Finetune Like You Pretrain: Boosting Zero-shot Adversarial Robustness in Vision-language Models

Supplementary Material

6. More Dataset Information

In the main paper (Tab. 5), we employ several popular variants of ImageNet that share the pre-defined classes partially or entirely, but have distinctly different data domains, to reflect the limitations of a leveraging a large extensive dataset with labelled classes as a proxy. These variants include ImageNet-R [14], ImageNet-A [15], ImageNet-S [49], ObjectNet [1]:

- **ImageNet-R**(endition) contains images of different renditions such as embroidery, paintings, toys, *etc*. It has 30,000 images from 200 pre-defined classes, which is a subset of the 1,000 classes of ImageNet. We use the textual prompt `'This is an artistic rendering of [CLS].'` when evaluating on this dataset.
- **ImageNet-A** contains natural image samples that standard models fail to classify. It has 7,500 images that belong to 200 pre-defined classes, which is a subset of the ImageNet classes. For evaluation, we employ the same textual prompt `'This is a photo of a [CLS].'` as in ImageNet.
- **ImageNet-Sketch** contains 50,000 images that are human-drawn black-and-white sketches. It has 1,000 pre-defined classes, which are the exact categories from ImageNet. For evaluation, we use the textual prompt of `'This is a sketch of a [CLS].'`.
- **ObjectNet** includes 50k real photographs of objects unusually arranged, such as varied camera angles, object poses, and diverse backgrounds. It has a total of 313 classes, of which 113 classes overlap with ImageNet. For evaluation, we employ the same textual prompt `'This is a photo of a [CLS].'` as in ImageNet.

7. Other Ablation Studies

In the main paper, we conduct ablative studies on the regularisation terms (Sec. 4.4). In this section, we perform ablative studies on other training settings.

7.1. Data Amount and Batch Size

We implement AdvFLYP$_{full}$ with varying amounts of image-text pairs and three batch sizes (256, 512 and 1024), and evaluate the performance of the finetuned model on 14 downstream datasets. Fig. 3 reports the results. It can be seen that both adversarial robustness and clean accuracy steadily increase as the model is finetuned on an increasingly large amount of image-text pairs. In the main paper, we employ one million noisy image-text pairs collected from the web and fix the training data amount to 1M to reduce training time. The performance of AdvFLYP$_{full}$ is projected to continue to further improve if we enlarge the training data amount further. As opposed to previous AFT methods that finetune CLIP via a cross-entropy loss, where the batch size does not play an important role, AdvFLYP is a contrastive finetuning paradigm, where samples are contrasted with each other in a batch, and a large batch size benefits model performance because it provides more negative examples [40]. The original paper of CLIP adopts the batch size of around 32k during pretraining to guarantee sufficient negative examples within a single batch. In this work, we are not able to adopt the same batch size for adversarial finetuning due to hardware constraints. Nonetheless, we observe a different pattern of impact of the batch size in our adversarial contrastive finetuning paradigm. Experimentally, we find that a batch size smaller than 256 leads to poor performance in terms of robustness and clean accuracy on downstream datasets, which implies that a sufficiently large batch size is also crucial for effective contrastive learning in our AdvFLYP paradigm. However, when the batch size is further increased (512 and 1024), there is a trade-off between robustness and clean accuracy, with a larger batch size benefiting the clean accuracy while suppressing the robustness gains, to some extent. One possible reason is that for a larger batch with more examples, it is more demanding to optimise the batch-wise perturbations (Eq. (6)) that maximise the con-

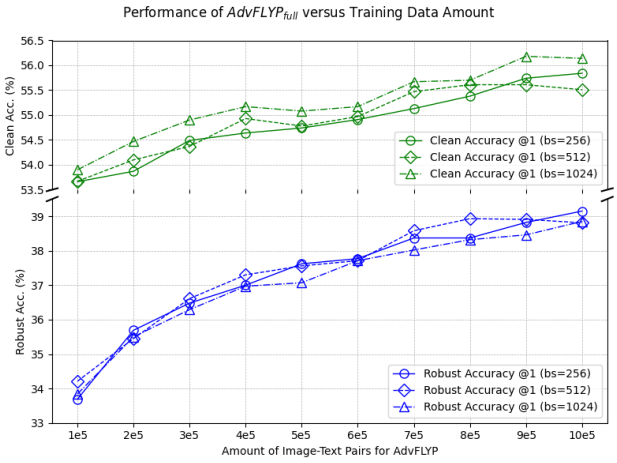


Figure 3. Performance of AdvFLYP$_{full}$ averaged over 14 downstream datasets versus the amount of image-text pairs from the web. The robust accuracy is evaluated under PGD-10 ($\epsilon = 1/255$).

