Novelty Assessment Report

Paper: AdPO: Enhancing the Adversarial Robustness of Large Vision-Language Models with Preference Optimization
PDF URL: https://openreview.net/pdf?id=FEMv4lHJ2C
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Large Vision-Language Models (LVLMs), such as GPT-4o and LLaVA, have recently witnessed remarkable advancements and are increasingly being deployed in real-world applications. However, inheriting the sensitivity of visual neural networks, LVLMs remain vulnerable to adversarial attacks, which can result in erroneous or malicious outputs. While existing efforts utilize adversarial fine-tuning to enhance robustness, they often suffer from significant performance degradation on clean inputs. In this paper, we propose AdPO, a novel adversarial defense strategy for LVLMs based on preference optimization. For the first time, we reframe adversarial training as a preference optimization problem, aiming to enhance the model's preference for generating normal outputs on clean inputs while rejecting the potential misleading outputs for adversarial examples. Notably, AdPO achieves this by solely modifying the image encoder, e.g., CLIP ViT, resulting in superior clean and adversarial performance in a variety of downstream tasks. Due to the computational cost of training large language models, we show that training on smaller LVLMs and transferring to larger ones achieves state-of-the-art performance with efficiency comparable to previous methods. Our comprehensive experiments confirm the effectiveness of the proposed AdPO which highlights the potential of preference-based learning in adversarially robust multimodal systems.

Core Task Landscape

This paper addresses: **Enhancing adversarial robustness of large vision-language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **31 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Defense Mechanisms and Robustness Enhancement**
- **Attack Strategies and Vulnerability Analysis**
- **Evaluation and Analysis Frameworks**
- **Surveys and Comprehensive Studies**
- **Related Topics and Auxiliary Studies**

Complete Taxonomy Tree

- Enhancing adversarial robustness of large vision-language models Survey Taxonomy
- Defense Mechanisms and Robustness Enhancement
  - Adversarial Training and Fine-Tuning Approaches
  - Vision Encoder Adversarial Fine-Tuning (3 papers)
    - [3] Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models (Christian Schlarmann, 2024) View paper
    - [13] On Enhancing Adversarial Robustness of Large Pre-trained Vision-Language Models (Jie Luo, 2024) View paper
    - [25] Sim-CLIP: Unsupervised Siamese Adversarial Fine-Tuning for Robust and Semantically-Rich Vision-Language Models (Hossain, 2024) View paper
  - Large-Scale Adversarial Pre-Training (1 papers)
    - [20] Double visual defense: Adversarial pre-training and instruction tuning for improving vision-language model robustness (Wang, 2025) View paper
  - Parameter-Efficient Adversarial Adaptation (1 papers)
    - [22] Enhancing Adversarial Robustness of Vision-Language Models through Low-Rank Adaptation (Yuheng Ji, 2025) View paper
  - Federated Adversarial Learning (1 papers)
    - [42] FedAPT: Federated Adversarial Prompt Tuning for Vision-Language Models (Kun Zhai, 2025) View paper
  - Multimodal Adversarial Training (1 papers)
    - [36] Boosting Adversarial Robustness of Vision-Language Pre-training Models against Multimodal Adversarial attacks (Y Wang, 2025) View paper
  - Jailbreak Defense (1 papers)
    - [26] Adversarial training for multimodal large language models against jailbreak attacks (Lu LiMing, 2025) View paper
  - Prompt-Based Defense Strategies
  - Adversarial Prompt Learning (2 papers)
    - [10] Adversarial prompt tuning for vision-language models (Zhang Jia-ming, 2024) View paper
    - [17] One Prompt Word is Enough to Boost Adversarial Robustness for Pre-Trained Vision-Language Models (Lin Li, 2024) View paper
  - Few-Shot Adversarial Prompt Frameworks (1 papers)
    - [5] Few-shot adversarial prompt learning on vision-language models (Zhou Yiwei, 2024) View paper

