PATH-CONSTRAINED MODEL MERGING FOR ROBUST VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) deployed in real-world settings need to handle both natural distribution shifts and adversarial perturbations. Model merging provides an efficient way to combine these two capabilities without additional training. However, we find that existing merging methods, while achieving clear OOD gains, do not reliably preserve the robustness already present in an adversarially robust model. We further observe that merging affects perturbation sensitivity differently across modules, and that correcting only a small number of sensitive parts can recover a large portion of the lost robustness. In this paper, we propose **Pa**th-**Co**nstrained Model Merging (PACOMERGE), a training-free and data-free post-merging approach. PACOMERGE operates along the merge path produced by an existing model merging method and uses the perturbation sensitivity of the adversarially robust model to define a constrained region along this path. Updates that remain within the constraint are kept unchanged, while those that exceed it receive only the necessary correction. We further show that the points satisfying the constraint form a continuous interval along the merge path, allowing PACOMERGE to retain as much of the original merge as possible while making only the minimum correction needed to recover robustness. Across five base mergers, PACOMERGE raises average robust accuracy to 18.24–27.78% on ImageNet and four distribution shifts, with consistent gains on zero-shot and LLaVA tasks while preserving multi-task merging performance.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly deployed in safety critical applications such as autonomous driving (Zhou et al., 2024), robotics (Ma et al., 2026), and medical image analysis (Qin et al., 2023). In real-world deployment, models often face more challenging test conditions, where changes in environments and data distributions lead to natural distribution shifts (Zhang et al., 2021), while subtle adversarial perturbations can also cause incorrect predictions (Schlarmann et al., 2024).

Existing approaches typically address these two challenges separately. Adversarial training (Mao et al., 2023; Schlarmann et al., 2024) can substantially improve robustness against worst-case perturbations, often at the cost of performance under natural distribution shifts (Zou & Liu, 2023). By contrast, domain generalization methods (Zhou et al., 2022; Addepalli et al., 2024) enhance generalization across diverse domains, without explicitly considering adversarial robustness.

Recently, model merging (Wortsman et al., 2022a; Yang et al., 2024) has become a more efficient alternative, directly combining checkpoints fine-tuned from the same pretrained model and integrating their capabilities without additional training. This naturally raises the question: *given an adversarially robust checkpoint and another OOD generalization checkpoint, can existing model merging retain adversarial robustness while improving OOD generalization?*

Baseline mergers (Matena & Raffel, 2022; Ilharco et al., 2023; Yadav et al., 2023; Marczak et al., 2025) are mainly designed to combine capabilities learned from different downstream tasks. Our setting is different: the two source models solve the same task and share the same label space, but specialize in different forms of robustness, namely OOD generalization and adversarial robustness. Since both models are fine-tuned from the same pretrained model, directly merging them may appear straightforward. However, Fig. 1 shows that mergers with similar OOD gains can retain very

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

different levels of adversarial robustness, with some losing most of the robustness inherited from the robust model. This suggests that improving OOD generalization and retaining adversarial robustness are not automatically compatible under model merging. A key question is where this robustness degradation arises and whether it can be corrected without discarding useful merge updates.

In this work, we propose **Pa**th-**Co**nstrained Model Merging (PACOMERGE), a training-free and data-free post-merging approach. Rather than designing a new merging rule, PACOMERGE treats the output of an existing merger as a proposal and adjusts it along the original update direction from the adversarially robust model to the merged model. For each module, PACOMERGEidentifies the largest feasible step under a relative perturbation amplification constraint defined with respect to the robust model. Since the weight matrices within attention and MLP modules are functionally coupled, we assign a shared step to each module using the most conservative feasible value among its matrices. This preserves as much of the original merge update as possible while selectively correcting updates that introduce excessive perturbation amplification. We further show that the feasible points under this relative perturbation amplification constraint form a continuous interval along the merge path. This means that the selected step is the largest feasible update for each module and, equivalently, the minimum correction to the merged model along its original update direction.

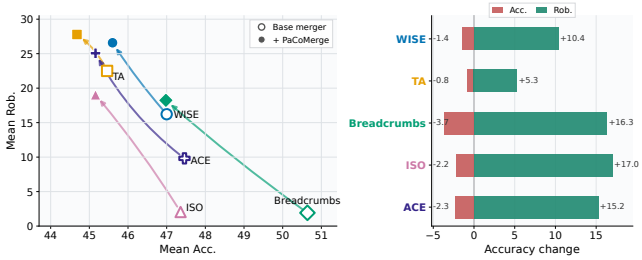


Figure 1: Across five base mergers, PaCoMerge improves adversarial robustness with modest changes in clean accuracy. Results use ViT-B/32 and report mean accuracy over ImageNet and four distribution shifts. Robust accuracy is evaluated with AutoAttack at $\epsilon = 1/255$. The right panel shows absolute changes from each base merger.

We evaluate PACOMERGE across five base mergers, multiple VLM backbones, distribution shifts, attacks, zero-shot benchmarks, and downstream VLM tasks. Across the five base mergers, PACOMERGE raises average robust accuracy to 18.24–27.78% on ImageNet and four distribution shifts. The gains further transfer to zero-shot and LLaVA tasks, while standard multi-task merging performance remains largely unchanged.

Contributions. Our core contributions are summarized as follows:

1. We study merging models with complementary OOD generalization and adversarial robustness, and identify substantial adversarial robustness degradation in existing mergers.
2. We propose PACOMERGE, a training- and data-free correction that selectively limits excessive perturbation amplification.
3. We show that the relative amplification constraint yields a continuous feasible interval, making the selected step the largest feasible update with minimum correction.
4. Across five base mergers and diverse VLM evaluations, PACOMERGE substantially improves adversarial robustness while largely preserving OOD and standard multi-task merging performance.

2 RELATED WORK

OOD and Adversarial Robustness. Vision models often suffer clear performance drops under both natural distribution shifts and adversarial perturbations. Previous studies have also shown that these two types of robustness are closely related (Li et al., 2024b; Yi et al., 2021; Zhu et al., 2023). For VLMs, FARE (Schlarmann et al., 2024), PMG-AFT (Wang et al., 2024), and APT (Li et al., 2024a) improve zero-shot robustness through adversarial fine-tuning or prompt tuning. More recent works improve CLIP robustness through test-time adaptation (Tong et al., 2025; Xing et al., 2025). CARE (Thai et al., 2026) further combines multiple robust experts through joint training. These approaches improve or jointly train robust models, whereas our goal is to preserve adversarial robustness when combining an already robust model with an OOD generalized model without additional training.

Spectral Norm and Perturbation Amplification. Spectral norms and Lipschitz continuity help study how neural networks respond to input perturbations. A larger spectral amplification can cause small input perturbations to grow as they pass through the network (Bartlett et al., 2017; Farnia et al., 2019). In model merging, SVC (Li et al., 2026) studies the over-accumulation of shared knowledge in multi-task model merging and attributes it to singular-value inflation along shared spectral directions. Liu et al. (Liu et al., 2026) analyze task interference in multi-task model merging and continual learning from the perspective of spectral norms. These works focus on cross-task knowledge composition. In contrast, we study the adversarially robust region along the merge path.

Data-free Model Merging. Most model merging methods focus on reducing conflicts between model updates. Model Soups (Wortsman et al., 2022a) uses simple parameter averaging. Task Arithmetic (Ilharco et al., 2023) and TIES-Merging (Yadav et al., 2023) combine models through task vectors and parameter conflict resolution. Iso (Marczak et al., 2025) further uses the subspace structure of model weights to improve merging. Seasoning Model Soups (Croce et al., 2023) interpolates models with different robustness properties to trade off robustness objectives, while PISTOLE (Dong et al., 2026) uses prediction stability and gradient information for parameter selection. In contrast, we study robustness retention through a correction that, for fixed hyperparameters, uses only model weights to analyze and correct adversarial robustness degradation caused by merging.

3 Motivation: Localized Robustness Degradation

We next examine where the robustness degradation observed in Fig. 1 arises and whether it requires modifying the entire merged model.

Perturbation amplification is highly localized. We compare the perturbation amplification of attention and MLP modules in the merged model relative to the adversarially robust source. As shown in Fig. 2, the increase in perturbation sensitivity is highly non-uniform across the network. Most Transformer modules remain close to the robust source, while only a small subset exhibits substantially larger amplification. Moreover, the affected modules differ across base mergers. This suggests that robustness degradation is not caused by a uniform shift of the entire model, and therefore cannot be effectively addressed by simply shrinking the whole merge update.

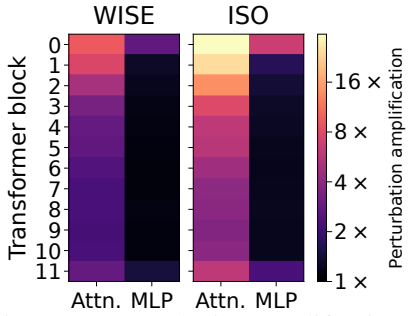


Figure 2: Perturbation amplification is concentrated in a few Transformer groups.

Selective correction is sufficient. The localized degradation suggests that recovering robustness may not require modifying the entire merged model. Fig. 3 confirms this intuition: correcting groups with large perturbation amplification recovers robustness with substantially fewer corrected groups. For WISE, correcting only four sensitive groups already recovers 7.3 points of robustness.

