CVPR JUNE 3-7, 2026 DENVER COLORADO

The Geometry of Robustness: Optimizing Loss Landscape Curvature and Feature Manifold Alignment for Robust Finetuning of Vision-Language Models

Shivang Chopra, Shaunak Halbe, Chengyue Huang, Brisa Maneechotesuwan, Zsolt Kira

Georgia Institute of Technology

Georgia Tech

Problem

Three-Way Trade-off in VLM Fine-tuning

Fine-tuning Vision-Language-Models (VLMs) improves downstream accuracy but often exposes a persistent three-way trade-off among in-distribution accuracy v/s OOD generalization v/s adversarial robustness.

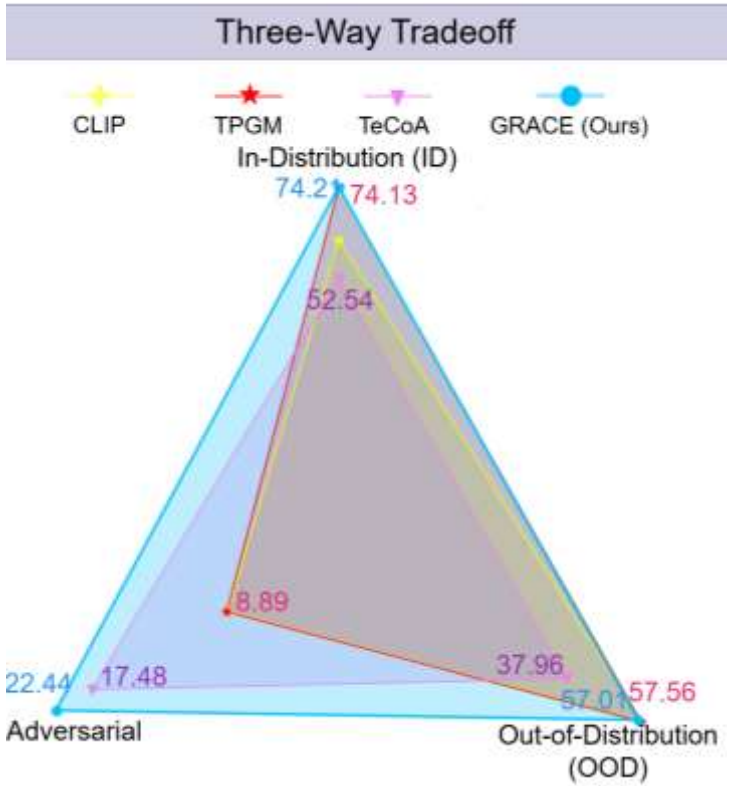


Figure 1: The VLM robustness three-way tradeoff. Existing robust fine-tuning strategies resolve at most two of {ID, OOD, adversarial} robustness simultaneously, leaving a gap in generalized robustness. GRACE is designed to close this gap.

In-distribution Accuracy v/s OOD Generalization v/s Adversarial Robustness

Tian et al. Trainable Projected Gradient Method for Robust Fine-tuning. CVPR 2023
Mao et al. Understanding Zero-shot Adversarial Robustness for Large-Scale Models. ICLR 2023

Georgia Tech

Key Objectives

Objective 1

- Understand why this trade-off arises in prior approaches and how to mitigate it.

Objective 2

- Develop a fine-tuning framework that improves ID accuracy, and adversarial robustness while preserving OOD generalization.

Georgia Tech

Theoretical Analysis

PAC-Bayesian Framework for Multi-Domain Robustness

Problem Setup

Consider fine-tuning a pre-trained VLM with parameters $\theta_0 \in \mathbb{R}^d$ to parameters θ using dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$.

For domain $s \in \mathcal{S} = \{\text{ID}, \text{OOD}, \text{Adv}\}$, define risk:

$$R_s(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}[\ell(f_\theta(x), y)]$$

The robust risk is $R_{\text{Rob}}(\theta) = \max_{s\in\mathcal{S}} R_s(\theta)$

Theorem 1: Robust PAC-Bayes Bound

$$R_{Rob}(\theta) \leq \hat{R}_{ID}(\theta) + \underbrace{\frac{\|\theta - \theta_0\|^2}{2n\sigma^2} + \frac{\ln(2n/\delta)}{2n}}_{\text{(A) Proximity from prior}} + \underbrace{\frac{\sigma^2}{2} \cdot Tr(\mathbb{E}[\nabla_\theta^2 R_{Rob}(\theta)])}_{\text{(B) Parameter-space sharpness}} + \underbrace{\max_{s,t\in\mathcal{S}} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t)}_{\text{(C) Cross-domain discrepancy}} + \lambda^*$$

Georgia Tech

Neyshabur et al. A Pac-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. ICLR 2018