- Mixture Prompt Tuning (1 papers)
  - [29] Enhancing Adversarial Robustness of Vision Language Models via Adversarial Mixture Prompt Tuning (Zhao ShiJi, 2025) View paper
- Test-Time Defense Methods
- Test-Time Prompt Tuning (2 papers)
  - [2] Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models (Xin Wang, 2025) View paper
  - [7] R-TPT: Improving Adversarial Robustness of Vision-Language Models through Test-Time Prompt Tuning (Li-jun Sheng, 2025) View paper
- Attention-Based Test-Time Defense (1 papers)
  - [16] Text-guided attention is all you need for zero-shot robustness in vision-language models (Yu Lu, 2024) View paper
- Zero-Shot Test-Time Defense (1 papers)
  - [18] On the Zero-shot Adversarial Robustness of Vision-Language Models: A Truly Zero-shot and Training-free Approach (Baoshun Tong, 2025) View paper
- Architectural and Modular Defense Approaches
- Alignment Module Perturbation (1 papers)
  - [11] Enhancing the robustness of vision-language foundation models by alignment perturbation (Cong Zhang, 2025) View paper
- Robust Vision Encoder Integration (1 papers)
  - [28] Robust-llava: On the effectiveness of large-scale robust image encoders for multi-modal large language models (Malik, 2025) View paper
- Hybrid Defense Strategies (1 papers)
  - [24] ArmorCLIP: a hybrid defense strategy for boosting adversarial robustness in vision-language models (Yuhan Liang, 2024) View paper
- Agentic Reasoning Defense (2 papers)
  - [39] ORCA: Agentic Reasoning For Hallucination and Adversarial Robustness in Vision-Language Models (C. Yu, 2025) View paper
  - [45] Hydra: An Agentic Reasoning Approach for Enhancing Adversarial Robustness and Mitigating Hallucinations in Vision-Language Models (Chung-En, 2025) View paper
- Specialized Defense Techniques
- Medical VLM Certified Robustness (1 papers)
  - [8] Promptsmooth: Certifying robustness of medical vision-language models via prompt learning (Shamshad, 2024) View paper
- Adversarial Detection (1 papers)
  - [6] Mirrorcheck: Efficient adversarial defense for vision-language models (Fares Samar, 2024) View paper
- Preference Optimization for Robustness ★ (1 papers)
  - [0] AdPO: Enhancing the Adversarial Robustness of Large Vision-Language Models with Preference Optimization (Anon et al., 2026) View paper
- Self-Calibrated Consistency Defense (1 papers)
  - [49] Self-Calibrated Consistency can Fight Back for Adversarial Robustness in Vision-Language Models (Liu Jiaxiang, 2025) View paper

• Attack Strategies and Vulnerability Analysis
- Multimodal Attack Paradigms (3 papers)
- [1] On the adversarial robustness of multi-modal foundation models (Christian Schlarmann, 2023) View paper
- [4] Revisiting the adversarial robustness of vision language models: a multimodal perspective (Zhou Wanqi, 2024) View paper
- [33] When Alignment Fails: Multimodal Adversarial Attacks on Vision-Language-Action Models (Yuping Yan, 2025) View paper
- Visual Adversarial Attacks (1 papers)
- [15] An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models (Luo, 2024) View paper
- Task-Specific Attack Analysis (1 papers)
- [14] Adversarial robustness for visual grounding of multimodal large language models (Gao, 2024) View paper

• Evaluation and Analysis Frameworks
- Robustness Benchmarking and Evaluation (5 papers)
- [19] On the robustness of medical vision-language models: Are they truly generalizable? (Imam, 2025) View paper
- [21] Evaluating robustness and diversity in visual question answering using multimodal large language models (Xixi Ga, 2024) View paper
- [27] On the robustness of large multimodal models against image adversarial attacks (Cui Xuanming, 2024) View paper
- [38] Evaluating the Adversarial Robustness of Vision-Language Models via Internal Feature Perturbations (Chaohu Liu, 2025) View paper
- [41] Adversarial Robustness of Vision in Open Foundation Models (Jonathon Fox, 2025) View paper
- Efficiency and Context Robustness Analysis (3 papers)
- [9] VLMInferSlow: Evaluating the Efficiency Robustness of Large Vision-Language Models as a Service (Yao Tian-liang, 2025) View paper
- [32] On the robustness of multimodal language model towards distractions (Liu Ming, 2025) View paper
- [50] PCRI: Measuring Context Robustness in Multimodal Models for Enterprise Applications (Agarwal, 2025) View paper
- Modality Robustness and Missing Modality Analysis (3 papers)
- [34] Enhance modality robustness in text-centric multimodal alignment with adversarial prompting (Tsai, 2025) View paper
- [37] Analyzing modality robustness in multimodal sentiment analysis (Hazarika, 2022) View paper
- [43] Are multimodal transformers robust to missing modality? (Mengmeng Ma, 2022) View paper
- Prompt Robustness Analysis (1 papers)
- [12] Towards robust prompts on vision-language models (Gu, 2023) View paper