(%)	PGD	CW	AA	*clean*
f_θ, g_ϕ	35.67	34.72	33.89	53.81
all	35.71	34.63	33.76	53.97
f_θ (ours)	**39.15**	**38.30**	**37.74**	**55.84**

Table 6. We unfreeze more modules of CLIP in our AdvFLYP$_{full}$. We evaluate the finetuned models using the attack methods of PGD, CW, and AutoAttack with $\epsilon = 1/255$.

(%)	PGD	AA	clean
AdvFLYP$_{2nd}$	32.71	31.25	49.01
AdvFLYP$_{full}$	**33.69**	**31.79**	**53.66**

Table 7. Adversarial robustness evaluated at $\epsilon = 1/255$ and clean accuracy averaged over 14 datasets. Both variants are trained with 100k image-text pairs.

trastive loss, leading to less difficult adversarial examples. Therefore, finetuning f_θ based on these examples leads to lesser robustness gains, while better retaining the zero-shot capabilities on clean images. In the main paper, we fix the batch size at 256.

7.2. More Tunable Modules

Despite the fact that we stick to the training recipe (mainly the training data distribution and training objective) of CLIP's pretraining, AdvFLYP differs from the pretraining process in that it only finetunes the pre-trained vision encoder f_θ of CLIP, and keeps all other modules of CLIP frozen. In this section, we unfreeze more modules of CLIP and implement AdvFLYP. Tab. 6 reports the performance of AdvFLYP$_{full}$ when we finetune the text encoder g_ϕ, and all modules in addition to the vision encoder f_θ. It can be seen that both variants degrade the performance of AdvFLYP$_{full}$ significantly, which indicates that the vision encoder f_θ is the component of central importance to robustfying CLIP. We speculate that the degradation is due to the unnecessary distortion of the text encoder g_ϕ. Considering that the adversarial images are only fed to the vision encoder f_θ, intuitively, one should only finetune f_θ to ensure that these adversarial images are aligned with the correct text supervision signals. Therefore, finetuning g_ϕ and other modules in addition to the vision encoder f_θ does not benefit the overall performance.

7.3. Regularisation Formulation

In the main paper, we formulate regularisation by computing (i) the deviation between adversarial and clean image features by the target model f_θ, and (ii) the deviation between the adversarial image features by the target model f_θ and the original CLIP F_{θ_0}. This applies to both logit-level (Eq. (12)) and feature-level (Eq. (8)) regularisation. Prior work on adversarial defence has employed the second term (ii) to defend neural networks [25]. As a preliminary experiment, we evaluate whether using the second term (ii) of Eq. (12) and Eq. (8) suffices to boost adversarial robustness of the target model. Specifically, we randomly sample 100k training data, on which we perform regularised AdvFLYP with only the (ii) terms for both logit and feature levels, and denote this variant as AdvFLYP$_{2nd}$. The results for this variant are reported in Tab. 7. From the table, it can be concluded that the first term (i) is also important, especially for the retention of clean accuracy, hence the complete formulation of the proposed regularisation in the main paper.

(%)	PGD	CW	AA	*clean*
TeCoA	**33.77**	**33.04**	**32.17**	51.35
naive FLYP	32.89	32.28	31.23	**51.66**
PMG-AFT	**35.85**	**34.87**	**33.89**	**54.25**
naive FLYP + $\mathcal{L}_{logit}$	35.28	34.35	33.33	53.87

Table 8. We implement 'naive FLYP' in our adversarial finetuning context and compare to the previous standard AFT paradigm TeCoA [30]. PMG-AFT [51] is equivalent to TeCoA + $\mathcal{L}_{logit}$. We evaluate the robustness of the finetuned models using PGD, CW, and AutoAttack with $\epsilon = 1/255$ and report the average results over 14 downstream datasets.