Predicted Failure Modes

Loss of proximity from Pretrained Model

Sharp Minima in the Loss Landscape

Unstable Features Across Distribution Shifts

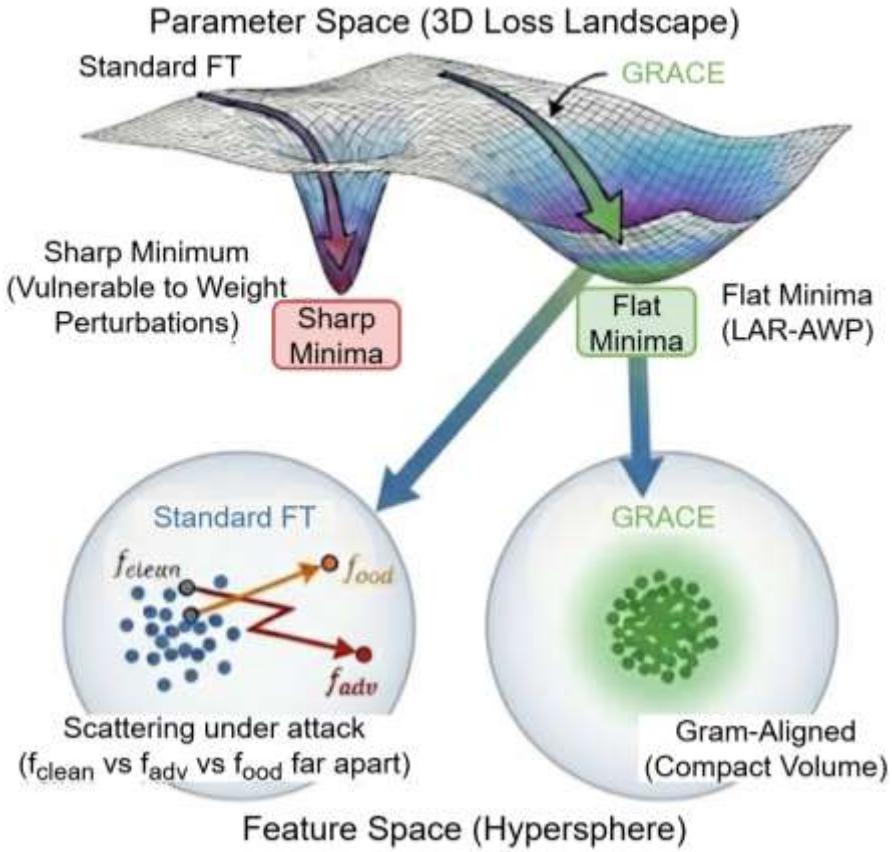
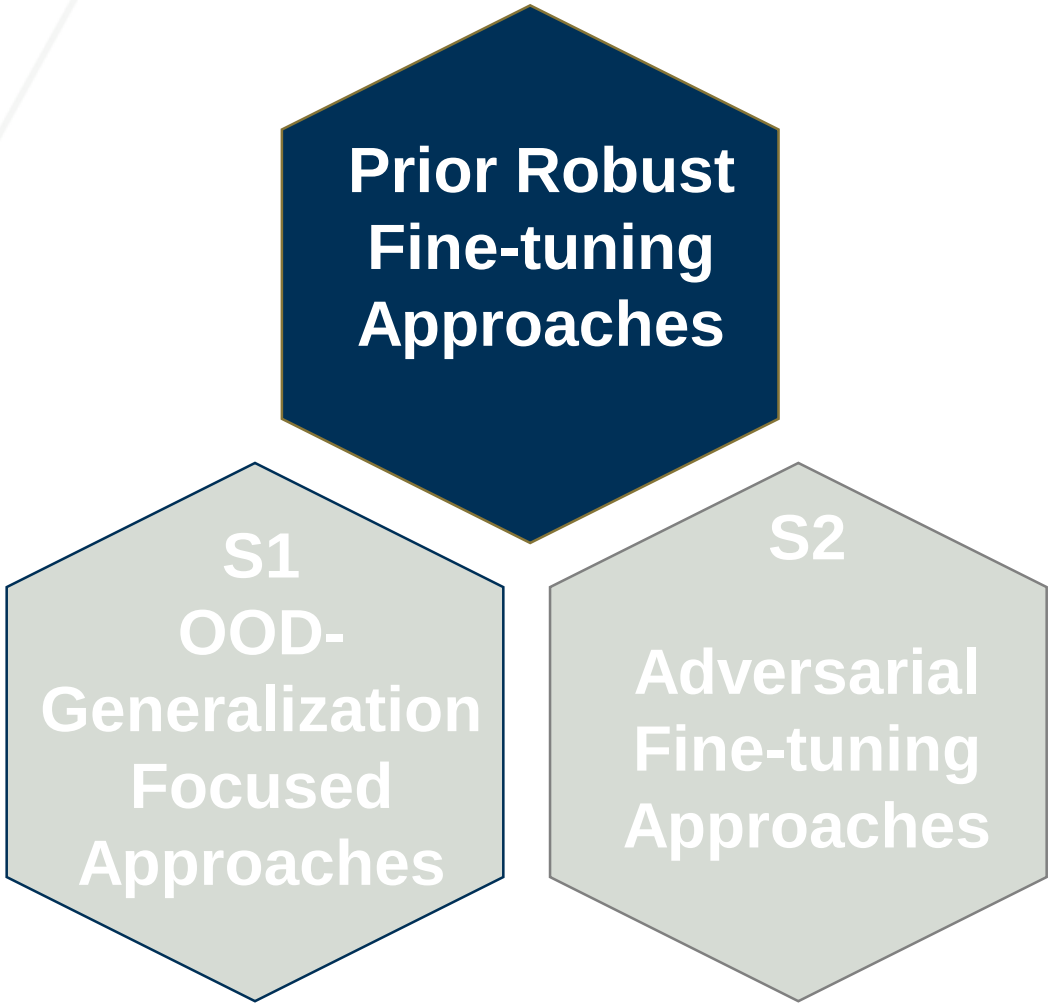