• Surveys and Comprehensive Studies (5 papers)
- [30] Adversarial attacks of vision tasks in the past 10 years: A survey (Chiyu Zhang, 2025) View paper
- [31] Safety of multimodal large language models on images and texts (Liu Xin, 2024) View paper
- [35] A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks (Tian Pu, 2024) View paper
- [46] A survey of safety on large vision-language models: Attacks, defenses and evaluations (Ye, 2025) View paper

- [47] A Survey of Recent Advances in Adversarial Attack and Defense on Vision-Language Models (MI Hossain, 2025) View paper
- Related Topics and Auxiliary Studies
  - Bias and Fairness in VLMs (1 papers)
  - [23] A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning (Berg, 2022) View paper
  - Domain-Specific VLM Applications (2 papers)
  - [44] Adversarial Robustness Analysis of Vision-Language Models in Medical Image Segmentation (Anjila Budathoki, 2025) View paper
  - [48] Hyperspectral Image Analysis in Single-Modal and Multimodal setting using Deep Learning Techniques (Pande, 2024) View paper
  - Architecture and Design Improvements (1 papers)
  - [40] Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design (Rishika Bhagwatkar, 2024) View paper

Narrative

Core task: Enhancing adversarial robustness of large vision-language models. The field is organized around five main branches that collectively address how to make multimodal systems more resilient to adversarial perturbations. Defense Mechanisms and Robustness Enhancement encompasses a wide range of techniques—from prompt-based methods like PromptSmooth[8] and R-TPT[7] to architectural modifications such as ArmorCLIP[24] and training-time interventions including Robust-LLaVA[28]—all aimed at hardening models against attacks. Attack Strategies and Vulnerability Analysis explores how adversaries can exploit weaknesses in vision-language models, examining both image-level perturbations and text-based jailbreaking approaches. Evaluation and Analysis Frameworks provide systematic ways to measure robustness across diverse settings, while Surveys and Comprehensive Studies offer broad perspectives on safety and adversarial challenges in multimodal systems. Related Topics and Auxiliary Studies connect robustness research to broader concerns such as missing modality handling and domain-specific applications.

Within Defense Mechanisms, a particularly active line of work focuses on specialized techniques that adapt models at inference or training time without full retraining. Preference optimization methods, prompt tuning strategies like Adversarial Prompt Tuning[10] and Few-shot Adversarial Prompt[5], and ensemble-based defenses represent contrasting trade-offs between computational overhead and robustness gains. AdPO[0] sits within this specialized defense cluster, emphasizing preference optimization to align model behavior under adversarial conditions. Compared to prompt-smoothing approaches such as PromptSmooth[8] that aggregate predictions over perturbed prompts, or test-time adaptation methods like Tapt[2] that refine representations dynamically, AdPO[0] leverages preference signals to guide the model toward more robust decision boundaries. This positions it alongside works like Alignment Perturbation[11] that also explore alignment-based defenses, yet AdPO[0] distinctively integrates preference learning into the robustness enhancement pipeline, offering a complementary angle to purely prompt-based or architectural defenses.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on reframing adversarial training through preference optimization to balance robustness and clean performance. Siblings represent alternative defense paradigms: detection-based approaches that identify adversarial samples, certification methods providing formal guarantees, and test-time consistency mechanisms. All address VLM robustness but differ fundamentally in their defense philosophy—optimization vs. detection vs. certification vs. consistency.