	(%)	AutoAttack ($\epsilon = 1/255$)	Clean	Avg.
	CLIP	0.05	62.03	31.04
(a)	TeCoA	32.17	51.35	41.76
(b)	TeCoA + $\mathcal{L}_{logit}$	**33.89**	**54.25**	**44.07**
(c)	TeCoA + $\mathcal{L}_{feat}$	32.94	52.31	42.63
(d)	TeCoA + $\mathcal{L}_{logit}$ + $\mathcal{L}_{feat}$	33.88	54.22	44.05

Table 9. Results of the models finetuned with different combinations of regularisation levels on top of TeCoA [30]. The combination (*b*) is equivalent to PMG-AFT [51]. The reported results are the average accuracy over 14 downstream datasets.

8. More Experimental Results

8.1. Robustness under Higher Attack Budgets

We report the full tables of robustness evaluated under the attack strength of $\epsilon = 2/255$ and $\epsilon = 4/255$ in Tab. 10 and Tab. 11, respectively. It can be seen that our AdvFLYP still consistently outperforms previous the previous AFT paradigm TeCoA [30] and its regularisation-based advancements PMG-AFT [51] and TGA-ZSR [56] under strong attack budgets, showing the reliability of our simple paradigm. In contrast, the method that formulates regularisation based on text-guided attention (TGA-ZSR [56]) is shown to be

