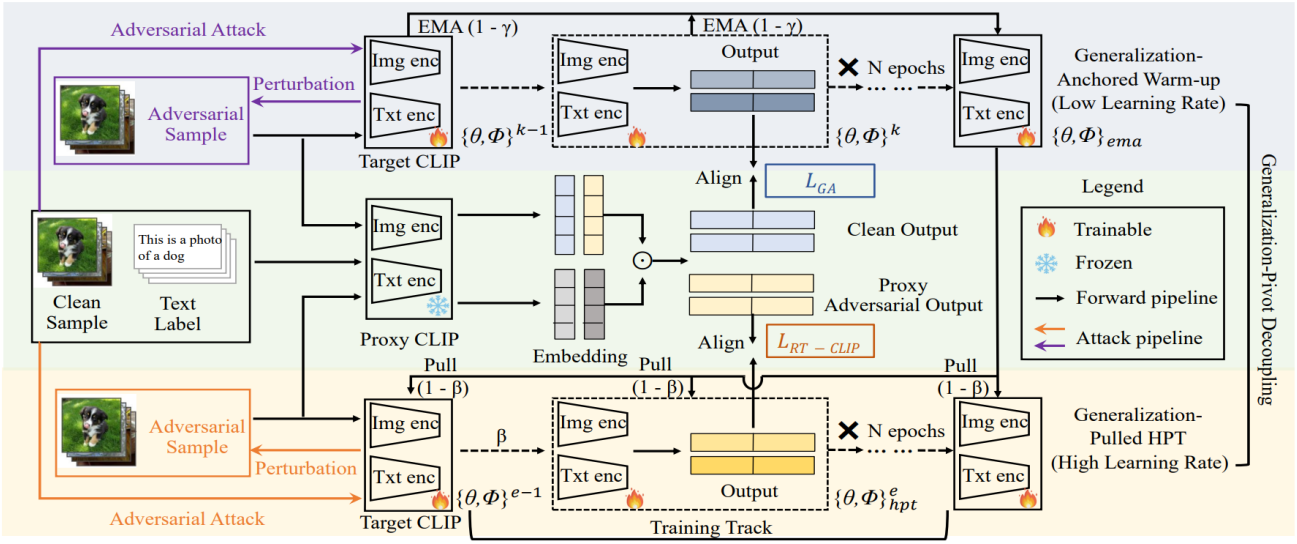
Proxy Robustness in Vision Language Models is Effortlessly Transferable

Xiaowei Fu, Fuxiang Huang, Lei Zhang

Frontiers of Computer Science, DOI: 10.1007/s11704-026-50951-1

Problems & Ideas

- Problems of conventional adversarial defense approaches:
  - Traditional adversarial defense either relies on computationally expensive adversarial training or requires a pre-trained robust teacher model. Both are impractical for large-scale Vision-Language Models.
- Ideas: We find that a standard CLIP model exhibits intrinsic robustness to adversarial examples generated by a different CLIP, and propose a lightweight framework that transfers this proxy robustness via heterogeneous proxy transfer.



The pipeline of the proposed HPT-GPD. In the generalization-anchored warm-up, the zero-shot generalization ability is maintained at a low learning rate; then, the proxy robust transfer is performed at a high learning rate, while the model obtained in the warm-up is used for generalization pulling.

Main Contributions

- Contributions:
  - We reveal that vanilla CLIP can provide proxy adversarial robustness for heterogeneous CLIP models, establishing the foundation for the adversarial robust transfer of VLMs;
  - We propose a heterogeneous proxy transfer framework via generalization pivot decoupling, which effortlessly enhances the zero-shot adversarial robustness of the target CLIP while preserving its natural generalization ability.
  - Experiments on 15 downstream datasets demonstrate that the proposed method offers signifficant improvements in both zero-shot adversarial robustness and natural generalization for CLIP.

Metrics	Method	CIFAR10	CIFAR100	STL-10	SUN397	Food101	OxfordPet	Flower102	DTD	EUROSAT	FGVC	TinyImageNet	ImageNet	Caltech101	Caltech256	StanfordCars	PCAM	Average	Time (s)
Adv acc	CLIP	42.18	17.57	67.77	10.49	3.28	8.98	4.88	18.75	3.84	0.39	0.39	6.09	15.82	23.76	2.34	51.61	17.38	0
	FT-Standard	24.21	12.30	47.65	5.71	2.65	1.56	7.61	17.57	0.13	0.00	0.39	6.87	**20.31**	23.82	1.75	48.15	13.79	406
	FT-TeCoA [14]	40.82	24.41	70.70	19.21	14.45	28.13	23.05	28.13	12.57	3.13	19.33	16.48	24.02	40.56	12.69	53.68	26.96	1112
	PMG [15]	67.06	33.31	77.91	11.13	13.91	42.00	27.58	29.63	21.25	3.30	15.15	9.99	14.82	41.21	13.32	57.65	29.95	1817
	HPT-GPD (ours)	**71.14**	**40.10**	**83.88**	**24.56**	**22.24**	**50.45**	**28.04**	**32.34**	**25.69**	**3.69**	**19.62**	**18.53**	19.11	**50.06**	**14.90**	**58.15**	**35.16**	1486
Clean acc	CLIP	87.43	60.39	97.08	57.79	82.93	86.95	65.72	39.90	41.12	20.50	56.06	58.24	19.42	79.43	51.60	53.52	59.88	0
	FT-Standard	88.67	58.78	95.89	46.61	70.07	75.00	41.60	41.79	38.60	8.57	58.59	51.28	29.10	79.88	37.10	54.74	54.76	406
	FT-TeCoA [14]	66.79	41.01	89.25	47.01	52.81	70.31	36.13	35.94	18.88	7.81	48.83	43.67	28.32	72.98	37.89	37.89	46.99	1112
	PMG [15]	82.80	52.28	92.99	58.78	70.52	84.46	56.77	35.11	28.58	15.75	46.41	54.86	20.98	76.82	48.04	57.70	55.18	1817
	HPT-GPD (ours)	90.74	64.83	95.60	58.79	75.05	81.68	50.11	34.26	38.32	14.07	61.56	53.35	22.87	79.63	45.00	58.14	57.75	1486

Average zero-shot adversarial accuracy (i.e., Adv acc) and clean accuracy (i.e., Clean acc) under PGD-10 attack.