Figure 2: Visual illustration of the predicted geometric failure modes resulting in the three-way trade-off.

Georgia Tech

Analyzing Prior Robust Finetuning Approaches



Method	(A) Proximity $\|\theta - \theta_0\|^2$	(B) Sharpness $\mathrm{Tr}(\nabla^2 R)$	(C) Stability $d_{H\Delta H}$
Vanilla FT	✗	✗	✗
WiSE-FT [36]	✓	✗	✗
FLYP [5]	✓	✗	~
TPGM [28]	✓	✗	✗
SPD [29]	✓	✗	✗
TeCoA [17]	✗	~	~
FARE [26]	✗	~	✓
PMG-AFT [34]	✓	~	✓
LAAT [13]	✗	✓	✓
GRACE (Ours)	✓	✓	✓

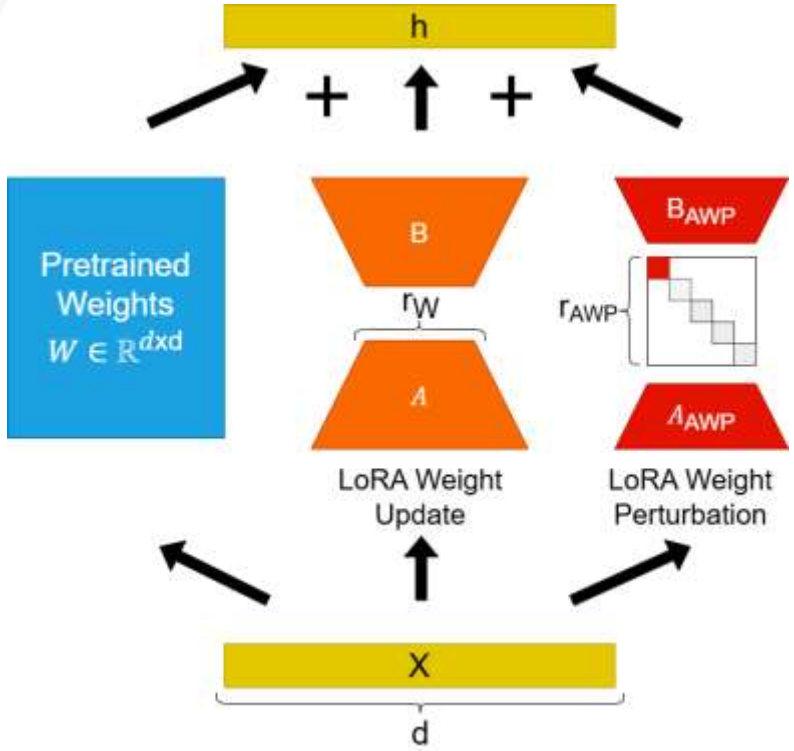
Table 1: Analysis of prior robust fine-tuning approaches based on their optimization objectives. Methods addressing each term explicitly through their loss function or training procedure are marked (✓), those with implicit or partial coverage are marked (~), and those without explicit optimization are marked (✗).

Georgia Tech

GRACE: Gram-aligned Robustness via Adaptive Curvature Estimation

Objective 1: Maintain Proximity to Pretrained Model

LoRA Finetuning



$$W(\theta) = W(\theta_0) + B_W A_W, \qquad A_W \in \mathbb{R}^{r \times n_2},\ B_W \in \mathbb{R}^{n_1 \times r}$$