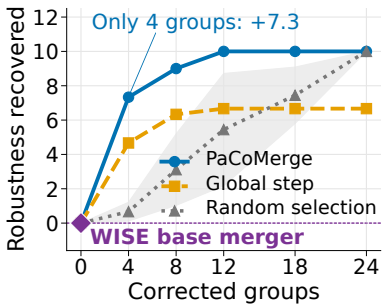
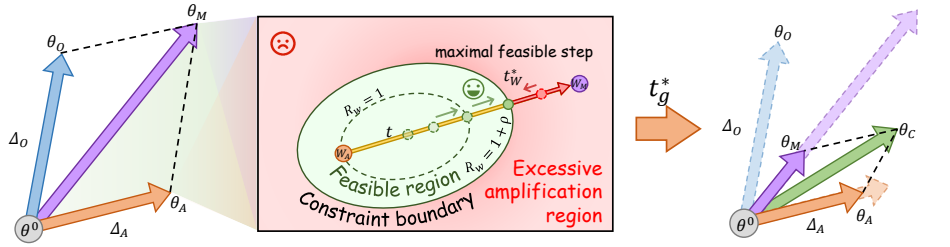


Figure 3: Selective correction rapidly recovers robustness by targeting high-amplification groups.

Together, these observations suggest that robustness degradation during merging is driven disproportionately by a small number of sensitive updates. This motivates a selective correction strategy: retain the original merge update wherever possible, while restricting only the modules that introduce excessive perturbation amplification.

4 Method

PaCoMerge is a data-free post-merging method designed to reduce adversarial robustness degradation while preserving the benefits of the base merger $\mathcal{M}(\theta_A, \theta_O)$. For fixed ρ and η, the correction



(a) Base merging (b) Constrained merge path (c) Selective correction

Figure 4: Overview of our method. (a) A base merger combines models with complementary robustness. (b) We identify the feasible region along the merge path and obtain the group-wise step $t_g^\star$. (c) Sensitive groups are selectively corrected while the remaining merge updates are retained.

uses only model weights. Rather than modifying the update direction produced by the base merger, PACOMERGE only controls how far each module moves along it. We first define the merge path in Sec. 4.2, introduce a relative perturbation amplification constraint in Sec. 4.3, and then derive a module-wise selective projection in Sec. 4.4. An overview of our proposed framework is shown in Fig. 4.

4.1 PRELIMINARY

Problem Setup. In this work, we focus on the CLIP (Radford et al., 2021) visual encoder, which is also widely used as the vision backbone in LVLMs such as LLaVA (Liu et al., 2023). Recent work has studied the robustness of CLIP-like visual encoders under distribution shifts and adversarial perturbations (Mao et al., 2023; Schlarmann et al., 2024; Chi et al., 2025). Let ϕ_θ denote the visual encoder and $\{\psi_k\}_{k=1}^K$ denote fixed text embeddings for K classes. For an input x, the normalized class logits are $s_k^\theta(x) = \langle \bar{\phi}_\theta(x), \bar{\psi}_k \rangle$, and the prediction is given by $h_\theta(x) = \arg\max_k s_k^\theta(x)$. We consider a source distribution $\mathcal{D}_S$ and a shifted target distribution $\mathcal{D}_T$. Let θ_A denote an adversarially robust model and θ_O a model with stronger OOD generalization. Specifically, θ_A is trained with adversarial perturbations bounded by ϵ, whereas θ_O exhibits stronger clean performance under distribution shift. A base merging method $\mathcal{M}(\cdot)$ produces a merged model $\theta_M = \mathcal{M}(\theta_A, \theta_O)$ to combine adversarial robustness and OOD generalization.

Spectral Norm and Adversarial Sensitivity. Prior work has connected spectral and Lipschitz properties to adversarial robustness (Cissé et al., 2017; Tsuzuku et al., 2018; Huang et al., 2021). For a linear transformation $f(x) = Wx$, its Lipschitz constant under the ℓ_2 norm is given by the spectral norm of W:

$$\|W\|_2 = \sup_{\boldsymbol{u} \neq \mathbf{0}} \frac{\|W\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} = \sigma_{\max}(W), \tag{1}$$

where $\sigma_{\max}(W)$ denotes the largest singular value. For an input perturbation $\boldsymbol{\delta}$,

$$\|W(x+\boldsymbol{\delta}) - Wx\|_2 = \|W\boldsymbol{\delta}\|_2 \leq \sigma_{\max}(W)\|\boldsymbol{\delta}\|_2. \tag{2}$$

Thus, the spectral norm measures the maximum sensitivity of a linear transformation to input perturbations.

4.2 MERGE PATH

For any weight matrix W, let W_A and W_M denote its parameters in θ_A and θ_M, respectively, and define the merge update relative to the robust model as $\Delta W = W_M - W_A$. We then parameterize the corresponding merge path as

$$W(t) = W_A + t\Delta W, \qquad t \in [0, 1]. \tag{3}$$

Here, $t = 0$ corresponds to the robust source W_A, while $t = 1$ recovers the merged weight W_M. This leads to a direct question: *Given the update direction produced by the base merger, how far can we move along it without introducing excessive perturbation amplification?* We address this question using a matrix-level relative perturbation amplification measure.

4.3 Relative Perturbation Amplification Constraint

Since PACOMERGE is data-free, no input-dependent sensitivity signal is available during merging. Moreover, parameter distance alone cannot characterize perturbation amplification, since the same parameter change may affect different input directions differently. Based on this observation, we compare how W_A and $W(t)$ amplify the same non-zero direction $\boldsymbol{u}$. Here, $\boldsymbol{u}$ denotes an arbitrary direction in the input representation space of W, and can also be interpreted as a local perturbation entering this linear transformation. Compared with $W_A\boldsymbol{u}$, we define the regularized squared relative perturbation amplification as

$$R_W(t) = \sup_{\boldsymbol{u}\neq\mathbf{0}} \frac{\|W(t)\boldsymbol{u}\|_2^2 + \eta\|\boldsymbol{u}\|_2^2}{\|W_A\boldsymbol{u}\|_2^2 + \eta\|\boldsymbol{u}\|_2^2}, \tag{4}$$

where $\eta > 0$ stabilizes the ratio when $W_A\boldsymbol{u}$ is small and prevents near-null directions of W_A from dominating. At $t = 0$, we have $W(t) = W_A$ and thus $R_W(0) = 1$. As we move along the merge path, some directions may be excessively amplified, weakening the local stability inherited from W_A. To limit this additional amplification, we set the following constraint:

$$R_W(t) \leq 1 + \rho, \tag{5}$$

where $\rho \geq 0$ controls the allowed relative perturbation amplification along the path. All t satisfying Eq. (5) form the feasible region, providing a data-free, matrix-level criterion for limiting excessive perturbation amplification relative to the robust model. It constrains local perturbation propagation, but does not directly guarantee end-to-end adversarial robustness for the nonlinear Transformer.

For computational convenience, we rewrite Eq. (4) by defining

$$C_A = W_A^\top W_A + \eta I, \qquad C_t = W(t)^\top W(t) + \eta I, \qquad R_W(t) = \lambda_{\max}\left(C_A^{-1/2} C_t C_A^{-1/2}\right), \tag{6}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. The detailed derivation is provided in Appendix C.1. We next determine how far each module can move along the merge path under this constraint.

4.4 Selective Projection

Under the constraint in Eq. (5), for each weight matrix W, we first find the maximum feasible step:

$$t_W^\star = \max\{t \in [0,1] : R_W(t) \leq 1 + \rho\}. \tag{7}$$

Here, $t_W^\star$ denotes the maximum step allowed along the update direction defined in Eq. (3). A trivial choice is to compute an independent $t_W^\star$ in Eq. (7) for every weight matrix W. However, an attention module is nonlinear, where Q, K, and V are coupled through $\mathrm{Softmax}(QK^\top)V$. Similar coupling also arises in MLP modules through the two linear transformations and the intermediate nonlinear activation. As a result, even if the Q, K, and V matrices individually satisfy the matrix-level amplification constraint, adjusting them independently may disrupt their coupled behavior. Another simple choice is to use a single global step given by the smallest matrix-level step: $t_{\text{global}}^\star = \min_W t_W^\star$. However, this choice can be overly conservative, as a single sensitive matrix may restrict the update of the entire model. We therefore share one interpolation step within each attention or MLP module. For the g-th module with weight matrices $\mathcal{W}_g$, we choose the largest step that is feasible for all matrices:

$$\mathbf{t}_\mathbf{g}^\star = \max\{\mathbf{t} \in [\mathbf{0},\mathbf{1}] : \mathbf{R}_\mathbf{W}(\mathbf{t}) \leq \mathbf{1} + \rho,\ \forall \mathbf{W} \in \mathcal{W}_\mathbf{g}\}. \tag{8}$$

As shown in Sec. 4.5, the feasible set for each matrix is an interval, so the module-wise step simplifies to $t_g^\star = \min_{W\in\mathcal{W}_g} t_W^\star$. For each $W \in \mathcal{W}_g$, let W_A and W_M denote its weights in the robust and merged models. Substituting $t_g^\star$ into Eq. (3) gives

$$\widehat{W} = W_A + t_g^\star(W_M - W_A), \qquad W \in \mathcal{W}_g. \tag{9}$$

Thus, each matrix remains on its original merge path from W_A to W_M. Other parameters, including biases, normalization parameters, and embeddings, remain unchanged from θ_M.

4.5 Analysis

We prove two properties of PACOMERGE. First, the relative amplification constraint directly bounds the additional propagation of perturbations through each constrained linear transformation. Second, the feasible region forms an interval along the merge path, making the module-wise step the exact minimum-correction solution under the shared-path constraint.

Perturbation Propagation. The relative amplification constraint also admits a direct perturbation-propagation interpretation. For any perturbation $\boldsymbol{\delta}$, $R_W(t) \leq 1+\rho$ implies

$$\|W(t)\boldsymbol{\delta}\|_2^2 \leq (1+\rho)\|W_A\boldsymbol{\delta}\|_2^2 + \rho\eta\|\boldsymbol{\delta}\|_2^2. \tag{10}$$

Thus, the additional amplification of any local perturbation is explicitly bounded relative to the robust model.

