark spans a wide spectrum of visual challenges, from general and fine-grained recognition to scene, texture, and action classification, as well as satellite imagery. The datasets include ImageNet (Deng et al. 2009), Caltech101 (Fei-Fei, Fergus, and Perona 2004), FGVCAircraft (Maji et al. 2013), OxfordPet (Parkhi et al. 2012), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014), StanfordCars (Krause et al. 2013), SUN397 (Xiao et al. 2010), DTD(Cimpoi et al. 2014), UCF101 (Soomro, Zamir, and Shah 2012), and EuroSAT (Helber et al. 2019).

**Baseline.** We compare our method against several baselines, which are grouped into the following categories:

- **Zero-shot CLIP**: The original CLIP model without any fine-tuning (Radford et al. 2021).
- **Single-modal Prompt Tuning**: This category includes the text-only method APT (Li et al. 2024) and the visual-only method AdvVP (Mao et al. 2022).
- **Multi-modal Prompt Tuning**: Methods that learn prompts for both modalities, including AdvVLP (Zhou et al. 2024a), AdvMaPLe (Khattak et al. 2023), and FAP (Zhou et al. 2024a).

**Implementation details.** For a fair comparison, we adopt experimental settings consistent with prior work. We use the ViT-B/32 CLIP architecture for all experiments and report all results as an average over three random seeds. We train for 5 epochs for the cross-dataset evaluation and 10 epochs for all other settings. We use an SGD optimizer with a momentum of 0.9, and an initial learning rate of 0.0035 managed by a cosine annealing scheduler with a one-epoch warmup. We use 2 learnable tokens as prompts, which are inserted into the first 9 blocks of both the vision and text encoders. For training, adversarial examples are generated with a 2-step $\ell_\infty$-PGD attack with a attack budget $\epsilon = 1/255$ and a step size $\alpha = 1/255$. While for evaluation, we use a stronger 100-step PGD attack. For our proposed loss terms, we set the hyperparameters $\alpha = 5.0$, $\beta = 0.05$ and $\gamma = 0.03$.

## Adversarial Few-Shot Learning

We assess the model's generalization from scarce data by conducting few-shot fine-tuning with 1, 2, 4, 8, and 16 samples per class. As shown in Figure 3, the prior state-of-the-art method FAP, exhibits a critical weakness in data-scarce scenarios. While generally effective, its performance