$$W_{\text{pert}}(\theta, \Delta) = W(\theta_0) + B_W A_W + B_{AWP} A_{AWP}$$

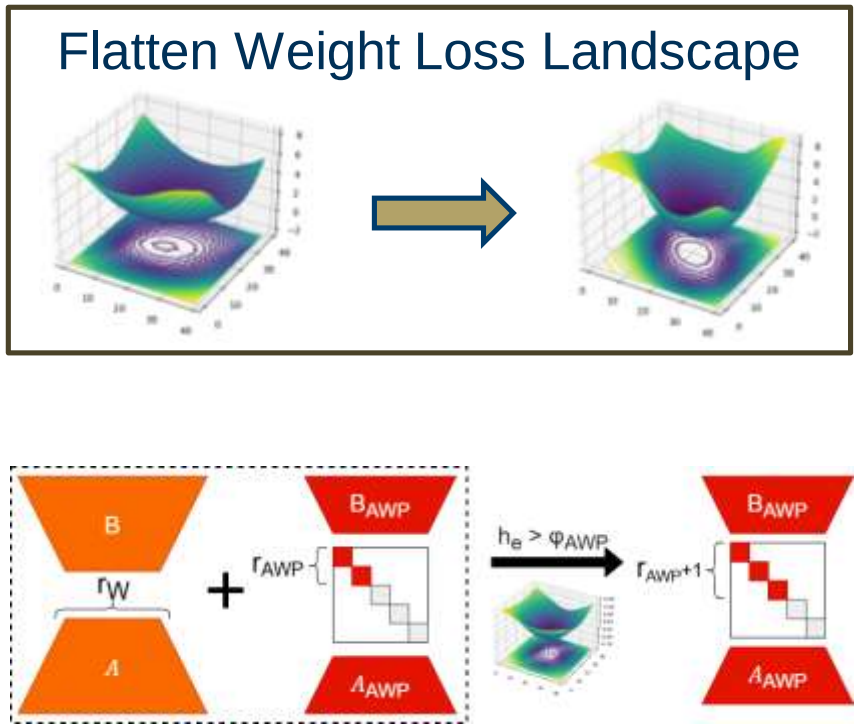
Figure 5: Schematics of the low-rank weight updates and adversarial weight perturbations.

Georgia Tech

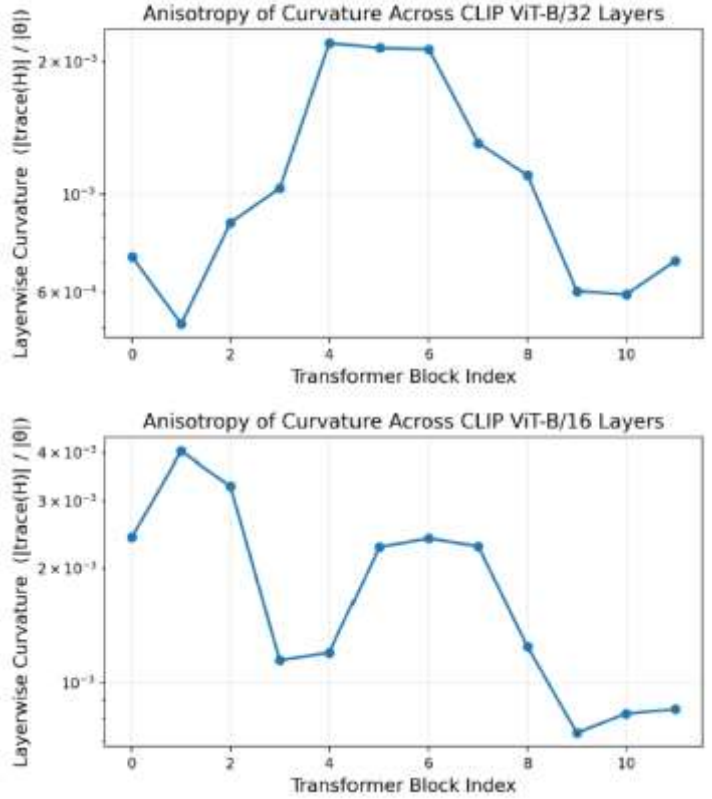
GRACE: Gram-aligned Robustness via Adaptive Curvature Estimation

Objective 2: Flatten Weight Loss Landscape

Layer-wise Adaptive Low-Rank Adversarial Weight Perturbation (LAR-AWP)



(a)



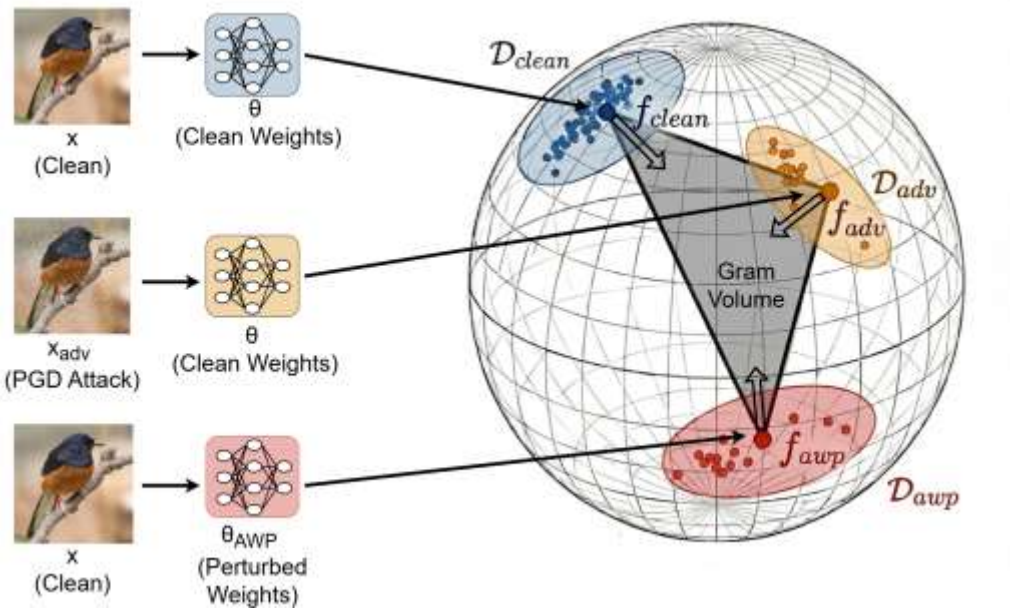
(b)