	(%)	CIFAR10	CIFAR100	STL10	Caltech101	Caltech256	OxfordPets	Flowers102	Food101	StanfordCars	SUN397	Country211	FGVCAircraft	EuroSAT	DTD	avg.
PGD	CLIP	0.00	0.00	0.71	1.71	0.84	0.00	0.02	0.01	0.00	0.02	0.00	0.00	0.00	0.21	0.25
	FARE	1.11	0.85	15.21	22.14	14.02	3.41	1.12	0.72	0.15	1.79	0.04	0.00	2.69	5.11	4.88
	TeCoA	11.66	6.75	46.08	51.03	37.80	31.32	14.15	9.03	3.44	13.14	0.92	1.08	9.85	14.20	17.89
	PMG-AFT	12.75	7.68	47.84	51.15	38.41	**32.98**	15.09	10.34	4.42	13.29	0.89	1.59	**10.31**	14.52	18.66
	TGA-ZSR	2.99	1.34	27.29	33.30	23.64	19.13	4.15	4.26	1.08	5.67	0.21	0.09	0.56	6.60	9.31
	AdvFLYP	19.32	9.63	52.24	52.96	41.13	27.61	18.13	12.40	9.15	16.51	0.99	1.80	0.60	**16.70**	19.94
	AdvFLYP$_{full}$	**26.15**	**13.27**	**57.28**	**55.23**	**42.99**	29.54	**18.95**	**14.54**	**9.35**	**17.48**	**1.08**	**1.92**	3.88	16.65	**22.02**
CW	CLIP	0.00	0.00	0.72	6.33	0.71	0.00	0.00	0.01	2.30	0.02	0.00	0.00	0.00	0.37	0.75
	FARE	1.36	1.16	17.02	28.48	16.15	3.82	1.50	1.15	2.52	2.56	0.06	0.00	2.34	4.63	5.91
	TeCoA	12.20	6.96	47.42	54.35	39.23	33.66	13.95	9.82	4.84	14.40	0.96	1.41	9.23	12.98	18.67
	PMG-AFT	12.46	7.39	49.05	55.77	39.89	**34.97**	13.50	11.56	5.55	14.11	0.87	1.14	**10.13**	13.14	19.25
	TGA-ZSR	3.37	1.68	28.49	37.71	24.45	18.64	4.94	4.96	3.11	6.69	0.34	0.27	0.54	5.90	10.08
	AdvFLYP	20.07	10.44	53.31	57.20	42.68	30.53	**18.78**	14.51	**12.19**	17.99	1.12	**2.13**	0.63	**16.22**	21.27
	AdvFLYP$_{full}$	**24.64**	**12.30**	**57.89**	**58.19**	**44.12**	32.43	18.56	**16.81**	11.33	**18.34**	**1.13**	1.92	4.69	15.37	**22.69**
AutoAttack	CLIP	0.00	0.06	0.00	0.05	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.03	0.08	0.05	0.02
	FARE	0.35	0.57	9.72	17.07	10.86	1.72	0.75	0.49	0.14	0.97	0.01	0.03	1.97	3.30	3.42
	TeCoA	9.97	5.83	44.46	49.31	36.09	29.63	12.44	7.58	2.71	11.26	0.66	0.72	8.97	12.39	16.57
	PMG-AFT	10.34	6.17	45.92	49.64	36.53	**30.64**	12.15	8.60	3.00	10.98	0.59	0.87	**9.86**	12.77	17.00
	TGA-ZSR	0.00	0.00	0.01	0.06	0.02	0.00	0.03	0.00	0.00	0.01	0.01	0.00	0.10	0.05	0.02
	AdvFLYP	17.87	8.88	51.29	51.91	39.96	26.38	**17.01**	11.33	**8.13**	15.10	**0.78**	**1.32**	0.34	**15.85**	19.01
	AdvFLYP$_{full}$	**23.02**	**11.18**	**56.22**	**53.90**	**41.29**	27.42	16.95	**13.18**	7.72	**15.33**	0.74	1.29	1.83	14.84	**20.35**
$\overline{\text{AVG}}$	CLIP	0.00	0.02	0.48	2.69	0.52	0.00	0.01	0.01	0.77	0.01	0.00	0.01	0.03	0.21	0.34
	FARE	0.94	0.86	13.99	22.56	13.68	2.98	1.12	0.79	0.94	1.77	0.04	0.01	2.33	4.34	4.74
	TeCoA	11.28	6.51	45.99	51.57	37.71	31.53	13.51	8.81	3.66	12.94	0.85	1.07	9.35	13.19	17.71
	PMG-AFT	11.85	7.08	47.60	52.18	38.27	**32.86**	13.58	10.17	4.32	12.80	0.78	1.20	**10.10**	13.48	18.31
	TGA-ZSR	2.12	1.01	18.60	23.69	16.04	12.59	3.04	3.07	1.40	4.12	0.19	0.12	0.40	4.18	6.47
	AdvFLYP	19.09	9.65	52.28	54.02	41.26	28.17	17.98	12.74	**9.82**	16.53	0.96	**1.75**	0.52	**16.26**	20.07
	AdvFLYP$_{full}$	**24.60**	**12.25**	**57.13**	**55.77**	**42.80**	29.80	**18.15**	**14.84**	9.47	**17.05**	**0.98**	1.71	3.47	15.62	**21.69**

Table 10. Classification accuracy (%) on 14 downstream datasets tested with three adversarial attack algorithms ($\epsilon = 2/255$). We highlight the **best** and second best result.

less effective on higher attack strengths, implying that it may have overfit to the attack strength ($\epsilon = 1/255$) during adversarial finetuning. On average, under the attack budget of $\epsilon = 2/255$, our basic AdvFLYP paradigm and its regularised variant AdvFLYP$_{full}$ achieve the robust accuracy of 20.07% and 21.69%, outperforming PMG-AFT (18.31%) by a relative margin of 9.61% and 18.46% over 14 downstream datasets and various attack methods, respectively. When evaluated under the budget of $\epsilon = 4/255$, AdvFLYP and AdvFLYP$_{full}$ achieve an average robustness of 5.87% and 5.93%, both outperforming PMG-AFT (4.00%). Results show that the paradigm of AdvFLYP is a competitive adversarial finetuning paradigm for VLMs despite its sheer simplicity, compared to the standard practice of finetuning VLMs on a large and extensive dataset with labelled classes such as ImageNet.