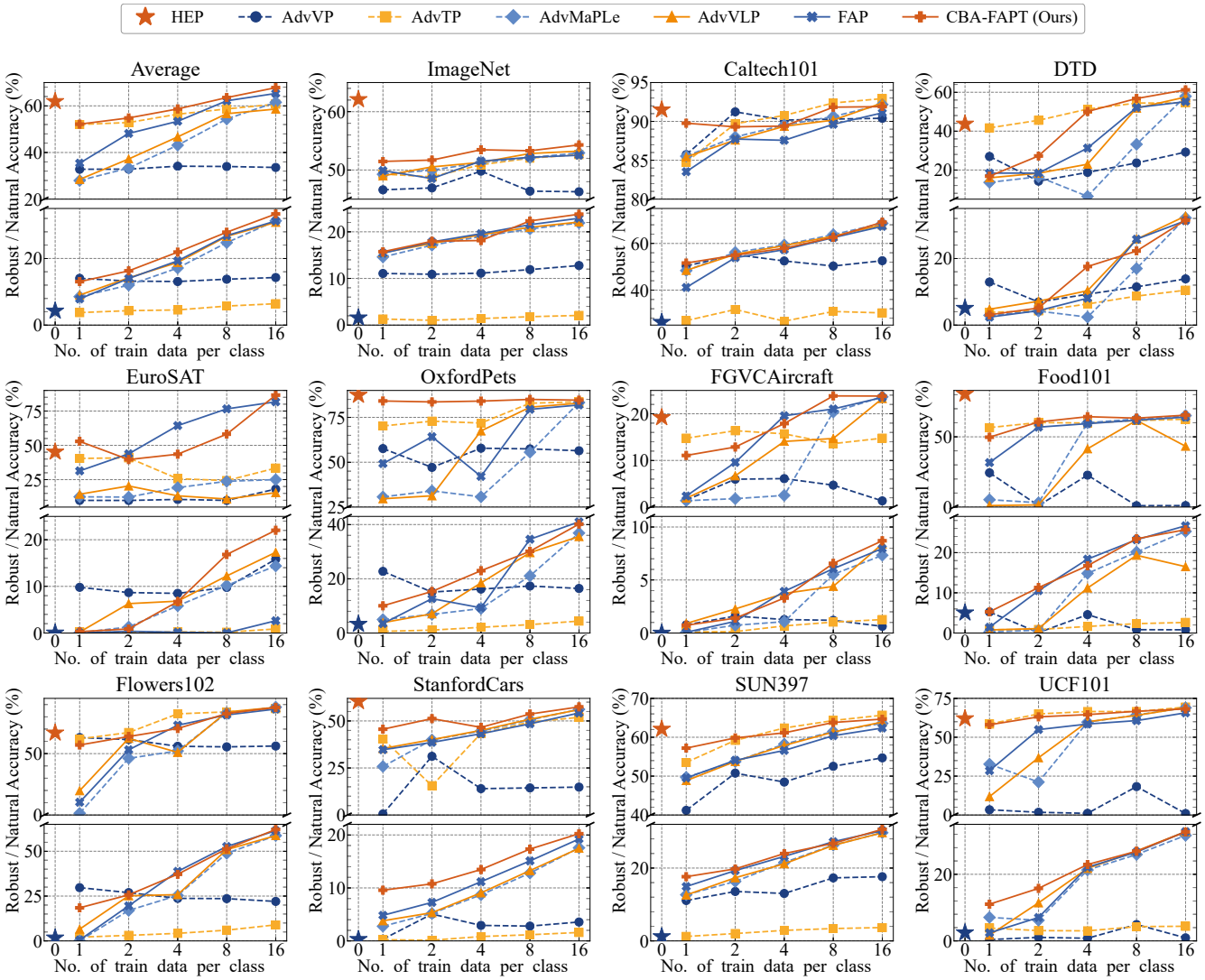



Figure 3: Accuracy (%) of adversarial few-shot learning on 11 datasets. Robust Accuracy is reported beneath each subfigure, Natural Accuracy above.

degrades sharply in extreme few-shot settings (1 and 2 samples), where its robust accuracy on datasets like Flowers102 can drop to nearly zero. In contrast, our method demonstrates superior performance under these challenging conditions. Our method achieves a higher average accuracy on both clean and adversarial data across all settings. Furthermore, our method addresses the common trade-off between robust and natural accuracy. While typical adversarial training degrades performance on clean data, our approach partially recovers this performance, striking a superior balance between the two metrics.

## Adversarial Cross-Dataset Evaluation

To further assess our method's generalization, we evaluate its performance in a cross-dataset setting. Specifically, we fine-tune the model on ImageNet in a 16-shot setting and subsequently evaluate its zero-shot robustness on the 10 remaining datasets. The results in Table 1 demonstrate that our approach achieves the highest robust accuracy on 7 out of the 11 datasets. Compared to the original CLIP model, where adversarial training typically causes a drop in zero-shot natural accuracy, our method can partially recover the model's natural accuracy, achieving a balance between robust and natural accuracy.

## Adversarial Base-to-New Generalization

Following FAP, we adopt the same experimental setup to evaluate adversarial base-to-new generalization. Specifically, we divide the dataset classes into base classes and new classes, train on the base classes using a 16-shot setting, and then evaluate the accuracy on both the base and new classes. This setup poses a more difficult test than standard

| **Nat Acc.** | ImageNet | Caltech 101 | DTD | Euro SAT | Oxford Pets | FGVC Aircraft | Food 101 | Flowers 102 | Stanford Cars | SUN397 | UCF101 | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CLIP** | **62.10** | **91.50** | **43.70** | **45.20** | **87.40** | **19.20** | **80.50** | **66.90** | **60.40** | **62.10** | **62.00** | **61.91** |
| **AdvVP** | 44.87 | 85.47 | 30.23 | 25.17 | 74.20 | 7.13 | 56.53 | 43.17 | 27.27 | 41.97 | 44.60 | 43.69 |
| **AdvVLP** | 53.23 | 87.33 | 33.43 | 18.37 | 78.80 | 10.70 | 55.80 | 49.77 | 38.70 | 52.81 | 51.50 | 48.22 |
| **AdvMaPLe** | 52.93 | 88.23 | 30.87 | 17.60 | 77.87 | 11.10 | 56.67 | 52.90 | 36.70 | 52.53 | 50.97 | 48.03 |
| **FAP** | 52.53 | 87.80 | 30.93 | 15.30 | 78.20 | 10.70 | 55.83 | 51.20 | 38.70 | 52.47 | 51.73 | 47.76 |
| **CBA-FAPT** | 53.87 | 86.77 | 30.33 | 21.53 | 81.20 | 13.43 | 58.07 | 53.87 | 42.63 | 53.40 | 51.23 | 49.67 |