Figure 6: (a) Rank curriculum for the adversarial weight perturbations. (b) Layerwise curvature anisotropy in CLIP models.

GRACE: Gram-aligned Robustness via Adaptive Curvature Estimation

Objective 3: Reduce Feature Distortion due to Distribution Shifts

Gram-Volume Feature Alignment



$$G_i = \begin{bmatrix} \langle f_{\text{ID}}, f_{\text{ID}} \rangle & \langle f_{\text{ID}}, f_{\text{Adv}} \rangle & \langle f_{\text{ID}}, f_{\text{AWP}} \rangle \\ \langle f_{\text{Adv}}, f_{\text{ID}} \rangle & \langle f_{\text{Adv}}, f_{\text{Adv}} \rangle & \langle f_{\text{Adv}}, f_{\text{AWP}} \rangle \\ \langle f_{\text{AWP}}, f_{\text{ID}} \rangle & \langle f_{\text{AWP}}, f_{\text{Adv}} \rangle & \langle f_{\text{AWP}}, f_{\text{AWP}} \rangle \end{bmatrix} + \varepsilon I$$

$$\mathcal{L}_{\text{GV}} = \sqrt{|\det(G_i)|}.$$

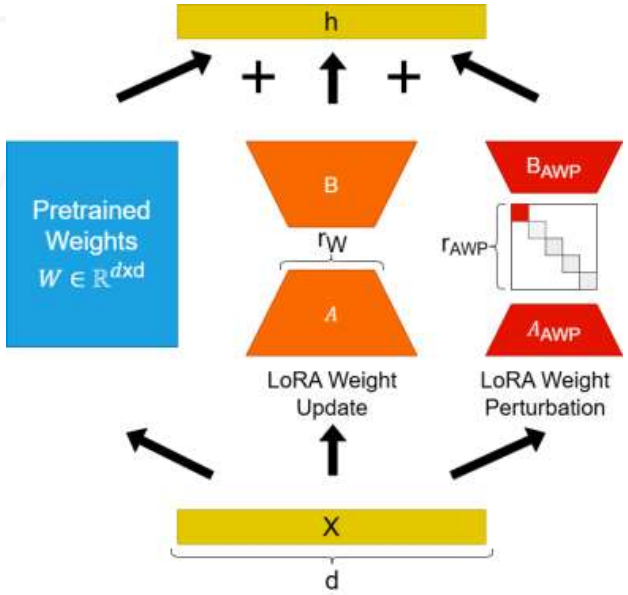
Figure 7: Gram-Volume feature alignment to encourage feature alignment across distribution shifts.