8.2. Naive FLYP for AFT

This work draws inspiration from FLYP [10], which finds that finetuning CLIP with a contrastive loss as employed in the pretraining process helps to improve generalisation to *out-of-distribution* (OOD) data. FLYP challenges previous standard finetuning practices that finetune CLIP with a cross-entropy loss on *in-distribution* (ID) data, and leverages a contrastive loss instead. However, they still operate on classification-oriented ID data, and ignore the overlap of classes present in a batch. In contrast, AdvFLYP aims to underscore the importance of following the data distribution and the training objective of VLMs' pretraining to boost its zero-shot adversarial robustness. Therefore, despite the fact that AdvFLYP stems from the same spirit as FLYP, its motivation and implementation differ fundamentally from those of FLYP, and is no simple extension of FLYP in the context of adversarial robustness. In this section, we naively apply FLYP for adversarial finetuning. Specifically, we finetune

	(%)	CIFAR10	CIFAR100	STL10	Caltech101	Caltech256	OxfordPets	Flowers102	Food101	StanfordCars	SUN397	Country211	FGVCAircraft	EuroSAT	DTD	avg.
PGD	CLIP	0.00	0.00	0.04	0.61	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.06
	FARE	0.06	0.02	1.42	5.05	2.04	1.55	0.02	0.02	0.00	0.03	0.00	0.00	0.00	0.37	0.76
	TeCoA	0.72	0.79	9.39	20.42	12.46	**3.08**	2.00	0.63	0.06	2.06	**0.10**	0.00	5.25	4.63	4.40
	PMG-AFT	0.53	0.89	9.91	20.02	12.24	2.81	1.74	0.76	0.06	1.92	0.07	**0.03**	**6.34**	4.84	4.44
	TGA-ZSR	0.04	0.01	2.10	6.37	3.47	2.10	0.05	0.10	0.00	0.22	0.01	0.00	0.00	0.80	1.09
	AdvFLYP	2.08	1.92	19.50	**27.69**	17.20	2.94	**3.27**	1.23	**0.85**	**3.30**	0.06	**0.03**	0.01	6.44	6.18
	AdvFLYP$_{full}$	**2.91**	**2.97**	**22.19**	27.28	**17.29**	2.73	2.91	**1.34**	0.57	3.11	0.08	0.00	0.07	**6.81**	**6.45**
CW	CLIP	0.00	0.00	0.04	4.94	0.07	0.00	0.00	0.00	2.41	0.00	0.00	0.00	0.00	0.11	0.54
	FARE	0.01	0.08	1.55	8.15	1.87	0.03	0.00	0.02	2.19	0.04	0.00	0.00	0.00	0.32	1.02
	TeCoA	0.80	1.00	11.02	25.58	14.50	**4.80**	2.19	0.87	1.52	3.11	**0.14**	0.03	5.54	4.10	5.37
	PMG-AFT	0.43	0.98	10.66	25.85	14.14	4.58	1.61	0.93	1.74	2.71	0.13	0.03	**6.75**	3.72	5.30
	TGA-ZSR	0.05	0.05	1.86	8.61	3.12	0.33	0.05	0.08	1.60	0.33	0.00	0.00	0.00	0.48	1.18
	AdvFLYP	2.19	2.32	20.36	**33.55**	**18.97**	4.42	**3.80**	1.76	**3.50**	**4.36**	0.11	**0.12**	0.00	**6.33**	7.27
	AdvFLYP$_{full}$	**2.51**	**2.94**	**22.89**	33.54	18.65	4.74	3.40	**1.97**	2.72	4.09	0.12	0.09	0.05	5.32	**7.36**
AutoAttack	CLIP	0.00	0.06	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03	0.08	0.05	0.02
	FARE	0.00	0.04	0.00	1.30	0.37	0.00	0.02	0.00	0.00	0.00	0.00	**0.03**	0.04	0.11	0.14
	TeCoA	0.10	0.24	4.28	14.34	8.59	**1.64**	1.06	0.28	0.11	0.92	**0.05**	0.00	0.08	3.14	2.49
	PMG-AFT	0.04	0.26	3.76	13.39	7.56	1.14	0.80	0.31	0.01	0.76	**0.05**	**0.03**	**0.38**	3.14	2.26
	TGA-ZSR	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.00	0.01	0.01	0.00	0.10	0.05	0.02
	AdvFLYP	0.50	0.91	10.98	**20.96**	**12.78**	1.55	**2.28**	**0.65**	**0.51**	**1.67**	**0.05**	**0.03**	0.01	**5.27**	**4.15**
	AdvFLYP$_{full}$	**0.65**	**1.42**	**12.65**	20.02	11.46	1.09	1.76	0.63	0.26	1.30	0.04	0.00	0.04	4.47	3.98
$\overline{\text{AVG}}$	CLIP	0.00	0.02	0.02	1.85	0.06	0.00	0.01	0.00	0.80	0.00	0.00	0.01	0.03	0.09	0.21
	FARE	0.02	0.05	0.99	4.83	1.42	0.53	0.01	0.01	0.73	0.03	0.00	0.01	0.01	0.27	0.64
	TeCoA	0.54	0.68	8.23	20.11	11.85	**3.17**	1.75	0.59	0.56	2.03	**0.10**	0.01	3.62	3.95	4.09
	PMG-AFT	0.33	0.71	8.11	19.75	11.31	2.84	1.38	0.67	0.61	1.80	0.08	0.03	**4.49**	3.90	4.00
	TGA-ZSR	0.03	0.02	1.32	5.00	2.20	0.81	0.04	0.06	0.53	0.19	0.01	0.00	0.03	0.44	0.76
	AdvFLYP	1.59	1.72	16.95	**27.40**	**16.32**	2.97	**3.12**	1.21	**1.62**	**3.11**	0.07	**0.06**	0.01	**6.01**	5.87
	AdvFLYP$_{full}$	**2.02**	**2.44**	**19.24**	26.94	15.80	2.85	2.69	**1.31**	1.19	2.84	0.08	0.03	0.06	5.53	**5.93**