| **Rob Acc.** | ImageNet | Caltech 101 | DTD | Euro SAT | Oxford Pets | FGVC Aircraft | Food 101 | Flowers 102 | Stanford Cars | SUN397 | UCF101 | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CLIP** | 1.57 | 26.23 | 5.07 | 0.03 | 3.27 | 0.00 | 5.03 | 1.73 | 0.30 | 1.20 | 2.47 | 4.26 |
| **AdvVP** | 11.67 | 48.07 | 12.93 | 4.57 | 19.03 | 0.83 | 9.70 | 16.20 | 2.90 | 12.77 | 10.47 | 13.56 |
| **AdvVLP** | 22.10 | 62.97 | **18.60** | 10.67 | 40.83 | 2.73 | 17.83 | 25.23 | 10.97 | 21.67 | 22.10 | 23.25 |
| **AdvMaPLe** | 21.90 | 64.90 | 17.50 | 10.53 | 42.83 | 2.73 | 18.54 | **28.73** | 10.43 | 21.90 | 23.20 | 23.93 |
| **FAP** | 22.90 | **65.43** | 16.93 | 9.97 | 43.77 | 2.77 | 19.60 | 27.23 | 11.80 | 22.40 | **23.77** | 24.23 |
| **CBA-FAPT** | **23.43** | 63.70 | 18.33 | **11.47** | **45.57** | **3.03** | **19.80** | 27.43 | **13.17** | **22.43** | 23.27 | **24.69** |

Table 1: Cross-dataset generalization from ImageNet. We report natural and robust (PGD-100) accuracy. Bold and underline denote the best and second-best results respectively.

| Method | Base Class | | New Class | |
|---|---|---|---|---|
| | Base Nat | Base Adv | New Nat | New Adv |
| **AdvVP** | 31.68 | 14.43 | 30.39 | 13.36 |
| **AdvVLP** | 58.95 | 32.37 | 46.92 | 21.61 |
| **AdvVLP** | 58.95 | 32.37 | 46.92 | 21.61 |
| **AdvMaPLe** | 60.38 | 30.69 | 46.18 | 20.25 |
| **FAP** | 70.52 | 38.05 | 49.58 | 21.86 |
| **CBA-FAPT** | **70.74** | **38.22** | **55.74** | **25.44** |

Table 2: Base-to-new generalization performance, averaged over 11 datasets. We report natural and robust accuracy on both base and new classes.

| Loss | Base Class | | New Class | |
|---|---|---|---|---|
| | Base Nat | Base Adv | New Nat | New Adv |
| $\mathcal{L}_{\text{adv}}$ | 60.38 | 30.69 | 46.18 | 20.25 |
| $+\mathcal{L}_{\text{nat}}$ | 69.11 | 35.72 | 51.74 | 23.38 |
| $+\mathcal{L}_{\text{VAA}}$ | 69.32 | 36.45 | 53.42 | 24.17 |
| **CBA-FAPT** | **70.74** | **38.22** | **55.74** | **25.44** |

Table 3: Ablation study of CBA-FAPT components on base-to-new generalization.

benchmarks, as it challenges the model to generalize across two dimensions: adversarial perturbations and the semantic shift to unseen classes. The results for this challenging setup are reported in Table 2. The results demonstrate a clear advantage for our method, which surpasses the state-of-the-art (SOTA) on both base and new classes. This superiority is especially pronounced in its generalization performance on the unseen new classes. Quantitatively, our method boosts the natural accuracy on new classes by 6.16% and the robust accuracy by 3.58%, demonstrating that CBA-FAPT has superior base-to-new generalization capabilities.

## Ablation Study

We conduct an ablation study on the base-to-new generalization setting to validate the contribution of each component in CBA-FAPT. Starting with a baseline of only the adversarial cross-entropy loss, we incrementally add each of our proposed components to isolate their individual contributions. At each step, we evaluate both natural and robust accuracy on the base and new classes. As shown in Table 3, incorporating the natural loss term substantially boosts natural accuracy on both class sets, while also providing a slight improvement to robust accuracy. Finally, incorporating our cross-modal attribution alignment module further boosts performance across all four metrics, achieving the best overall results. These results confirm the effectiveness of our cross-modal attribution alignment as a regularizer for enhancing generalization.

## Conclusion

We propose CBA-FAPT, a novel framework for robust few-shot prompt tuning. Its core mechanism is a novel bidirectional attribution map, which fuses forward attention with backward gradients to faithfully represent the model's decision logic. By enforcing consistency between the attribution maps of clean and adversarial examples, our framework stabilizes the model's internal logic against attacks, thus enhancing its robustness. Extensive experiments on 11 datasets demonstrate that CBA-FAPT achieves a SOTA robustness-accuracy trade-off, while also setting a new benchmark in challenging few-shot, cross-dataset, and base-to-new settings. Ultimately, by shifting the focus from correcting final outputs to stabilizing the model's internal logic, our work offers a more principled and interpretable path toward robust and reliable VLMs.