Pathwise Constrained Projection. For a weight matrix W, let $\mathcal{T}_W(\rho) = \{t \in [0,1] : R_W(t) \leq 1+\rho\}$ denote its feasible steps along the merge path.

Theorem 1 (Pathwise Constrained Projection)**.** *For every weight matrix W, $\mathcal{T}_W(\rho) = [0, t_W^\star]$. Consequently, for a module g with weight matrices $\mathcal{W}_g$,*

$$\mathcal{T}_g(\rho) = \bigcap_{W\in\mathcal{W}_g} \mathcal{T}_W(\rho) = [0, t_g^\star], \qquad t_g^\star = \min_{W\in\mathcal{W}_g} t_W^\star. \tag{11}$$

Moreover, $t_g^\star$ solves

$$\min_{t\in[0,1]} \sum_{W\in\mathcal{W}_g} \|W_A + t(W_M - W_A) - W_M\|_F^2 \quad \text{s.t.} \quad R_W(t) \leq 1+\rho,\ \forall W \in \mathcal{W}_g. \tag{12}$$

The complete proof is provided in Appendix C.3.

Theorem 1 shows that $t_g^\star$ retains the largest jointly feasible portion of the merge update while making the minimum correction along the original merge path.

5 Experiments

We evaluate PACOMERGE from three perspectives: its effectiveness across different base mergers, its generalization across datasets and downstream tasks, and the role of selective projection in recovering robustness. We further evaluate robustness under different attacks and perturbation budgets.

5.1 Experiment Setup

Source Models. We consider two CLIP vision encoders derived from the same pretrained model. We use the zero-shot OpenAI CLIP checkpoint as the generalization source θ_O, and its adversarially fine-tuned FARE (Schlarmann et al., 2024) checkpoint as the robustness source θ_A. FARE is trained on ImageNet-1K (Deng et al., 2009) to improve adversarial robustness while preserving the pretrained CLIP representation.

Architectures. We evaluate PACOMERGE on CLIP ViT-B/32 and ViT-L/14 (Dosovitskiy et al., 2021), applying the correction to each base-merger checkpoint without additional training.

Evaluation Benchmarks. We evaluate PACOMERGE in three settings. (1) *OOD robustness under attack*: ImageNet (Deng et al., 2009) and four distribution shifts, including ImageNet-V2 (Recht et al., 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019). (2) *Zero-shot robustness*: 11 classification benchmarks covering diverse visual domains, including Caltech101 (Fei-Fei et al., 2004), StanfordCars (Krause et al., 2013), CIFAR-10/100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), FGVC-Aircraft (Maji et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), PCAM (Veeling et al., 2018), Oxford-IIIT Pets (Parkhi et al., 2012), and STL-10. (3) *VLM integration*: LLaVA-1.5-7B (Liu et al., 2023) with the merged OpenCLIP ViT-L/14 vision encoder on COCO (Lin et al., 2014), Flickr30k, VQAv2 (Goyal et al., 2017), and TextVQA. Adversarial robustness is evaluated under white-box attacks at multiple perturbation budgets.

Baselines. We apply PACOMERGE to five representative base merging methods: WiSE-FT (Wortsman et al., 2022b), Task Arithmetic (TA) (Ilharco et al., 2023), Model Breadcrumbs (Davari & Belilovsky, 2024), Isotropic Merging (ISO) (Marczak et al., 2025), and ACE (Xu et al., 2026). All methods use the same pair of source checkpoints. For classification, we additionally report the two source models, OpenAI CLIP and FARE, as reference points.

Detailed implementation and evaluation settings are provided in Appendix A.

Table 1: Clean and robust accuracy (%) of ViT-B/32 on ImageNet and four distribution shifts under AutoAttack ($\epsilon = 1/255$). Avg. averages all clean and robust results.

Method	ImageNet		ImageNet-V2		ImageNet-R		ImageNet-S		ImageNet-A		Avg.
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	
Source models											
OpenAI	**63.5**	0.0	**55.2**	0.0	**67.7**	0.2	**40.8**	0.1	**31.6**	0.0	25.91
FARE	51.7	**39.7**	44.2	**31.1**	49.9	**40.6**	31.0	**23.9**	6.4	**3.6**	**32.21**
Base mergers and PaCoMerge outputs											
WISE	**60.8**	22.5	**53.5**	17.7	**63.1**	25.1	39.4	14.2	**18.2**	1.6	31.61
+ PaCoMerge	59.9	**35.3**	51.1	**28.9**	62.9	**40.5**	**40.1**	**24.2**	14.0	**4.0**	**36.09**
TA	**59.9**	31.2	**51.3**	26.1	**62.1**	33.7	38.3	18.3	**15.7**	3.2	33.98
+ PaCoMerge	58.9	**37.5**	50.6	**30.4**	61.6	**42.2**	**38.8**	**24.9**	13.5	**3.9**	**36.23**
Breadcrumbs	**62.5**	1.4	**54.0**	0.8	**67.7**	5.1	**42.1**	2.3	**26.9**	0.0	26.28
+ PaCoMerge	60.0	**22.6**	51.1	**17.8**	65.0	**30.2**	41.3	**18.7**	17.5	**1.9**	**32.61**
ISO	**60.1**	0.7	**50.4**	0.3	**64.8**	4.9	38.0	4.3	**23.5**	0.0	24.70
+ PaCoMerge	57.7	**23.4**	49.0	**17.5**	62.4	**31.7**	**41.7**	**20.5**	15.0	**2.1**	**32.10**
ACE	**62.0**	14.5	**52.3**	10.7	**63.3**	15.9	39.8	7.1	**19.9**	0.9	28.64
+ PaCoMerge	58.1	**33.2**	50.3	**26.7**	62.8	**38.3**	**40.3**	**23.4**	14.3	**3.7**	**35.11**

Table 2: Vision-language evaluation of LLaVA-1.5-7B with merged OpenCLIP ViT-L/14 vision encoders. We report clean and adversarial performance at $\epsilon = 2/255$ and $4/255$. Captioning is measured by CIDEr and VQA by accuracy (%). Arrows show changes after applying PACOMERGE; bold marks the better result within each pair.

Method	COCO (CIDEr)			Flickr30k (CIDEr)			VQAv2 (Acc.)			TextVQA (Acc.)		
	Clean	2/255	4/255	Clean	2/255	4/255	Clean	2/255	4/255	Clean	2/255	4/255
WISE	**118.59**	64.85	46.81	**79.02**	41.92	31.02	**70.08**	37.40	31.84	**33.52**	14.76	11.74
+PaCoMerge	118.27 ↓ 0.32	**76.14** ↑ 11.29	**59.41** ↑ 12.60	76.53 ↓ 2.49	**54.47** ↑ 12.55	**42.99** ↑ 11.97	69.68 ↓ 0.40	**41.54** ↑ 4.14	**35.22** ↑ 3.38	32.96 ↓ 0.56	**16.68** ↑ 1.92	**13.18** ↑ 1.44
TA	**117.73**	69.77	56.03	**78.61**	48.21	37.82	**69.10**	39.46	34.96	**32.94**	15.88	12.32
+PaCoMerge	117.41 ↓ 0.32	**79.40** ↑ 9.63	**64.17** ↑ 8.14	75.42 ↓ 3.19	**55.79** ↑ 7.58	**46.48** ↑ 8.66	69.00 ↓ 0.10	**43.14** ↑ 3.68	**35.06** ↑ 0.10	31.24 ↓ 1.70	**16.84** ↑ 0.96	**12.50** ↑ 0.18
Breadcrumbs	**120.59**	37.75	23.34	**81.93**	24.29	15.00	**71.32**	29.96	25.96	**35.52**	11.78	8.10
+PaCoMerge	119.78 ↓ 0.81	**70.40** ↑ 32.65	**46.99** ↑ 23.65	78.77 ↓ 3.16	**50.35** ↑ 26.06	**31.83** ↑ 16.83	70.42 ↓ 0.90	**37.50** ↑ 7.54	**31.72** ↑ 5.76	33.72 ↓ 1.80	**16.50** ↑ 4.72	**12.00** ↑ 3.90
ISO	118.42	35.74	20.67	**80.42**	23.70	13.27	**70.12**	29.58	25.82	**35.66**	11.18	9.24
+PaCoMerge	**118.87** ↑ 0.45	**70.32** ↑ 34.58	**50.11** ↑ 29.44	77.71 ↓ 2.71	**49.67** ↑ 25.97	**34.92** ↑ 21.65	69.66 ↓ 0.46	**38.82** ↑ 9.24	**34.04** ↑ 8.22	32.90 ↓ 2.76	**15.92** ↑ 4.74	**11.58** ↑ 2.34
ACE	**119.37**	61.35	38.26	**81.48**	40.92	26.57	**70.78**	35.18	30.86	**33.62**	13.88	10.88
+PaCoMerge	118.34 ↓ 1.03	**77.29** ↑ 15.94	**60.34** ↑ 22.08	76.73 ↓ 4.75	**54.31** ↑ 13.39	**43.03** ↑ 16.46	69.76 ↓ 1.02	**41.26** ↑ 6.08	**36.04** ↑ 5.18	32.78 ↓ 0.84	**16.78** ↑ 2.90	**12.76** ↑ 1.88

5.2 MAIN RESULTS

Robustness under Distribution Shift. Table 1 shows that existing mergers differ substantially in how well they retain the adversarial robustness of FARE. While they generally recover clean performance, robustness retention varies considerably: TA remains relatively robust, whereas Breadcrumbs and ISO lose most of the robustness inherited from FARE. PACOMERGE recovers larger gains for mergers suffering more severe robustness degradation, consistent with its selective projection design that restricts only updates causing excessive perturbation amplification while preserving the remaining merge updates. Overall, PACOMERGE improves adversarial robustness across all five base mergers while largely preserving clean performance, yielding an overall average gain of 2.25–7.40 points. The gains on ImageNet-A are smaller, where FARE reaches only 3.6% robust accuracy, leaving limited source robustness for PACOMERGE to recover.