Table 11. Classification accuracy (%) on 14 downstream datasets tested with three adversarial attack algorithms ($\epsilon = 4/255$). We highlight the **best** and second best result.

CLIP's vision encoder f_θ on ImageNet. Instead of employing a cross-entropy loss as in TeCoA [30] and PMG-AFT [51], we follow the implementation of FLYP and leverage the contrastive loss. As in FLYP, we ignore the fact that some classes may overlap in one batch. We report the results in Tab. 8. It can be seen that performing FLYP naively does not lead to improved robustness or better clean accuracy in the context of adversarial finetuning. In contrast, by performing real contrastive finetuning on adversarial web images and their corresponding texts, our AdvFLYP paradigm achieves significantly improved robustness compared to previous AFT methods.

8.3. Other Vision Backbones

In the main paper, we focus on the CLIP ViT-B/32 model, following the practice of prior work [30, 51, 56]. The AdvFLYP paradigm can be readily employed to boost the adversarial robustness of other CLIP backbones and other CLIP-style VLMs. In this section, we conduct a preliminary experiment on CLIP ViT-B/16 with 100k image-text pairs to show that AdvFLYP achieves consistent improvement over previous paradigms. Results reported in Tab. 12 show that the AdvFLYP paradigm consistently outperforms previous AFT

(%)	PGD	AA	clean
PMG-AFT	31.59	28.92	54.46
AdvFLYP$_{full}$	**35.13**	**33.62**	**56.07**

Table 12. Robustness ($\epsilon = 1/255$) and clean accuracy of CLIP ViT-B/16 averaged over 14 datasets.

paradigms on other vision backbones.

9. Training Data Analysis

The current *de facto* standard paradigm for finetuning VLMs to achieve zero-shot adversarial robustness is largely based on the adversarial training (AT) principles of classical adversarial learning [28], which involve a dataset of labelled classes. This paradigm is reasonable in the sense that the finetuned CLIP is to be deployed in downstream classification datasets. However, we believe that adversarial finetuning should not be considered as a separate process from the pretraining of VLMs. In the pretraining of CLIP, the encoders are trained to match a batch of noisy web images with their corresponding texts. Therefore, we propose to finetune the model to match the adversarial images with corresponding texts over web-scale image-text data. Our aim is to rethink the current standard AFT paradigm and present a new paradigm that is simpler, more intuitive and yet more effective than the standard AFT paradigm. This section investigates the impact of the training data in more depth.