Similarities: - All methods aim to enhance adversarial robustness of vision-language models - Each approach seeks to maintain model utility while defending against attacks - All represent alternatives or complements to standard adversarial training

Differences: - Preference Optimization modifies training objectives, while Adversarial Detection operates at inference without model modification - Medical VLM Certified Robustness provides formal guarantees through randomized smoothing, whereas Preference Optimization offers empirical robustness improvements - Self-Calibrated Consistency Defense exploits attack fragility at test-time, while Preference Optimization embeds robustness during training - Preference Optimization explicitly targets the clean-robust performance tradeoff, while siblings focus on detection accuracy, certification radius, or consistency metrics

Suggested Search Directions: - Hybrid approaches combining preference optimization with detection or certification mechanisms - Comparative studies on training-time vs. test-time defense effectiveness for VLMs - Domain-specific preference optimization (e.g., medical VLMs with certified guarantees)

Sibling Subtopics

- **Adversarial Detection** (leaves: 1, papers: 1)
- Scope: Techniques that detect adversarial samples by generating images from captions and comparing embeddings or features.
- Exclude: Excludes methods that modify model behavior rather than detect attacks; see Adversarial Training or Test-Time Defense.
- **Medical VLM Certified Robustness** (leaves: 1, papers: 1)
- Scope: Methods providing certified robustness guarantees for medical vision-language models using randomized smoothing or similar techniques.
- Exclude: Excludes empirical defenses without certification and general VLM methods; see Adversarial Training or Prompt-Based Defense.
- **Self-Calibrated Consistency Defense** (leaves: 1, papers: 1)
- Scope: Approaches exploiting semantic and viewpoint fragility of attacks through self-calibrated consistency mechanisms.
- Exclude: Excludes training-based and prompt-based methods; see Adversarial Training or Test-Time Defense.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: AdPO: Adversarial defense strategy based on preference optimization

Description: The authors propose AdPO, a novel adversarial defense method that reframes adversarial training as a preference optimization problem. This approach enhances LVLMs' preference for generating correct outputs on clean inputs while rejecting misleading outputs on adversarial examples, representing the first application of preference optimization techniques to adversarial training.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Positive enhanced preference alignment for text-to-image models

URL: View paper

Brief Assessment

Positive Enhanced Preference[73] focuses on aligning text-to-image models with human preferences to improve generative quality, not on adversarial defense for vision-language models. The application domains and objectives are fundamentally different.

---

2. Structured preference modeling for reinforcement learning-based fine-tuning of large models

URL: View paper

Brief Assessment

Structured Preference Modeling[75] focuses on general reinforcement learning fine-tuning of large language models using preference modeling for policy optimization, not on adversarial defense for vision-language models. The candidate addresses standard RLHF optimization without any discussion of adversarial training, robustness, or vision-language security.

---

3. Structured preference optimization for vision-language long-horizon task planning

URL: View paper

Brief Assessment

Structured Preference Optimization[77] focuses on vision-language long-horizon task planning in embodied environments, not adversarial defense for vision-language models. The two works address fundamentally different problems despite both using preference optimization techniques.

---

4. Aligning modalities in vision large language models via preference fine-tuning

URL: View paper

Brief Assessment

Aligning Modalities Preference[78] focuses on reducing hallucinations in vision-language models through preference optimization for modality alignment, not adversarial robustness. The candidate addresses a fundamentally different problem (hallucination vs. adversarial attacks) using preference optimization in a different context.

---

5. TCPO: Thought-Centric Preference Optimization for Effective Embodied Decision-making

URL: View paper

Brief Assessment

TCPO[80] focuses on preference optimization for embodied decision-making in interactive environments (alfworld, gymcards), not adversarial defense for vision-language models. The technical domains and objectives are fundamentally different.

---

6. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning

URL: View paper

Brief Assessment

Pref-GRPO[76] focuses on text-to-image generation using pairwise preference rewards for stable reinforcement learning, not adversarial defense for vision-language models. The application domains and technical objectives are fundamentally different.