Georgia Tech

GRACE: Gram-aligned Robustness via Adaptive Curvature Estimation

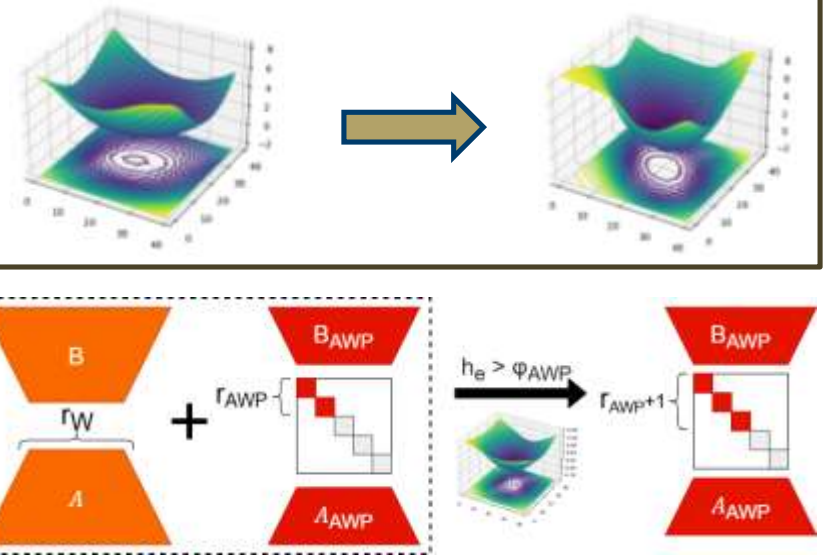
LoRA Finetuning

Maintain Proximity to Pretrained Model



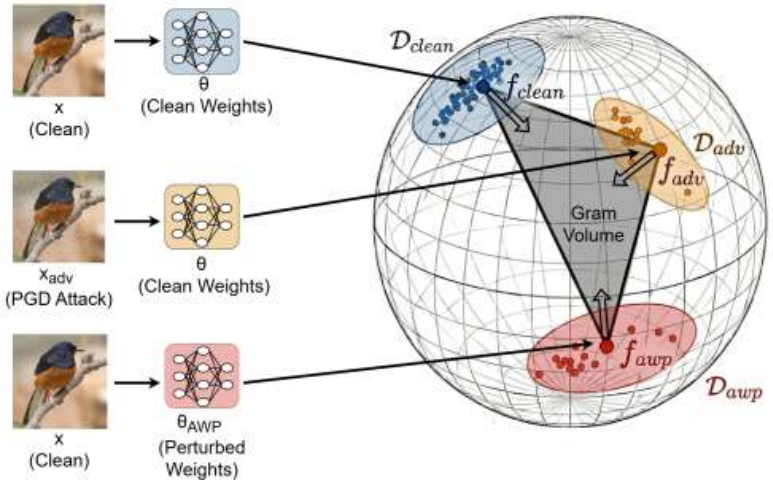
Layer-wise Adaptive Low-Rank Adversarial Weight Perturbation (LAR-AWP)

Flatten Weight Loss Landscape



Gram-Volume Feature Alignment

Reduce Feature Distortion



$$\mathcal{L}_{\text{GRACE}} = \mathcal{L}_{\text{task}} + \lambda_{\text{LAR}}\,\mathcal{L}_{\text{LAR-AWP}} + \lambda_{\text{GV}}\,\mathcal{L}_{\text{GV}}$$

Georgia Tech

Overall Training Algorithm

Algorithm 1 GRACE: Gram-aligned Robustness via Adaptive Curvature Estimation

Require: Pretrained CLIP model F_{θ_0}; training data $\mathcal{D}$; LoRA rank r; perturbation radius ρ; PGD budget c; tradeoff parameters $\lambda_{\text{LAR}}, \lambda_{\text{GV}}$.

Initialize LoRA parameters $\Theta = \{A_W, B_W\}_W$; freeze backbone weights.
Initialize $r_{\text{AWP}}^{(W)} \leftarrow 0$ and $h_W \leftarrow 0$ for all layers W.
for each training iteration **do**
  Sample minibatch $\{(x_i, y_i)\}_{i=1}^{B}$.
  Compute $f_{\text{ID}}(x_i)$ and $\mathcal{L}_{\text{task}}$. ▷ Clean forward pass
  Obtain $x_i^{\text{Adv}} \approx \arg\max_{\|\delta\| \le \epsilon} \mathcal{L}(F_\theta(x_i + \delta), y_i)$ using PGD. ▷ Input-space adversary
  Compute $f_{\text{Adv}}(x_i)$ from x_i^{Adv}.
  for $t = 1$ to T_{AWP} **do**
    $(A_{\text{AWP}}, B_{\text{AWP}}) \leftarrow (A_{\text{AWP}}, B_{\text{AWP}}) + \eta \nabla_\Delta \mathcal{L}(F_{\theta,\Delta}(x_i^{\text{Adv}}), y_i)$. ▷ Low-rank weight-space ascent
    Project Δ onto $\|\Delta\| \le \rho$ using layerwise rank masks $r_{\text{AWP}}^{(W)}$.
  end for
  Compute $f_{\text{AWP}}(x_i)$ using $W_{\text{pert}}(\theta, \Delta)$. ▷ Perturbed-weight forward pass
  Form G_i from $(f_{\text{ID}}(x_i), f_{\text{Adv}}(x_i), f_{\text{AWP}}(x_i))$.
  Compute $\mathcal{L}_{GV} = \sqrt{|\det(G_i)|}$. ▷ Feature-volume alignment
  if iteration $\bmod K = 0$ **then**
    Estimate $h_W \approx \mathbb{E}[\|\nabla_W \mathcal{L}\|^2]$ using validation mini-batch gradients.
    Update $r_{\text{AWP}}^{(W)}$ from curvature percentile bins. ▷ Sharper layers receive higher rank
  end if
  Update Θ by minimizing

$$\mathcal{L}_{\text{GRACE}} = \mathcal{L}_{\text{task}} + \lambda_{\text{LAR}} \mathcal{L}_{\text{LAR-AWP}} + \lambda_{\text{GV}} \mathcal{L}_{\text{GV}}.$$

  ▷ Outer LoRA update
end for

Georgia Tech

Datasets

In-Distribution Dataset	OOD Datasets	Natural Adversarial Datasets	Zero-Shot Datasets
ImageNet	ImageNet-v2, ImageNet-R, ImageNet-S	ImageNet-A, ImageNet-A-Plus	CalTech101, Cars, DTD, EuroSAT, FGVC, Flowers, Oxford Pets, STL-10

Table 4: Datasets used to test OOD and adversarial robustness.