9.1. Impact on Non-Adversarial Finetuning

Finetuning the model weights of pretrained VLMs, even with clean non-adversarial data, can already compromise the generalisation of the model. We conduct a preliminary experiment to reveal the importance of following CLIP's pretraining data distribution. Specifically, we collect 100k noisy image-text pairs from the web, and utilise them to finetune f_θ without creating adversarial images. As a reference, we randomly sample 100 images per class from ImageNet, resulting in 100k labelled images in total. We employ this subset of images to finetune f_θ. For both toy datasets, we finetune for 10 epochs with the learning rate of $5e-5$. We report the clean accuracy in Tab. 13. From this toy preliminary experiment, it can be seen that despite its extensive and less noisy nature, finetuning f_θ on the clean images of ImageNet already causes a slight degradation of generalisation. In contrast, when finetuning f_θ on web-scale image-text pairs, the zero-shot performance of the model even slightly improves. This highlights the importance of following the pretraining data distribution when modifying the model weights of VLMs, whereas treating finetuning as a separate process from pretraining of VLMs and modifying model weights is not an ideal choice.

9.2. Image-Text Pairs from ImageNet

In this section, we employ a generative VLM Qwen2.5-VL-3B-Instruct to generate a semantically-rich textual description for each image in ImageNet. We use the prompt of `'Describe this image with no more than 50 words'`. We provide a captioned example of an training image in Fig. 4. It can be seen that the chosen generative VLM is able to produce highly informative and coherent textual descriptions for images. We then leverage the ImageNet dataset with the textual descriptions to perform AFT. As in our proposed paradigm, we employ the contrastive loss with the batch size of 256, and impose regularisation on both logit and feature level. Results reported in Tab. 14 show that performing AFT on the captioned ImageNet with a contrastive loss mitigates the memorisation of the finetuning data (with a lower reported number on ImageNet), while improving the robustness across downstream datasets to a limited extent. In comparison, AdvFLYP$_{full}$ outperforms the variant that performs AFT on captioned images of ImageNet, despite leveraging noisy web-scale image-text pairs, showing the

In comparison, our AdvFLYP$_{full}$ paradigm consistently outperforms both baselines, showing the superiority of our simple paradigm of following the pretraining behaviour in AFT instead of treating them as separate processes.

(a)

A man wearing a brown beanie and green jacket holds a large, colorful carp fish by the riverbank. The fish has a vibrant orange and blue pattern on its scales.

(b)

Figure 4. An example of an image from ImageNet and its description generated by Qwen2.5-VL-3B-Instruct.

10. Discussion on Feature Regularisation

Wang et al. [51] propose logit-level regularisation on top of TeCoA, showing that it boosts the generalisation of robustness and clean accuracy across downstream datasets by penalising the logit discrepancy between the adversarial logits by the target model f_θ and adversarial logits by the frozen pretrained vision encoder F_{θ_0} (first term of Eq. (12)), and the logit discrepancy between adversarial logits by the target model and the clean logits by the target model (second term of Eq. (12)). We conduct a trial experiment by imposing feature-level regularisation on top of TeCoA, and report the results in Tab. 9. The results reveal different behaviour of the TeCoA [30] and our AdvFLYP paradigms. Penalising discrepancy on the logit level achieves significantly large improvement over TeCoA (compare *a* and *b*) in terms of both downstream robustness and clean accuracy, whereas imposing regularisation on the feature level brings marginal effects (compare *a* and *c*). In contrast, as reported in Tab. 4 in the main paper, logit- and feature-level penalties

(%)	ImageNet	CIFAR10	CIFAR100	STL10	Caltech101	Caltech256	OxfordPets	Flowers102	Food101	StanfordCars	SUN397	Country211	FGVCAircraft	EuroSAT	DTD	avg.
CLIP	59.75	85.05	57.18	96.41	86.19	82.04	87.27	65.62	83.83	52.13	58.87	15.26	20.16	38.32	40.11	62.03
web data	59.69	84.31	58.77	95.92	86.57	82.42	87.33	65.56	83.06	52.94	60.67	15.59	21.15	34.58	41.12	62.14
IN	67.19	84.33	58.14	96.52	85.02	82.06	87.16	62.29	77.99	46.61	59.27	13.69	17.43	35.83	40.11	60.46

Table 13. The clean accuracy of the models finetuned on 100k image-text web data and a 100k-subset of ImageNet, respectively. Both baselines are finetuned with clean non-adversarial images.