---

7. SAMPO: Visual Preference Optimization for Intent-Aware Segmentation with Vision Foundation Models

URL: View paper

Brief Assessment

SAMPO[79] applies preference optimization to visual segmentation tasks in medical imaging, not to adversarial training for vision-language models. The candidate focuses on intent-aware segmentation using sparse visual prompts, while the original contribution addresses adversarial robustness in LVLMs through preference-based learning.

---

8. Calibrated self-rewarding vision language models

URL: View paper

Brief Assessment

Calibrated Self-rewarding[72] focuses on addressing hallucination and modality misalignment in vision-language models through self-rewarding with visual constraints, not adversarial robustness. The candidate applies preference optimization to improve alignment between image and text modalities, while the original applies it to adversarial training for robustness against adversarial attacks.

---

9. A preference-driven paradigm for enhanced translation with large language models

URL: View paper

Brief Assessment

Preference-driven Translation[71] applies preference optimization to machine translation quality improvement, not to adversarial defense in vision-language models. The domains and objectives are fundamentally different.

---

10. Modality-balancing preference optimization of large multimodal models by adversarial negative mining

URL: View paper

Brief Assessment

Modality-balancing Preference[74] focuses on addressing modality imbalance in multimodal models through preference optimization with adversarial negative mining, not on adversarial defense or robustness training for vision-language models against adversarial attacks.

Contribution 2: Dual optimization strategy combining PIO and AIO

Description: The authors introduce two complementary optimization components: Preferred Image Optimization increases probability of correct outputs under clean inputs while decreasing erroneous outputs under adversarial images, and Adversarial Image Optimization explicitly optimizes for correct responses under adversarial inputs. This dual approach serves as a general adversarial training framework applicable beyond specific algorithms or models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Enhancing infrared small target detection robustness with bi-level adversarial framework

URL: View paper

Brief Assessment

Bi-level Adversarial Framework[70] addresses infrared small target detection robustness through adversarial corruption generation and detector optimization, not preference-based learning for vision-language models. The technical domains and optimization objectives differ fundamentally.

---

2. Optimizing Robustness and Accuracy in Mixture of Experts: A Dual-Model Approach

URL: View paper

Brief Assessment

Dual-Model Mixture[68] addresses robustness-accuracy trade-offs in Mixture of Experts architectures through dual-model strategies, not vision-language models. The optimization components target different technical domains and model structures.

---

3. Balancing generalization and robustness in adversarial training via steering through clean and adversarial gradient directions

URL: View paper

Brief Assessment

Steering Gradient Directions[63] focuses on balancing clean and adversarial gradient directions through orthogonal projection and interpolation in adversarial training for image classification, not on preference optimization for vision-language models as in the original paper.

---

4. Improving the accuracy-robustness trade-off of classifiers via adaptive smoothing

URL: View paper

Brief Assessment

Adaptive Smoothing Tradeoff[67] focuses on mixing output probabilities of standard and robust classifiers at inference time for vision tasks, not on dual optimization during adversarial training for vision-language models as in the original paper.

---

5. Robustness and accuracy could be reconcilable by (proper) definition

URL: View paper

Brief Assessment

Reconcilable by Definition[66] focuses on redefining robust error through local equivariance rather than dual optimization components. The paper does not propose complementary optimization strategies like PIO and AIO for adversarial training.

---

6. R&D-Agent-Quant: A Multi-Agent Framework for Data-Centric Factors and Model Joint Optimization

URL: View paper

Brief Assessment

R&D-Agent-Quant[69] focuses on quantitative finance factor-model optimization, not adversarial robustness in vision-language models. The dual optimization in the candidate refers to factor and model development in financial markets, not clean/adversarial image training.

---

7. Fortify the guardian, not the treasure: Resilient adversarial detectors

URL: View paper

Brief Assessment

Fortify Guardian[64] focuses on adversarial detector robustness through dual optimization of detector and classifier against adaptive attacks, while the original paper addresses vision-language model robustness through preference optimization combining clean and adversarial image training. These are fundamentally different problem domains and optimization approaches.