Vision-Language Robustness. Table 2 evaluates the merged encoders in LLaVA-1.5-7B. PACOMERGE consistently improves adversarial performance across all five base mergers, four downstream tasks, and both perturbation budgets, while largely preserving clean performance. For example, at $\epsilon = 4/255$, ISO improves from 20.67 to 50.11 CIDEr on COCO and from 13.27 to 34.92 on Flickr30k. This demonstrates that the robustness recovery transfers to downstream vision-language tasks without additional training.

Table 3: Clean/robust accuracy (%) on 11 zero-shot classification benchmarks with ViT-B/32. Each entry reports clean/robust accuracy under AutoAttack at $\epsilon = 2/255$.

Method	Cal101	Cars	C10	C100	DTD	EuroSAT	Aircraft	Flowers	PCAM	Pets	STL10	Avg.
Source models												
OpenAI	82.0/0.4	61.0/0.0	89.6/1.3	61.1/1.1	43.7/0.0	49.6/0.0	21.3/0.0	63.6/0.0	52.9/0.0	85.1/0.0	97.4/0.8	32.3
FARE	79.8/73.7	37.5/23.6	67.6/56.6	46.9/37.2	28.8/23.1	17.8/14.4	11.5/6.1	33.4/25.0	61.8/56.5	77.1/66.3	89.7/84.5	46.3
Base mergers and PaCoMerge outputs												
WISE	82.6/59.3	55.8/9.6	86.5/59.0	62.5/32.1	38.1/14.3	29.3/10.1	17.1/0.7	52.7/15.9	53.8/45.3	82.8/42.8	95.7/80.6	46.7
+PaCoMerge	82.0/69.8	53.0/21.7	82.8/64.6	59.9/39.1	37.2/23.1	23.3/13.9	16.0/4.4	50.0/28.7	55.2/49.9	82.1/61.6	95.3/86.7	**50.0**
TA	81.8/67.8	52.8/16.6	82.9/63.1	59.7/37.3	36.9/20.6	22.8/12.6	15.8/2.3	50.6/24.9	56.2/48.3	82.2/56.6	95.3/85.2	48.7
+PaCoMerge	81.2/70.9	51.1/23.1	81.4/65.9	58.8/39.8	36.0/24.2	21.5/14.2	15.6/4.9	48.4/29.2	55.9/50.9	81.6/64.2	94.4/87.3	**50.0**
Breadcrumbs	83.0/16.8	57.9/0.1	90.1/21.6	64.7/10.9	40.4/1.7	37.2/0.0	18.7/0.0	58.5/1.0	52.6/3.0	83.6/1.6	97.1/40.5	35.5
+PaCoMerge	82.4/62.3	54.6/12.3	84.4/56.2	58.2/30.6	37.8/16.6	34.2/5.1	17.8/2.0	52.5/19.9	53.4/47.2	82.7/46.0	95.6/80.6	**46.9**
ISO	82.7/12.7	54.6/0.1	87.7/18.5	58.4/9.3	37.6/1.2	34.3/0.0	16.7/0.0	55.4/0.8	52.5/0.9	83.0/0.7	96.9/28.6	33.3
+PaCoMerge	81.5/63.2	50.8/13.5	78.9/52.5	50.4/29.0	35.3/17.4	27.3/10.6	15.8/2.0	49.4/20.0	54.6/47.4	81.0/45.8	94.8/78.4	**45.4**
ACE	82.9/49.7	53.2/4.2	85.5/51.6	57.7/25.9	38.1/10.7	27.8/1.2	16.6/0.2	51.9/9.2	54.3/36.3	82.9/28.9	95.8/75.7	42.7
+PaCoMerge	82.0/69.8	51.6/20.1	79.2/59.2	51.3/33.2	35.5/21.8	24.5/13.5	16.1/3.8	48.3/26.8	56.2/49.6	81.4/59.0	94.6/84.4	**48.3**

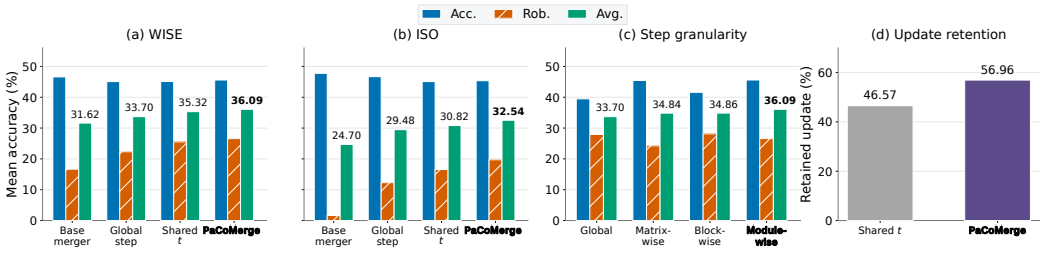


Figure 5: Selective correction on ViT-B/32. (a) and (b) Comparison with global and shared-step correction. (c) Effect of correction granularity. (d) Retained fraction of the base-merger update.

Zero-Shot Robustness. Table 3 evaluates PACOMERGE on 11 zero-shot classification benchmarks beyond the ImageNet distribution family. PACOMERGE improves the overall average for every base merger by 1.3–12.1 points, with the largest gains for mergers suffering severe robustness degradation. For example, ISO improves from 33.3 to 45.4 overall. This consistent improvement across diverse unseen domains suggests that the correction generalizes beyond ImageNet-derived shifts.

5.3 ROBUSTNESS AND ABLATION STUDIES

Sensitivity-Guided Correction. Table 4 compares correction criteria at a matched correction norm. PaCoMerge achieves the highest robust accuracy (26.58%) and overall average (36.09%).

Robustness across Attacks. Table 5 further shows consistent gains under FGSM, PGD-CE, PGD-CW, and AutoAttack across perturbation budgets from $1/255$ to $4/255$, indicating that the improvement is not specific to a particular attack setting.

Sensitivity to ρ. Figure 6 shows a smooth trade-off as ρ varies: relaxing the robustness constraint increases clean accuracy while gradually reducing robust accuracy. The observed curve suggests a gradual clean–robust trade-off over the evaluated range of ρ.

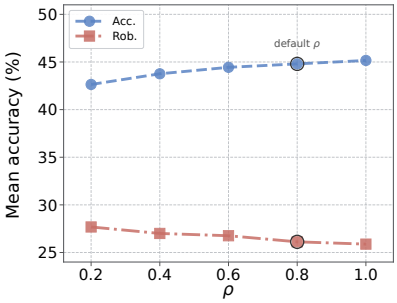


Figure 6: Sensitivity to the robustness tolerance ρ on ViT-B/32. Clean and robust accuracy (%) are averaged over ImageNet and four distribution shifts under AutoAttack.

5.4 UNDERSTANDING SELECTIVE CORRECTION

We next examine why selective correction is important for preserving the benefits of the base merger.

Table 4: Comparison of correction criteria at a matched correction norm.

Method	Acc.	Rob.	Avg.
WISE	**47.16**	16.02	31.59
$R_W(1)$ Top-4	45.35	23.90	34.63
Magnitude Top-4	44.92	20.74	32.83
Raw spectral Top-4	45.18	22.56	33.87
PaCoMerge	45.47	**26.68**	**36.08**

Table 5: Robust accuracy (%) of ViT-B/32 across attack objectives and perturbation budgets.

Method	Attack objective ($\epsilon = 1/255$)			AutoAttack budget		
	FGSM	PGD-CE	PGD-CW	1/255	2/255	4/255
FARE	29.48	29.44	28.84	28.44	21.76	10.28
WISE	23.60	18.80	19.44	16.48	4.04	0.32
+ PaCoMerge	**29.76**	**28.72**	**28.00**	**27.04**	**13.40**	**2.68**
ISO	14.68	3.32	3.48	1.64	0.24	0.04
+ PaCoMerge	**24.36**	**21.60**	**21.56**	**19.72**	**6.48**	**0.92**

Figures 5(a) and (b) compare PACOMERGE with less selective alternatives on WISE and ISO. A global correction recovers robustness but sacrifices more of the clean performance, while using a single shared step improves the balance but still underperforms PACOMERGE. This trend is consistent across both mergers. Figure 5(c) further studies the correction granularity. Moving from global to finer-grained correction improves the clean–robust balance, with module-wise correction achieving the highest overall average. Importantly, this gain does not come from simply moving further toward the robust source. As shown in Fig. 5(d), PACOMERGE retains 56.96% of the base-merger update, compared with 46.57% for a shared interpolation step. Together, these results show that selective correction preserves more useful merger updates while suppressing the components most harmful to robustness.

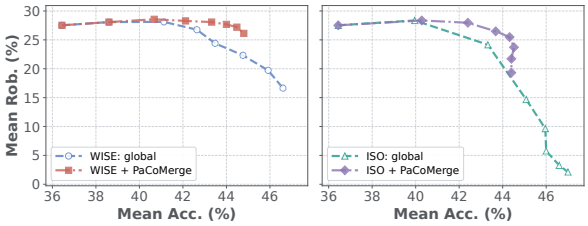


Figure 7: Clean–robust frontiers for WISE (left) and ISO (right). Both curves sweep t along the global interpolation path; PACOMERGE corrects each proposal at fixed $\rho = 0.8$. Axes: mean clean (x) and robust (y) accuracy (%).

Comparison with Global Interpolation. Figure 7 compares global interpolation with PACOMERGE applied to each interpolated proposal for WISE and ISO. Across much of the evaluated range, PACOMERGE achieves higher robust accuracy at comparable clean accuracy, especially where global interpolation loses robustness. This pattern is consistent with the update-retention analysis in Fig. 5(d), suggesting that preserving useful merge components helps achieve a better balance than globally shrinking the entire update.