(%)	ImageNet	CIFAR10	CIFAR100	STL10	Caltech101	Caltech256	OxfordPets	Flowers102	Food101	StanfordCars	SUN397	Country211	FGVCAircraft	EuroSAT	DTD	avg.
PMG-AFT	37.17	39.23	19.40	76.79	71.14	59.34	62.42	30.69	29.86	14.48	29.26	2.59	4.65	12.83	21.76	33.89
AFT$_{capIN}$	34.34	49.71	26.08	82.36	71.82	60.37	64.95	34.33	37.27	19.28	28.96	2.71	4.77	11.12	23.30	36.93
AdvFLYP$_{full}$	29.47	50.15	25.99	78.50	72.47	61.31	59.88	37.81	37.53	24.61	34.44	3.46	5.73	11.46	24.95	37.74

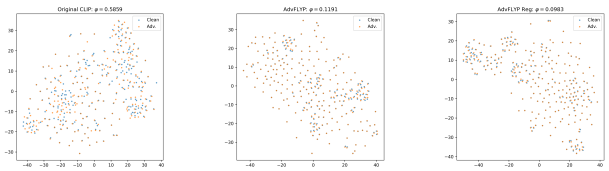
Table 14. Robustness of the models finetuned on class-labelled ImageNet (PMG-AFT), VLM-captioned ImageNet (AFT$_{capIN}$), and our AdvFLYP$_{full}$, evaluated with AutoAttack ($\epsilon = 1/255$).

play different roles, with $\mathcal{L}_{logit}$ benefiting transferability of robustness gains across downstream datasets and $\mathcal{L}_{feat}$ facilitating preservation of zero-shot capabilities on clean images. We believe there are two main reasons. Firstly, the prior TeCoA paradigm caters to the classification task, where the logit is the key element. Additionally, they produce adversarial images that maximise the cross-entropy loss w.r.t. a pre-defined set of categories. This may not cause a significant shift of the embeddings in the latent space. In contrast, our AdvFLYP creates adversarial images based on noisy web images w.r.t. their texts, which can result in considerable embedding shifts. Finetuning f_θ with these distorted adversarial embeddings can contaminate the vision encoder. Therefore, penalising the deviation of image features with $\mathcal{L}_{feat}$ is crucial for retaining CLIP's zero-shot performance on clean images.

To further investigate the effects of imposing regularisation over AdvFLYP, we analyse the cosine deviation of clean and adversarial features of AdvFLYP and the regularised variant AdvFYLP$_{full}$. We define the cosine deviation as follows:

$$\varphi = \arccos \frac{f_\theta(x)^\intercal f_\theta(x+\delta)}{\|f_\theta(x)\| \cdot \|f_\theta(x+\delta)\|} \tag{14}$$

where a larger φ indicates greater cosine deviation of adversarial image features from their clean counterparts in the latent space, and vice versa. Specifically, we sample 256 images from ImageNet and employ the t-SNE algorithm visualise the adversarial and clean image features for AdvFLYP and AdvFLYP$_{full}$. As can be seen from Fig. 5, adversarial image features deviate significantly from their clean counterparts in the latent space of the original CLIP, with average cosine deviation at $\varphi = 0.5859$, whereas AdvFLYP and its regularised variant AdvFLYP$_{full}$ effectively mitigating such deviation, as evidenced by the narrowed gap between adversarial and clean features. Imposing regularisation (AdvFLYP$_{full}$) further mitigates the cosine deviation as φ is decreased to 0.0983, in comparison to $\varphi = 0.1191$ achieved by AdvFLYP with no regularisation terms.



(a) Original CLIP (b) AdvFLYP (c) AdvFLYP$_{full}$

Figure 5. t-SNE visualisation of adversarial and clean image features in the latent space.