---

8. Trade-off between robustness and accuracy of vision transformers

URL: View paper

Brief Assessment

Vision Transformers Tradeoff[65] addresses robustness-accuracy trade-offs in vision transformers through adapter-based architecture modifications, not through dual optimization strategies for adversarial training in vision-language models. The candidate focuses on architectural solutions (accuracy and robustness adapters with gated fusion) for ViTs, while the original proposes preference optimization methods (PIO and AIO) for LVLMs.

---

9. Solving the robustness puzzle: The joint impact of optimization approach, robustness metrics, and scenarios on water resources management under deep â□¦

URL: View paper

Brief Assessment

Water Resources Robustness[62] focuses on water resources management under deep uncertainty, examining robustness metrics and optimization approaches in that domain. This is fundamentally different from the original paper's dual optimization strategy (PIO and AIO) for adversarial training in vision-language models.

---

10. On the duality between sharpness-aware minimization and adversarial training

URL: View paper

Brief Assessment

Sharpness Adversarial Duality[61] focuses on the relationship between sharpness-aware minimization (SAM) and adversarial training (AT) in terms of weight vs. input perturbation, not on dual optimization components for vision-language models combining preferred and adversarial image optimization.

Contribution 3: Transfer learning approach from smaller to larger LVLMs

Description: The authors demonstrate that adversarial training can be performed on smaller LVLM models (e.g., TinyLLaVA) and the resulting robust image encoder can be transferred to larger models. This strategy achieves computational efficiency comparable to previous methods while reducing overfitting risks and enabling fair comparison with prior CLIP-based approaches.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Improving adversarial robustness using knowledge distillation guided by attention information bottleneck

URL: View paper

Brief Assessment

Attention Information Bottleneck[57] focuses on knowledge distillation for adversarial robustness in general neural networks, not on transfer learning from smaller to larger LVLMs specifically. The candidate's approach uses knowledge distillation as a defense mechanism, while the original paper's contribution is about computational efficiency through model size transfer in the LVLM domain.

---

2. KDRSFL: A knowledge distillation resistance transfer framework for defending model inversion attacks in split federated learning

URL: View paper

Brief Assessment

KDRSFL[52] focuses on knowledge distillation in split federated learning for defending model inversion attacks, not on transfer learning from smaller to larger vision-language models for adversarial robustness.

---

3. Improving Adversarial Robustness Through Adaptive Learning-Driven Multi-Teacher Knowledge Distillation

URL: View paper

Brief Assessment

Multi-Teacher Distillation[58] focuses on transferring adversarial robustness from multiple teacher CNNs to a student CNN using knowledge distillation on clean data for image classification tasks, not on transferring robust image encoders from smaller to larger vision-language models.

---

4. Common knowledge learning for generating transferable adversarial examples

URL: View paper

Brief Assessment

Common Knowledge Learning[51] focuses on transfer-based adversarial attacks across different DNN architectures (e.g., CNNs to Transformers) by distilling knowledge from multiple teacher models into a student model for generating transferable adversarial examples. This is fundamentally different from the original paper's contribution of transferring adversarially-trained image encoders from smaller to larger LVLMs for computational efficiency in adversarial defense.

---

5. Reinforced compressive neural architecture search for versatile adversarial robustness

URL: View paper

Brief Assessment

Reinforced Compressive Search[56] focuses on neural architecture search for finding robust sub-networks within teacher networks, not on transfer learning from smaller to larger vision-language models for adversarial robustness.

---

6. On the benefits of knowledge distillation for adversarial robustness

URL: View paper

Brief Assessment

Knowledge Distillation Benefits[53] focuses on distilling adversarial robustness from larger teacher models to smaller student models for compression purposes, not on training smaller models first and then transferring to larger ones as described in the original contribution.

---

7. On transfer of adversarial robustness from pretraining to downstream tasks

URL: View paper

Brief Assessment

Transfer Pretraining Robustness[54] focuses on transferring adversarial robustness from pretrained representations to downstream tasks via linear probing, not on training smaller LVLMs and transferring to larger ones. The candidate examines how robust representations affect downstream task robustness, while the original contribution specifically addresses computational efficiency by training on smaller LVLM models (e.g., TinyLLaVA) and transferring the robust image encoder to larger models.