5.5 Generalization across Backbones and Source Models

We further evaluate PACOMERGE on ViT-L/14 and ConvNeXt-B. Table 6 shows consistent robustness gains, from 33.38 to 45.60 and from 34.80 to 39.80, respectively, improving the clean–robust average to 54.73 and 53.50. We additionally evaluate FARE–BMA, TeCoA–OpenAI CLIP, and TeCoA–BMA source pairs (Appendix B.1, Tables 11–13). Across all three combinations, PACOMERGE improves mean robust accuracy for each of the five base mergers, with the largest gains for Breadcrumbs and ISO. These results support generalization across the evaluated source pairs.

Table 6: **Generalization across backbones.** PACOMERGE improves the clean–robust average on both ViT-L/14 and ConvNeXt-B. Robust accuracy is evaluated with AutoAttack at $\epsilon = 1/255$.

Method	ViT-L/14			ConvNeXt-B		
	Acc.	Rob.	Avg.	Acc.	Rob.	Avg.
Source models						
Natural	**71.12**	0.00	35.56	**69.60**	0.00	34.80
FARE	56.34	**46.98**	**51.66**	61.60	**50.80**	**56.20**
Base merger and PACOMERGE *output*						
WISE	**68.32**	33.38	50.85	**67.80**	34.80	51.30
+PACOMERGE	63.86	**45.60**	**54.73**	67.20	**39.80**	**53.50**

6 Conclusion

We study model merging between OOD generalized and adversarially robust VLMs under the same task and label space, and show that existing mergers can substantially degrade adversarial robustness despite preserving strong OOD performance. Motivated by the localized nature of this degradation, we introduce PACOMERGE, a training-free and data-free post-merging method that constrains source-relative perturbation amplification along the original merge path. We further show that the resulting feasible region forms a continuous interval, allowing PACOMERGE to make the minimum correction along the original update direction. Across diverse base mergers, VLM backbones, distribution shifts, attacks, and downstream tasks, PACOMERGE consistently recovers adversarial robustness while largely preserving the benefits of the original merge.

AI USE STATEMENT

We used generative AI tools solely for language polishing and assistance with writing research code. All other aspects of the research, including problem formulation, method design, theoretical analysis, experimental design and execution, and interpretation of results, were carried out by the authors. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was inspected and tested by the authors. The authors take full responsibility for the final manuscript, implementation, and experimental results.

ETHICS STATEMENT

This work studies model merging for improving the adversarial robustness of vision-language models using existing model checkpoints and publicly available benchmarks. The study does not involve recruiting human participants or collecting new personal data. Adversarial attacks are used in controlled experiments to evaluate model robustness. The research focuses on methodological development and benchmark evaluation and does not involve deployment in high-stakes settings.

REPRODUCIBILITY STATEMENT

Section 4 specifies the merge path, relative perturbation amplification constraint, and module-wise projection procedure of PACOMERGE, with theoretical derivations and proofs provided in Appendix C. Section 5.1 describes the experimental setup, while Appendix A reports source-model details, base-merger configurations, numerical implementation, hyperparameter settings, and evaluation protocols. Appendix B provides additional experiments, including hyperparameter sensitivity analyses and strengthened adversarial evaluations. Given fixed hyperparameters and source and merged checkpoints, PACOMERGE computes the correction directly from model weights without additional training or data. We will publicly release the code and configuration files to facilitate reproducibility.

REFERENCES

Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R. Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *CVPR*, 2024.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NeurIPS*, 2017.

Zhixiang Chi, Li Gu, Huan Liu, Ziqiang Wang, Yanan Wu, Yang Wang, and Konstantinos N Plataniotis. Learning to adapt frozen clip for few-shot test-time domain adaptation. In *ICLR*, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*. PMLR, 2020.

Francesco Croce, Sylvestre-Alvise Rebuffi, Evan Shelhamer, and Sven Gowal. Seasoning model soups for robustness to adversarial and natural distribution shifts. In *CVPR*, 2023.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *ECCV*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Junhao Dong, Xinghua Qu, Cong Zhang, Qi Rong Sua, Nguyen Duc Thai, Wenbo Pan, Xinfeng Li, Tongliang Liu, Piotr Koniusz, and Yew-Soon Ong. Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In *ICLR*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable adversarial training via spectral normalization. In *ICLR*, 2019.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, 2004.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021b.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. In *NeurIPS*, 2021.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR workshops*, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Lin Li, Haoyan Guan, Jianing Qiu, and Michael W. Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024a.

Lin Li, Yifei Wang, Chawin Sitawarin, and Michael W. Spratling. Oodrobustbench: a benchmark and large-scale analysis of adversarial robustness under distribution shift. In *ICML*, 2024b.

Yayuan Li, Ze Peng, Jian Zhang, Jintao Guo, Yue Duan, and Yinghuan Shi. When shared knowledge hurts: Spectral over-accumulation in model merging. In *ICML*, 2026.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Shangge Liu, Yuehan Yin, Yinghuan Shi, Lei Wang, and Wenbin Li. When muon meets task interference: A spectral perspective on continual learning and model merging. *arXiv preprint arXiv:2608.27518*, 2026.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 37(7):3031–3051, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartlomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. *ICML*, 2025.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *NeurIPS*, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *CVPR*, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. In *ICLR*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ICML*, 2024.

Nguyen Duc Thai, Junhao Dong, Sua Qi Rong, Hua Yu, and Yew-Soon Ong. Unifying adversarially robust model experts in vision-language models. *arXiv preprint arXiv:2607.27897*, 2026.

Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *CVPR*, 2025.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *NeurIPS*, 2018.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, 2018.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022b.

Songlong Xing, Zhengyu Zhao, and Nicu Sebe. CLIP is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of CLIP. In *CVPR*, 2025.

Bo Xu, Haotian Wu, Hehai Lin, Weiquan Huang, Beier Zhu, Yao Shu, and Chengwei Qin. Ace-merging: Data-free model merging with adaptive covariance estimation. In *CVPR*, 2026.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *NeurIPS*, 2023.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*, 2024.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. Improved OOD generalization via adversarial training and pretraing. In *ICML*, 2021.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *CVPR*, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pp. 1–20, 2024.

Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. In *ICCV*, 2023.

Xin Zou and Weiwei Liu. On the adversarial robustness of out-of-distribution generalization models. In *NeurIPS*, 2023.

PATH-CONSTRAINED MODEL MERGING FOR ROBUST VISION-LANGUAGE MODELS (*Supplementary Material*)

Anonymous authors
Paper under double-blind review

ABSTRACT

This supplementary material provides additional details and results for PACOMERGE. Appendix A describes the implementation, base merging methods, and experimental configurations. Appendix B reports additional evaluations across source-model pairs and settings. Appendix C provides the derivations and proofs for the theoretical results in the main paper. Appendix D discusses the limitations of our study and outlines directions for future work.

A REPRODUCIBILITY DETAILS

This appendix collects the implementation settings and evaluation protocols used in our experiments.

Correction hyperparameters. Unless otherwise stated, all experiments use $\rho = 0.8$ and $\eta = 10^{-3}$. These defaults were chosen heuristically and are shared across backbones, source-model pairs, and base mergers. The tolerance ρ controls the allowable relative amplification, while η regularizes both terms of the amplification ratio. Figure 6 and Table 15 report sensitivity to ρ and η, respectively.

Scope of data-free correction. Once these hyperparameters are fixed, PACOMERGE computes the correction from model weights without data, labels, gradients, or further training. This describes the correction procedure, not the data requirements of hyperparameter selection.

A.1 MODELS AND CLASSIFIER CONSTRUCTION

Source models. Our main experiments merge visual encoders with complementary generalization and adversarial robustness. Table 7 summarizes the source-model pairs used across architectures.

Table 7: **Source models used across architectures.**

Architecture	Generalization source	Robust source
ViT-B/32	OpenAI CLIP	FARE
ViT-L/14	OpenAI CLIP	FARE
ConvNeXt-B	OpenCLIP LAION-2B	FARE

For ViT-B/32, we instantiate OpenCLIP `ViT-B-32` with `openai` weights and use the released FARE checkpoint `vitb32_fare_eps_4.pt`. For ViT-L/14, we use OpenAI CLIP ViT-L/14 together with the corresponding FARE checkpoint `fare_eps_4.pt`. For ConvNeXt-B, the generalization source is OpenCLIP `convnext_base_w` with `laion2b_s13b_b82k_augreg` weights, and the robust source is the corresponding released FARE checkpoint.[1]

[1] `hf-hub:chs20/FARE4-convnext_base_w-laion2B-s13B-b82K-augreg`.

Table 8: **PACOMERGEimplementation settings.** Default settings are described in Appendix A.

Setting	Value
Tolerance ρ (default)	0.8
Ridge parameter η	10^{-3}
Search iterations	24
Feasibility tolerance	10^{-6}
Boundary computation	Float64
Attention group	$\{W_{\mathrm{QKV}}, W_O\}$
MLP group	$\{W_{\mathrm{fc}}, W_{\mathrm{proj}}\}$
Default granularity	Module-wise

Classification protocol. We use the official OpenCLIP evaluation preprocessing for all CLIP classification experiments. Images are bicubically resized and center-cropped to 224×224, converted to RGB tensors in $[0, 1]$, and normalized with mean $(0.48145466, 0.4578275, 0.40821073)$ and standard deviation $(0.26862954, 0.26130258, 0.27577711)$. CIFAR-10, CIFAR-100, and STL-10 are attacked at their native resolution, with resizing performed inside the differentiable model.