## Acknowledgments

This research was supported by the National Key R&D Program of China (No. 2024YFB3108700), the National Natural Science Foundation of China (No. 62372195, and 62462054), CCF-Huawei Populus Grove Fund, the Foundation of Yunnan Key Laboratory of Service Computing (No. YNSC24116), the Key Laboratory of Equipment Data Security and Guarantee Technology, Ministry of Education (No. 2024020300), and the Key Laboratory of Computing Power Network and Information Security, Ministry of Education (No. 2024PY010).

## References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847.

Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; and Wang, Z. 2020. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 699–708.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Conde, M. V.; and Turgutlu, K. 2021. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3956–3960.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devillers, B.; Choksi, B.; Bielawski, R.; and VanRullen, R. 2021. Does language help generalization in vision models? *arXiv preprint arXiv:2104.08313*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 178–178.

Feng, J.; Yang, L. T.; Zhu, Q.; and Choo, K.-K. R. 2020. Privacy-Preserving Tensor Decomposition Over Encrypted Data in a Federated Cloud Environment. *IEEE Transactions on Dependable and Secure Computing*, 17(4): 857–868.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, Y.; Zhang, J.; Yang, P.; Sun, Z.; and Shen, X. 2025. PPBR: Privacy-Preserving and Byzantine-Robust Edge-Assisted Hierarchical Federated Learning in Mobile Networks. *IEEE Transactions on Mobile Computing*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Huang, Q.; Dong, X.; Chen, D.; Chen, Y.; Yuan, L.; Hua, G.; Zhang, W.; and Yu, N. 2023. Improving adversarial robustness of masked autoencoders via test-time frequency-domain prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1600–1610.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

Kuang, H.; Liu, H.; Lin, X.; and Ji, R. 2024. Defense against adversarial attacks using topology aligning adversarial training. *IEEE Transactions on Information Forensics and Security*, 19: 3659–3673.

Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; and Liang, P. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.

Levi, M.; and Kontorovich, A. 2024. Splitting the difference on adversarial training. In *33rd USENIX Security Symposium (USENIX Security 24)*, 3639–3656.

Li, L.; Guan, H.; Qiu, J.; and Spratling, M. 2024. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24408–24419.

Lin, W.; and Byrne, B. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mao, C.; Geng, S.; Yang, J.; Wang, X.; and Vondrick, C. 2022. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505.

Qiang, Y.; Pan, D.; Li, C.; Li, X.; Jang, R.; and Zhu, D. 2022. Attcat: Explaining transformers via attentive class activation tokens. *Advances in Neural Information Processing Systems*, 35: 5052–5064.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Raghunathan, A.; Xie, S. M.; Yang, F.; Duchi, J.; and Liang, P. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*.

Ren, B.; Yang, L. T.; Nie, X.; Feng, J.; Deng, X.; and Zhu, C. 2025a. Zero-Shot Fault Diagnosis for Smart Process Manufacturing via Tensor Prototype Alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 13983–13994.

Ren, B.; Yi, Y.; Zhang, Q.; and Liu, D. 2025b. Zero-Shot Image Recognition via Learning Dual Prototype Accordance Across Meta-Domains. *IEEE Transactions on Image Processing*, 34: 6361–6373.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

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

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Vaishnavi, P.; Eykholt, K.; and Rahmati, A. 2022. Transferring adversarial robustness through robust representation matching. In *31st USENIX Security Symposium (USENIX Security 22)*, 2083–2098.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492.

Yang, F.; Xia, M.; Xia, S.; Ma, C.; and Hui, H. 2024. Revisiting the Robust Generalization of Adversarial Prompt Tuning. *arXiv preprint arXiv:2405.11154*.

Yu, L.; and Xiang, W. 2023. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24355–24363.

Yu, L.; Zhang, H.; and Xu, C. 2024. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37: 96424–96448.

Zhang, H.; Luo, F.; Wu, J.; He, X.; and Li, Y. 2023. LightFR: Lightweight Federated Recommendation with Privacy-preserving Matrix Factorization. *ACM Transactions on Information Systems*, 41(4).

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

Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; and Chan, A. B. 2024. Gradient-based visual explanation for transformer-based clip. In *International Conference on Machine Learning*, 61072–61091. PMLR.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2023. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36: 54111–54138.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13041–13049.

Zhou, Y.; Xia, X.; Lin, Z.; Han, B.; and Liu, T. 2024a. Few-shot adversarial prompt learning on vision-language models. *Advances in Neural Information Processing Systems*, 37: 3122–3156.

Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; and Sun, T. 2022b. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17907–17917.

Zhou, Z.; Li, M.; Liu, W.; Hu, S.; Zhang, Y.; Wan, W.; Xue, L.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024b. Securely fine-tuning pre-trained encoders against adversarial examples. In *2024 IEEE Symposium on Security and Privacy (SP)*, 3015–3033.