---

8. Adversarially robust transfer learning

URL: View paper

Brief Assessment

Adversarially Robust Transfer[60] focuses on transferring adversarial robustness from source to target domains in image classification tasks, not on transferring from smaller to larger vision-language models. The candidate addresses robust feature extractors in CNNs, while the original contribution concerns computational efficiency in training LVLMs of different sizes.

---

9. Initialization matters for adversarial transfer learning

URL: View paper

Brief Assessment

Initialization Transfer Learning[55] focuses on transferring adversarially robust image encoders from smaller to larger vision models in computer vision tasks, not on large vision-language models (LVLMs) which combine visual and language components for multimodal understanding.

---

10. Adversarially robust distillation

URL: View paper

Brief Assessment

Adversarially Robust Distillation[59] focuses on knowledge distillation from larger teacher networks to smaller student networks for adversarial robustness in image classification, not on transfer learning from smaller to larger models in LVLMs.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] AdPO: Enhancing the Adversarial Robustness of Large Vision-Language Models with Preference Optimization View paper
- [1] On the adversarial robustness of multi-modal foundation models View paper
- [2] Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models View paper
- [3] Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models View paper
- [4] Revisiting the adversarial robustness of vision language models: a multimodal perspective View paper
- [5] Few-shot adversarial prompt learning on vision-language models View paper
- [6] Mirrorcheck: Efficient adversarial defense for vision-language models View paper
- [7] R-TPT: Improving Adversarial Robustness of Vision-Language Models through Test-Time Prompt Tuning View paper
- [8] Promptsmooth: Certifying robustness of medical vision-language models via prompt learning View paper
- [9] VLMInferSlow: Evaluating the Efficiency Robustness of Large Vision-Language Models as a Service View paper