For each class, we independently encode all prompt templates and ℓ_2-normalize the resulting text features. The normalized features are then averaged and normalized again to form the final class embedding. Image features are likewise ℓ_2-normalized. Both classification and adversarial objectives use cosine-similarity logits multiplied by a fixed factor of 100; the learned CLIP logit-scale parameter is not used. The same text classifier is used for the source, merged, and corrected models within each benchmark.

Prompt templates. The ImageNet family uses the standard 80-template OpenAI CLIP ensemble. With the dataset aliases used in our evaluation pipeline, Caltech101, CIFAR-10, CIFAR-100, DTD, EuroSAT, Flowers102, PCAM, and Oxford-IIIT Pets use the same 80-template ensemble. Stanford Cars uses eight templates, while FGVC-Aircraft and STL-10 use two templates each.

A.2 PACOMERGE IMPLEMENTATION

Balanced regularization. The term *balanced* refers to adding the same regularization term $\eta\|\boldsymbol{u}\|_2^2$ to both the numerator and denominator of Eq. (4). Equivalently, both input-side Gram matrices in Eq. (6) include ηI. This regularization is independent of the exact low-dimensional evaluation described below.

Functional grouping. For each eligible weight matrix W, we compute its maximum feasible step $t_W^\star$ under the source-relative amplification constraint. Matrices within the same functional group share

$$t_g^\star = \min_{W \in \mathcal{W}_g} t_W^\star, \tag{13}$$

and are updated along the original merge path as

$$W_C = W_A + t_g^\star (W_M - W_A). \tag{14}$$

In the CLIP implementation used in our experiments, the fused QKV projection is treated as one matrix W_{QKV} rather than three independent Q, K, and V matrices. Each attention group therefore contains $\{W_{\mathrm{QKV}}, W_O\}$, while each MLP group contains $\{W_{\mathrm{fc}}, W_{\mathrm{proj}}\}$. Biases, normalization parameters, embeddings, and other ineligible parameters remain at their base-merger values.

For group-level analyses, we define the endpoint amplification by the worst matrix within each functional group:

$$R_g(t) = \max_{W \in \mathcal{W}_g} R_W(t). \tag{15}$$

Thus, the heatmaps in Fig. 2 report $R_g(1)$ at the uncorrected merge endpoint.

Table 9: **Base-merger configurations for two-source experiments.**

Method	Hyperparameter	Value
WISE	Mixing coefficient	0.5
TA	Task-vector scale	0.6
Breadcrumbs	$(\lambda, q_{\text{keep}}, q_{\text{remove}})$	$(0.6, 0.15, 0.006)$
ISO	Task-vector scale	0.5
ACE	λ; (ϵ, h, r)	0.5; $(10^{-5}, 0.3, 0.3)$

Exact low-dimensional feasibility computation. Equation (6) is defined using input-side Gram matrices. Let $A = W_A$ and $B = W(t)$, with $A, B \in \mathbb{R}^{m\times n}$, where $m = d_{\text{out}}$ and $n = d_{\text{in}}$. When $n \le m$, the criterion can be evaluated directly using $C_A = LL^\top$ and the largest eigenvalue of $L^{-1}C_tL^{-\top}$. For wide matrices ($m < n$) and $\rho > 0$, an exact output-side *feasibility test* follows by defining

$$\alpha = \rho\eta > 0, \qquad \beta = 1 + \rho, \qquad M_\rho = \beta A^\top A + \alpha I_n \succ 0.$$

Since $C_A \succ 0$, congruence transformations give

$$\begin{aligned} R_W(t) \le \beta &\iff B^\top B + \eta I_n \preceq \beta(A^\top A + \eta I_n) \\ &\iff B^\top B \preceq M_\rho \\ &\iff \lambda_{\max}\big(BM_\rho^{-1}B^\top\big) \le 1. \end{aligned}$$

For the last equivalence, set $X = BM_\rho^{-1/2}$. Then $B^\top B \preceq M_\rho$ is equivalent to $X^\top X \preceq I_n$, and hence to $XX^\top \preceq I_m$, because the two Gram matrices of this same X share their nonzero eigenvalues. Thus, the feasible set and its boundary $t_W^\star$ are exactly those defined in Eq. (7).

The Woodbury identity further avoids forming the $n \times n$ matrix M_ρ. With $S = \alpha I_m + \beta AA^\top \succ 0$,

$$H_\rho(t) := BM_\rho^{-1}B^\top = \frac{1}{\alpha}\left[BB^\top - \beta BA^\top S^{-1}AB^\top\right].$$

All Gram products and linear solves on the right involve $m \times m$ matrices. This is an equivalent test of Eq. (5), not a direct substitution of $W^\top W$ by $WW^\top$ in Eq. (6). In particular, separately replacing the two input-side Gram matrices by $AA^\top + \eta I_m$ and $BB^\top + \eta I_m$ does not generally preserve the generalized eigenvalue ratio; the cross term involving $BA^\top$ is essential. The value $\lambda_{\max}(H_\rho(t))$ is used only for the threshold test and is not identified with $R_W(t)$. Amplification values such as $R_g(1)$ retain the original definition in Eq. (6).

Numerical computation. For fixed A, ρ, and η, the Cholesky factor of S can be reused throughout the boundary search. The product $S^{-1}AB^\top$ is evaluated by solving $SY = AB^\top$, without explicitly forming an inverse, giving $H_\rho(t) = \alpha^{-1}(BB^\top - \beta BA^\top Y)$. Before applying `torch.linalg.eigvalsh`, the matrix is symmetrized as $(H_\rho(t) + H_\rho(t)^\top)/2$; the same symmetrization applies to the direct input-side computation. For $\rho = 0$, the original input-side test is used, since the output-side inverse above requires $\rho > 0$. Boundary computation is performed in float64, and the projected weights are cast back to the dtype of the base-merger checkpoint. With ρ and η fixed, no gradients or data samples are required by the correction.

Extension to ConvNeXt. For ConvNeXt-B, we apply PACOMERGE to the two pointwise MLP projections within each residual block and treat them as a single functional group. All remaining parameters, including spatial convolutions, normalization layers, and the text tower, are kept at their base-merger values.

A.3 BASE MERGER INSTANTIATION

In the two-source experiments, we apply PACOMERGE to five representative base mergers: WiSE-FT (WISE), Task Arithmetic (TA), Model Breadcrumbs, Isotropic Merging (ISO), and ACE. The proposal-building configurations are summarized in Table 9.

These values are fixed by our proposal-building scripts and checkpoint metadata and remain unchanged throughout evaluation. In particular, no adversarial accuracy is queried when constructing

Table 10: **Main evaluation settings.**

Setting	Benchmarks	Adversarial evaluation
Distribution shift	ImageNet, V2, R, Sketch, A	AutoAttack, $\epsilon = 1/255$
Zero-shot	11 classification benchmarks	AutoAttack, $\epsilon = 2/255$
Vision-language	COCO, Flickr30k, VQAv2, TextVQA	$\epsilon = 2/255,\ 4/255$

or selecting the base-merger proposals. The same merger configurations are retained when evaluating alternative source pairs. The standard multi-task setting is described separately with Table 14.

A.4 Evaluation Protocol

Unless otherwise specified, the main classification experiments use ViT-B/32 and evaluate 1,000 examples per dataset. We use a fixed sampling seed of 0 and evaluate all compared methods and attacks on the same ordered subset. Clean and adversarial evaluations therefore use identical examples. Results averaged across multiple datasets assign equal weight to each dataset. All classification ablations, including hyperparameter sensitivity, attack-strength checks, and perturbation-budget comparisons, use 500 examples per dataset. The vision-language experiments use ViT-L/14 and evaluate 500 examples per task. The ViT-L/14 and ConvNeXt-B backbone comparison evaluates robust accuracy with AutoAttack at $\epsilon = 1/255$.

The main evaluation settings are summarized in Table 10.

AutoAttack. All classification attacks use an ℓ_∞ threat model, with perturbation budgets defined in the raw RGB $[0, 1]$ space. We use the official AutoAttack (Croce & Hein, 2020) implementation with its `standard` configuration, comprising APGD-CE, APGD-T, targeted FAB, and Square Attack. Robust accuracy is computed from per-example outcomes: an example counts as robust only if it is initially classified correctly and remains correct under every attack in the ensemble.

Additional attacks. For the attack-objective comparison, FGSM uses one step of size ϵ. PGD-CE and PGD-CW use 20 iterations, one uniformly random initialization within the perturbation set, and a step size of $\epsilon/4$. PGD-CW maximizes the largest incorrect logit relative to the ground-truth logit.

Strengthened AutoAttack and transfer evaluations. The strengthened evaluation retains the full AutoAttack ensemble. We increase the APGD budgets to 200 iterations, with ten APGD-CE restarts and five APGD-T restarts per target; targeted FAB and Square Attack retain their official standard settings. The per-example success criterion above applies to both the default and strengthened evaluations. Transfer attacks are generated using the specified OpenAI or FARE surrogate and subsequently evaluated as fixed adversarial inputs on the target model.

Vision-language evaluation. We evaluate LLaVA-1.5-7B by directly using an OpenCLIP ViT-L/14 vision tower at 224×224 resolution. The robust, merged, and corrected vision checkpoints retain their native 224-resolution configuration. Neither the checkpoints nor their positional embeddings are converted to 336 resolution. Thus, the experiments do not use the default `openai/clip-vit-large-patch14-336` vision tower. COCO and Flickr30k use the Karpathy test splits, VQAv2 uses the official `val2014` split, and TextVQA uses its validation split. For each task, 500 examples are sampled without replacement using NumPy seed 42. Evaluation uses batch size one, float16 precision, temperature zero, and one beam; generation is limited to 20 new tokens for captioning and five for VQA.