Georgia Tech

Results: OOD and Natural Adversarial Datasets

Method	ID		Domain Shift						Natural Adversarial				Statistics			
	ImageNet		ImageNet-V2		ImageNet-S		ImageNet-R		ImageNet-A		ImageNet-A-Plus		OOD Avg.		Nat Adv Avg	
	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv
CLIP	63.35	0.00	56.48	0.00	41.44	0.00	68.83	0.01	32.69	0.00	37.92	0.00	55.58	0.00	35.30	0.00
Vanilla FT	74.86	0.00	63.75	0.00	43.34	0.00	60.85	0.00	22.47	0.00	29.12	0.00	55.98	0.00	25.80	0.00
WISE-FT	70.20	0.00	61.06	0.00	43.34	0.00	63.78	0.00	32.60	0.00	39.74	0.00	56.06	0.00	**36.17**	0.00
FLYP	72.15	0.00	62.22	0.00	42.22	0.00	64.38	0.01	30.06	0.00	37.89	0.00	56.27	0.00	33.98	0.00
TPGM	74.13	0.00	63.20	0.00	42.71	0.00	64.79	0.00	32.38	0.00	38.73	0.00	**56.90**	0.00	35.56	0.00
SPD	73.68	0.00	62.60	0.00	42.56	0.00	64.60	0.00	32.07	0.00	37.46	0.00	56.59	0.00	34.76	0.00
TeCoA	52.54	24.40	41.10	17.71	25.89	13.06	44.54	21.01	5.11	0.66	9.86	1.07	37.18	17.26	7.48	0.86
FARE	47.38	13.94	39.00	9.77	28.63	11.05	47.79	17.05	5.97	0.54	10.38	0.68	38.47	12.62	8.18	0.61
PMG-AFT	58.20	26.80	43.50	19.70	31.10	14.80	50.30	22.10	8.22	1.12	13.07	1.94	41.63	18.87	10.64	1.53
LAAT	55.46	21.54	42.30	15.80	30.15	12.20	48.90	20.40	12.23	1.93	16.48	3.48	40.45	16.13	14.35	2.70
GRACE	74.21	25.44	62.60	20.51	42.11	17.75	58.53	23.39	20.02	3.32	25.33	4.86	54.41	**20.55**	22.67	**4.09**

Table 5: OOD and Natural Adversarial evaluation on ImageNet and its variants. (ViT-B/32)

Georgia Tech

Results: Zero-Shot Datasets

	Zero shot Datasets																Statistics	
	CalTech101		Cars		DTD		EuroSAT		FGVC		Flowers		Oxford Pets		STL-10		ZS Avg.	
Method	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv	Clean	Adv
CLIP	80.47	0.00	51.47	0.00	41.43	0.00	45.27	0.00	15.93	0.00	60.18	0.00	84.00	0.00	95.71	0.00	59.31	0.00
Vanilla FT	79.85	0.00	45.44	0.00	40.26	0.00	41.18	0.00	14.16	0.00	55.42	0.00	84.30	0.00	97.03	0.00	57.21	0.00
WISE-FT	81.47	0.00	53.18	0.00	42.07	0.00	46.57	0.00	16.08	0.00	60.66	0.00	85.90	0.00	97.26	0.00	**60.40**	0.00
FLYP	80.77	0.00	33.72	0.00	38.56	0.00	39.72	0.00	11.76	0.00	50.18	0.00	84.21	0.00	96.00	0.00	54.37	0.00
TPGM	80.49	0.00	47.69	0.00	40.85	0.00	43.37	0.00	13.86	0.00	57.18	0.00	85.14	0.00	97.25	0.00	58.23	0.00
SPD	79.80	0.00	46.31	0.00	40.53	0.00	42.51	0.00	15.03	0.00	57.70	0.00	84.84	0.00	97.38	0.00	58.01	0.00
TeCoA	74.03	51.40	10.68	1.50	24.04	10.70	18.72	11.00	4.56	0.50	25.59	6.60	67.07	29.10	85.30	55.40	38.75	20.78
FARE	76.97	45.80	33.87	1.30	27.87	12.20	14.77	11.00	8.73	0.60	31.72	5.30	75.44	20.40	87.87	55.50	44.66	19.01
PMG-AFT	77.10	53.20	29.50	2.00	30.20	12.90	22.30	11.50	7.90	0.80	33.40	7.30	72.10	31.50	88.90	56.80	45.18	**22.00**
LAAT	75.10	47.20	28.80	1.60	29.90	11.80	20.60	10.90	6.10	0.70	28.90	6.10	72.20	22.40	86.10	55.90	43.46	19.58
GRACE	79.23	49.30	52.10	1.90	41.60	12.60	46.10	11.20	15.50	0.70	60.05	7.00	85.20	30.10	97.10	56.00	59.61	21.10