- [10] Adversarial prompt tuning for vision-language models View paper
- [11] Enhancing the robustness of vision-language foundation models by alignment perturbation View paper
- [12] Towards robust prompts on vision-language models View paper
- [13] On Enhancing Adversarial Robustness of Large Pre-trained Vision-Language Models View paper
- [14] Adversarial robustness for visual grounding of multimodal large language models View paper
- [15] An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models View paper
- [16] Text-guided attention is all you need for zero-shot robustness in vision-language models View paper
- [17] One Prompt Word is Enough to Boost Adversarial Robustness for Pre-Trained Vision-Language Models View paper
- [18] On the Zero-shot Adversarial Robustness of Vision-Language Models: A Truly Zero-shot and Training-free Approach View paper
- [19] On the robustness of medical vision-language models: Are they truly generalizable? View paper
- [20] Double visual defense: Adversarial pre-training and instruction tuning for improving vision-language model robustness View paper
- [21] Evaluating robustness and diversity in visual question answering using multimodal large language models View paper
- [22] Enhancing Adversarial Robustness of Vision-Language Models through Low-Rank Adaptation View paper
- [23] A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning View paper
- [24] ArmorCLIP: a hybrid defense strategy for boosting adversarial robustness in vision-language models View paper
- [25] Sim-CLIP: Unsupervised Siamese Adversarial Fine-Tuning for Robust and Semantically-Rich Vision-Language Models View paper
- [26] Adversarial training for multimodal large language models against jailbreak attacks View paper
- [27] On the robustness of large multimodal models against image adversarial attacks View paper
- [28] Robust-llava: On the effectiveness of large-scale robust image encoders for multi-modal large language models View paper
- [29] Enhancing Adversarial Robustness of Vision Language Models via Adversarial Mixture Prompt Tuning View paper
- [30] Adversarial attacks of vision tasks in the past 10 years: A survey View paper
- [31] Safety of multimodal large language models on images and texts View paper
- [32] On the robustness of multimodal language model towards distractions View paper
- [33] When Alignment Fails: Multimodal Adversarial Attacks on Vision-Language-Action Models View paper
- [34] Enhance modality robustness in text-centric multimodal alignment with adversarial prompting View paper
- [35] A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks View paper
- [36] Boosting Adversarial Robustness of Vision-Language Pre-training Models against Multimodal Adversarial attacks View paper
- [37] Analyzing modality robustness in multimodal sentiment analysis View paper
- [38] Evaluating the Adversarial Robustness of Vision-Language Models via Internal Feature Perturbations View paper
- [39] ORCA: Agentic Reasoning For Hallucination and Adversarial Robustness in Vision-Language Models View paper
- [40] Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design View paper
- [41] Adversarial Robustness of Vision in Open Foundation Models View paper
- [42] FedAPT: Federated Adversarial Prompt Tuning for Vision-Language Models View paper
- [43] Are multimodal transformers robust to missing modality? View paper
- [44] Adversarial Robustness Analysis of Vision-Language Models in Medical Image Segmentation View paper
- [45] Hydra: An Agentic Reasoning Approach for Enhancing Adversarial Robustness and Mitigating Hallucinations in Vision-Language Models View paper
- [46] A survey of safety on large vision-language models: Attacks, defenses and evaluations View paper
- [47] A Survey of Recent Advances in Adversarial Attack and Defense on Vision-Language Models View paper
- [48] Hyperspectral Image Analysis in Single-Modal and Multimodal setting using Deep Learning Techniques View paper
- [49] Self-Calibrated Consistency can Fight Back for Adversarial Robustness in Vision-Language Models View paper
- [50] PCRI: Measuring Context Robustness in Multimodal Models for Enterprise Applications View paper
- [51] Common knowledge learning for generating transferable adversarial examples View paper
- [52] KDRSFL: A knowledge distillation resistance transfer framework for defending model inversion attacks in split federated learning View paper
- [53] On the benefits of knowledge distillation for adversarial robustness View paper
- [54] On transfer of adversarial robustness from pretraining to downstream tasks View paper
- [55] Initialization matters for adversarial transfer learning View paper
- [56] Reinforced compressive neural architecture search for versatile adversarial robustness View paper
- [57] Improving adversarial robustness using knowledge distillation guided by attention information bottleneck View paper
- [58] Improving Adversarial Robustness Through Adaptive Learning-Driven Multi-Teacher Knowledge Distillation View paper
- [59] Adversarially robust distillation View paper
- [60] Adversarially robust transfer learning View paper
- [61] On the duality between sharpness-aware minimization and adversarial training View paper
- [62] Solving the robustness puzzle: The joint impact of optimization approach, robustness metrics, and scenarios on water resources management under deep â¦ View paper
- [63] Balancing generalization and robustness in adversarial training via steering through clean and adversarial gradient directions View paper
- [64] Fortify the guardian, not the treasure: Resilient adversarial detectors View paper
- [65] Trade-off between robustness and accuracy of vision transformers View paper
- [66] Robustness and accuracy could be reconcilable by (proper) definition View paper
- [67] Improving the accuracy-robustness trade-off of classifiers via adaptive smoothing View paper
- [68] Optimizing Robustness and Accuracy in Mixture of Experts: A Dual-Model Approach View paper
- [69] R&D-Agent-Quant: A Multi-Agent Framework for Data-Centric Factors and Model Joint Optimization View paper
- [70] Enhancing infrared small target detection robustness with bi-level adversarial framework View paper
- [71] A preference-driven paradigm for enhanced translation with large language models View paper
- [72] Calibrated self-rewarding vision language models View paper
- [73] Positive enhanced preference alignment for text-to-image models View paper
- [74] Modality-balancing preference optimization of large multimodal models by adversarial negative mining View paper
- [75] Structured preference modeling for reinforcement learning-based fine-tuning of large models View paper
- [76] Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning View paper
- [77] Structured preference optimization for vision-language long-horizon task planning View paper

- [78] Aligning modalities in vision large language models via preference fine-tuning View paper
- [79] SAMPO: Visual Preference Optimization for Intent-Aware Segmentation with Vision Foundation Models View paper
- [80] TCPO: Thought-Centric Preference Optimization for Effective Embodied Decision-making View paper