Adversarial examples are generated by an untargeted ℓ_∞ APGD attack on the input image for 100 iterations. The attack starts from the clean image and uses an initial step size of ϵ with adaptive reduction. Its objective maximizes the teacher-forced autoregressive loss of the first reference caption for captioning tasks or the most frequent ground-truth answer for VQA. Gradients are propagated through the complete VLM to the image, while the model parameters and text inputs remain fixed.

Evaluation uncertainty. The reported accuracies are point estimates on the specified evaluation subsets. Using the same examples for compared methods permits paired comparisons, but does not

quantify uncertainty from sample selection or attack randomness. Small differences in the ablations should therefore be interpreted descriptively, rather than as evidence of statistical significance.

Aggregation metrics. For ImageNet and its four distribution shifts, mean clean and robust accuracy are defined by

$$\mathrm{MeanAcc} = \frac{1}{5}\sum_{d=1}^{5}\mathrm{Acc}_d, \qquad \mathrm{MeanRob} = \frac{1}{5}\sum_{d=1}^{5}\mathrm{Rob}_d. \tag{16}$$

When a single overall average is reported, we equally average all five clean and five robust values:

$$\mathrm{Avg.} = \frac{1}{10}\sum_{d=1}^{5}\left(\mathrm{Acc}_d + \mathrm{Rob}_d\right). \tag{17}$$

Analysis protocols. **Endpoint amplification.** The heatmaps in Fig. 2 use WISE and ISO proposals constructed from OpenAI CLIP and FARE on ViT-B/32. Each cell reports $R_g(1)$ from Eq. (15), i.e., the largest matrix-level endpoint amplification within the corresponding attention or MLP group.

Selective correction. For Fig. 3, functional groups are sorted by $R_g(1)$ in descending order. At each k, the top-k groups receive their complete PACOMERGEcorrection, while all unselected groups remain at the base-merger weights. Selecting an already feasible group with $t_g^\star = 1$ therefore leaves the checkpoint unchanged. The random baseline uses three fixed random orderings with seeds 7, 19, and 41; the curve reports their mean and the shaded region shows ± 1 standard deviation. Robustness recovery is the absolute change in mean robust accuracy relative to the uncorrected WISE proposal and is reported in percentage points.

Correction criteria. The controls in Table 4 rank functional groups using

$$S_g^{\mathrm{rel}} = \max_{W\in\mathcal{W}_g} R_W(1), \tag{18}$$

$$S_g^{\mathrm{mag}} = \left(\sum_{W\in\mathcal{W}_g}\|W_M - W_A\|_F^2\right)^{1/2}, \tag{19}$$

$$S_g^{\mathrm{raw}} = \max_{W\in\mathcal{W}_g}\frac{\sigma_{\max}^2(W_M)+\eta}{\sigma_{\max}^2(W_A)+\eta}. \tag{20}$$

Each control selects its top four functional groups and uses one shared interpolation step within the selected groups. The step is chosen such that its total correction norm matches that of PACOMERGE,

$$\|\theta_{\mathrm{control}} - \theta_M\|_2 = \|\theta_{\mathrm{PACOMERGE}} - \theta_M\|_2, \tag{21}$$

where the norm is computed over floating-point visual parameters. Unselected parameters remain at the base-merger values. The controls match correction magnitude, but not the number of selected parameters.

Correction granularity. Global correction shares a single feasible step across all eligible matrices. Matrix-wise correction assigns each matrix its own $t_W^\star$. For block-wise correction, all eligible attention and MLP matrices within Transformer block ℓ share

$$t_\ell = \min_{W\in\mathcal{W}_\ell} t_W^\star. \tag{22}$$

Module-wise correction instead shares a separate step within each attention or MLP functional group.

Retained update energy. Figure 5(d) measures how much of the original base-merger update remains after correction over the eligible attention/MLP matrices $\mathcal{E}$:

$$\mathrm{Retained}(\theta) = 100\,\frac{\sum_{W\in\mathcal{E}}\|W - W_A\|_F^2}{\sum_{W\in\mathcal{E}}\|W_M - W_A\|_F^2}. \tag{23}$$

This is an update-energy ratio rather than an average interpolation coefficient.

Table 11: Results with **FARE as the robust source and BMA as the generalization source**. We report clean (Acc.) and robust (Rob.) accuracy (%) on ImageNet and four distribution shifts under AutoAttack at $\epsilon = 1/255$. PACOMERGE consistently recovers robustness across all five base mergers while preserving most of their clean performance.

Method	ImageNet		ImageNet-V2		ImageNet-R		ImageNet-S		ImageNet-A		Avg.
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	
Source models											
BMA	65.40	0.20	55.00	0.20	68.40	0.00	40.00	0.00	32.40	0.00	26.16
FARE	49.80	39.00	41.00	30.00	51.00	40.40	26.20	16.80	7.80	4.40	30.64
Base mergers and PACOMERGE *outputs*											
WISE	**64.00**	29.00	**54.40**	22.80	**66.60**	26.60	36.00	15.80	**18.60**	1.60	33.54
+PACOMERGE	58.80	**39.40**	48.80	**31.40**	62.60	**43.20**	**38.20**	**25.20**	9.40	**3.00**	**36.00**
TA	**62.00**	34.60	**52.40**	24.00	**63.60**	39.20	**35.00**	13.40	**15.80**	2.00	34.20
+PACOMERGE	56.60	**40.20**	49.40	**30.00**	61.40	**46.00**	33.60	**19.80**	11.80	**3.80**	**35.26**
Breadcrumbs	**65.00**	2.80	**56.00**	2.20	**68.20**	7.00	**38.80**	2.00	**25.60**	0.20	26.78
+PACOMERGE	60.20	**35.40**	48.40	**24.60**	63.00	**41.00**	35.00	**18.00**	15.00	**3.20**	**34.38**
ISO	**62.80**	2.20	**51.40**	1.20	**68.00**	6.40	34.00	3.00	**22.60**	0.00	25.16
+PACOMERGE	58.40	**37.40**	48.60	**27.60**	64.00	**45.60**	**35.00**	**20.00**	16.00	**4.20**	**35.68**
ACE	**62.40**	21.20	**54.40**	12.80	**65.00**	19.20	**35.80**	5.20	**18.60**	0.80	29.54
+PACOMERGE	58.40	**39.20**	50.20	**28.00**	62.60	**46.00**	33.60	**19.40**	13.80	**3.80**	**35.50**

Clean–robust frontier. For Fig. 7, we fix the robust source θ_A and the original base-merger endpoint θ_M. For both WISE and ISO, we sweep

$$\theta_{\text{global}}(t) = \theta_A + t(\theta_M - \theta_A), \qquad t \in \{0, 0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 1\}.$$

For each t, the baseline evaluates $\theta_{\text{global}}(t)$ directly, and PACOMERGE corrects this same proposal relative to θ_A. We fix $\rho = 0.8$, $\eta = 10^{-3}$, the eligible weights, and module grouping; the proposal itself changes with t. All points use the same 500 examples per distribution and AutoAttack configuration at $\epsilon = 1/255$, with five distributions weighted equally. This paired comparison isolates the additional module-wise correction at each global interpolation point.

Sensitivity to ρ. The sensitivity analysis in Fig. 6 uses the equal-weight WISE proposal constructed from OpenAI CLIP and FARE on ViT-B/32. Only ρ is varied; all other settings are fixed. Results are equally averaged over ImageNet, ImageNet-V2, ImageNet-R, ImageNet-Sketch, and ImageNet-A using the 500-example ablation subsets.

B ADDITIONAL RESULTS

B.1 GENERALIZATION ACROSS SOURCE PAIRS

We further evaluate PACOMERGE with different combinations of robust and generalization source models to examine whether its effectiveness depends on a particular source pair. We consider FARE or TeCoA as the robust source, and OpenAI CLIP or BMA as the generalization source. Across all settings, PACOMERGE consistently improves adversarial robustness for all five base mergers, while preserving most of their clean accuracy. The gains are particularly pronounced for mergers that suffer severe robustness degradation, such as Breadcrumbs and ISO.

FARE + BMA. Table 11 shows that PACOMERGE improves robust accuracy for every base merger across all five evaluation sets. The largest gains occur for Breadcrumbs and ISO, whose average robust accuracy increases by 21.6 and 24.4 points, respectively. This pattern is consistent with the selective nature of PACOMERGE: mergers that lose more of the robustness of the source model receive larger corrections.

TeCoA + OpenAI CLIP. Table 12 gives a stronger test of source generalization by replacing FARE with TeCoA while retaining the original OpenAI CLIP checkpoint as the generalization

Table 12: Results with **TeCoA as the robust source and OpenAI CLIP as the generalization source**. Clean (Acc.) and robust (Rob.) accuracy (%) are reported on ImageNet and four distribution shifts under AutoAttack at $\epsilon = 1/255$.