Table 6: Clean and Adversarial evaluation on zero-shot datasets. (ViT-B/32)

Georgia Tech

Results: Unified Summary

Method	ID	Average OOD	Average Adversarial	Harmonic Mean
CLIP	63.35	57.44	8.82	20.46
Vanilla FT	74.86	56.59	8.95	21.01
WISE-FT	70.20	58.05	9.04	21.11
FLYP	72.15	55.32	8.49	20.03
TPGM	74.13	57.56	8.89	20.92
SPD	73.68	57.53	8.69	20.54
TeCoA	52.54	37.96	17.48	29.24
FARE	47.38	41.56	13.43	25.07
PMG-AFT	58.20	43.40	19.57	<u>32.85</u>
LAAT	55.46	41.95	17.90	30.69
GRACE (Ours)	74.21	57.01	22.44	**39.69**

Table 7: Unified results across zero-shot and fine-tuned settings (ViT-B/32)

Georgia Tech

Ablation Study and Computation Analysis

Variant	ID	Average OOD	Average Adversarial	Harmonic Mean
CLIP	63.3	57.4	8.8	20.4
LoRA-FT (no regs)	72.8	55.0	8.2	19.4
+ GV only	72.0	56.5	8.6	20.2
+ LAR-AWP (no curriculum)	71.0	53.0	17.2	32.9
+ LAR-AWP (with curriculum)	72.5	54.0	22.2	38.7
GRACE	74.2	57.0	22.4	39.6

Table 8: Ablation study testing the impact of various components on the overall GRACE framework. (ViT-B/32)

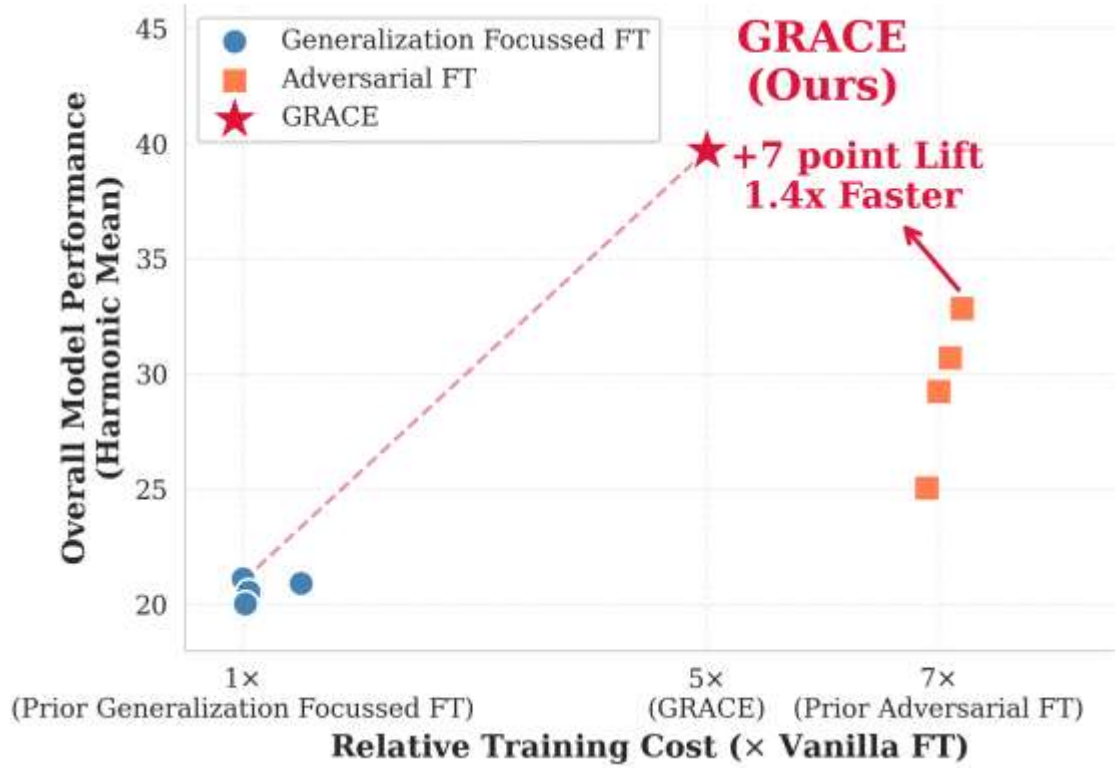


Figure 8: Pareto curve for GRACE. (ViT-B/32)

Georgia Tech

Conclusion

- Study the three-way tradeoff between:
  ID Accuracy v/s OOD Accuracy v/s Adversarial Robustness
- Theoretically analyze the reason behind this trade-off.
- Propose GRACE to effectively resolve the trade-off.

Georgia Tech