Method	ImageNet		ImageNet-V2		ImageNet-R		ImageNet-S		ImageNet-A		Avg.
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	
Source models											
OpenAI	62.80	0.00	54.60	0.00	69.20	0.00	36.20	0.00	31.40	0.00	25.42
TeCoA	56.60	46.40	44.20	34.80	47.20	38.40	26.40	19.60	6.20	3.00	32.28
Base mergers and PACOMERGE *outputs*											
WISE	**65.60**	37.00	**54.20**	26.80	**65.80**	37.40	38.20	16.40	**15.00**	1.80	35.82
+PACOMERGE	62.00	**47.40**	50.40	**34.60**	60.40	**47.40**	**40.60**	**25.60**	9.40	**4.40**	**38.22**
TA	**64.00**	45.60	**55.00**	33.00	**62.40**	41.80	**36.80**	21.20	11.60	3.40	37.48
+PACOMERGE	61.20	**48.00**	52.00	**34.80**	58.80	**46.60**	36.60	**25.60**	11.60	**4.00**	**37.92**
Breadcrumbs	**66.20**	4.80	**55.60**	2.00	**68.60**	6.60	**41.20**	2.80	**26.20**	0.20	27.42
+PACOMERGE	64.80	**42.80**	52.20	**30.80**	63.00	**45.80**	40.20	**22.20**	14.80	**4.60**	**38.12**
ISO	59.60	1.00	49.80	1.40	**65.20**	7.20	34.60	4.40	**21.20**	0.00	24.44
+PACOMERGE	**62.40**	**44.00**	**53.40**	**32.80**	62.00	**46.40**	**38.80**	**25.60**	13.60	**5.60**	**38.46**
ACE	**64.60**	33.40	**57.00**	22.40	**65.00**	33.20	**38.20**	14.20	**16.20**	1.60	34.58
+PACOMERGE	62.00	**47.20**	51.80	**34.60**	60.80	**47.80**	37.80	**26.00**	12.00	**4.00**	**38.40**

Table 13: Results with **TeCoA as the robust source and BMA as the generalization source**. We report clean (Acc.) and robust (Rob.) accuracy (%) on ImageNet and four distribution shifts under AutoAttack at $\epsilon = 1/255$.

Method	ImageNet		ImageNet-V2		ImageNet-R		ImageNet-S		ImageNet-A		Avg.
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	
Source models											
BMA	65.40	0.20	55.00	0.20	68.40	0.00	40.00	0.00	32.40	0.00	26.16
TeCoA	56.60	46.40	44.20	34.80	47.20	38.40	26.40	19.60	6.20	3.00	32.28
Base mergers and PACOMERGE *outputs*											
WISE	**66.20**	38.00	**57.80**	26.00	**62.60**	35.40	**38.40**	24.00	**15.60**	2.80	36.68
+PACOMERGE	62.00	**48.80**	50.00	**34.20**	60.80	**47.80**	36.60	**30.40**	10.40	**4.00**	**38.50**
TA	**64.60**	45.40	**56.60**	31.40	**61.40**	43.00	**37.20**	22.20	**12.80**	2.80	37.74
+PACOMERGE	61.80	**48.00**	52.20	**35.20**	58.40	**47.20**	36.80	**25.40**	10.80	**3.80**	**37.96**
Breadcrumbs	**65.60**	6.40	**57.40**	4.60	**69.00**	10.20	**41.60**	5.00	**25.80**	0.20	28.58
+PACOMERGE	63.40	**43.00**	52.60	**29.40**	62.80	**45.60**	38.80	**22.60**	14.00	**3.20**	**37.54**
ISO	61.40	4.00	**54.20**	2.80	**65.20**	9.40	36.60	5.00	**22.40**	0.20	26.12
+PACOMERGE	61.40	**45.60**	53.80	**33.40**	61.40	**46.40**	**38.40**	**25.80**	13.00	**4.20**	**38.34**
ACE	**65.80**	34.60	**58.40**	23.20	**66.40**	34.40	**38.40**	16.00	**16.40**	1.80	35.54
+PACOMERGE	62.20	**49.00**	53.00	**34.60**	60.00	**47.40**	37.80	**25.00**	11.00	**4.20**	**38.42**

source. Across all five mergers, PACOMERGE again improves robustness. The average robust-accuracy gain ranges from 2.8 points for TA to 28.1 points for ISO, showing that the correction remains effective under a different robust source.

TeCoA + BMA. As shown in Table 13, the same trend persists when both source checkpoints are changed. PACOMERGE improves average robust accuracy for all five mergers, with gains of 7.8, 3.0, 23.5, 26.8, and 10.0 points for WISE, TA, Breadcrumbs, ISO, and ACE, respectively. Together with the previous two settings, these results show that the correction is not tied to a particular robust or generalization source.

Multi-Task Merging. We follow the vision benchmark of Marczak et al. (2025), using OpenAI-pretrained CLIP ViT-B/32 and ViT-B/16 and the task-specific checkpoints distributed through the

Table 14: Average clean task accuracy (%) on the 8-, 14-, and 20-task merging benchmarks of Marczak et al. (2025). Results are averaged equally across tasks.

	ViT-B/32			ViT-B/16		
Method	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
WISE	66.26	64.34	60.98	**72.26**	**69.48**	**65.30**
+PaCoMerge	66.26	**64.39**	**61.09**	72.24	69.37	65.24
TA	71.15	65.38	60.48	**75.40**	**70.71**	**65.68**
+PaCoMerge	**71.34**	**65.58**	**60.67**	75.22	70.59	65.48
Breadcrumbs	77.18	63.80	49.90	**81.98**	**70.07**	**57.10**
+PaCoMerge	**77.31**	**64.12**	**50.37**	81.89	69.99	56.89
ISO	**86.38**	**81.43**	**77.61**	**91.17**	**86.29**	**82.32**
+PaCoMerge	86.34	81.21	77.43	91.10	86.17	82.19
ACE	70.04	66.36	62.57	**74.96**	**70.90**	**66.74**
+PaCoMerge	**70.18**	**66.40**	**62.71**	74.87	70.83	66.72

Table 15: Sensitivity to the fixed ridge parameter on the WISE/FARE proposal. Results are mean accuracy (%) over five ImageNet distributions with 500 examples each. All settings use balanced regularization as in Eq. (4), module-wise steps, and $\rho = 0.8$.

η	Acc.	Rob.	Avg.
10^{-4}	44.67	**27.80**	36.23
10^{-3}	45.80	27.06	**36.43**
10^{-2}	**46.23**	26.54	36.37

official ISO implementation (the Consensus Merging/TALL-Masks checkpoint collection). Let θ_0 denote their common pretrained initialization and $\Delta_i = \theta_i - \theta_0$ the update for task i. Each merger combines the T task checkpoints into a proposal θ_M; WISE uses equal-weight averaging, $\theta_M^{\text{WISE}} = \frac{1}{T}\sum_{i=1}^{T}\theta_i$.

Correction reference and procedure. We set $\theta_A = \theta_0$, using the original pretrained CLIP as the reference without an additional adversarially trained checkpoint. Each proposal is corrected by Eq. (9) along $W(t) = W_0 + t(W_M - W_0)$, with amplification measured against the pretrained matrix W_0. The base-merger and paired PACOMERGE rows report the proposal before and after correction under the same task evaluation.

Task sets and evaluation. The 8-task set consists of Cars, DTD, EuroSAT, GTSRB, MNIST, RESISC45, SUN397, and SVHN. The 14-task set adds CIFAR-100, STL-10, Flowers102, Oxford-IIIT Pets, PCAM, and FER2013. The 20-task set further adds EMNIST, CIFAR-10, Food101, Fashion-MNIST, RenderedSST2, and KMNIST. We use ISO's task-specific CLIP classification heads and report the unweighted mean of clean test accuracies over the T tasks (Table 14).

Task preservation. Across five mergers, two architectures, and three task scales, PACOMERGE changes mean task accuracy by between -0.22 and $+0.47$ percentage points (Table 14). This evaluates preservation of clean multi-task performance after correction.

B.2 Attack-Strength Sanity Check

We further examine whether the robustness gains of PACOMERGEare sensitive to attack strength. We consider a strengthened AutoAttack evaluation as well as surrogate-transfer attacks generated from OpenAI CLIP and FARE. We use the same PACOMERGEconfiguration as in the main experiments, with $\eta = 10^{-3}$, $\rho = 0.8$, balanced regularization as in Eq. (4), and functional module grouping. White-box and transfer evaluations each use 500 examples per distribution.

Table 16: Robustness under strengthened white-box and surrogate-transfer attacks on ViT-B/32 at $\epsilon = 1/255$. Values report mean robust accuracy (%) across five ImageNet distributions.

	White-box		Transfer	
Method	Default	Strengthened	OpenAI	FARE
WISE	16.42	16.42	43.08	42.53
+ PACOMERGE	**26.58**	**26.42**	**44.11**	**44.87**

Table 17: Robustness across perturbation budgets on ViT-L/14, using OpenAI and TeCoA as source models. Results use 500 ImageNet examples and AutoAttack. The same PaCoMerge checkpoint is evaluated at every budget, with fixed $\eta = 10^{-3}$ and $\rho = 0.8$. Bold marks the better result within the WISE/PaCoMerge pair.

		Rob. (ϵ)		
Method	Acc.	$1/255$	$2/255$	$4/255$
OpenAI	72.0	0.0	0.0	0.0
TeCoA	74.4	69.1	62.5	47.1
WISE	**80.2**	55.8	26.6	1.0
+PaCoMerge	79.3	**66.2**	**51.2**	**21.4**

As shown in Table 16, the robustness gain of PACOMERGEremains substantial under the strengthened white-box attack: its mean robust accuracy changes from 26.58% to 26.42%, while WISE remains at 16.42%. The corresponding gains over WISE are 10.16 and 10.00 percentage points. Under surrogate-transfer attacks, PACOMERGE improves over WISE by 1.03 points for the OpenAI CLIP surrogate and 2.34 points for the FARE surrogate.

Robustness at larger perturbation budgets. Table 17 extends the budget evaluation to ViT-L/14 with TeCoA as the robust source. At $\epsilon = 4/255$, PaCoMerge improves WISE robust accuracy from 1.0% to 21.4%, while clean accuracy decreases from 80.2% to 79.3%. The near-zero robust accuracy observed at this budget in the ViT-B/32–FARE experiments is therefore a limitation of that evaluated setting, rather than a uniform outcome across backbone–source combinations. PaCoMerge nevertheless remains below TeCoA's 47.1% robust accuracy, demonstrating partial recovery over the base merger rather than full preservation of source robustness.

Table 18: Effect of the robustness constraint on ViT-B/32 with WISE and the FARE anchor. Results are averaged over ImageNet and four distribution shifts, using 500 examples per distribution and AutoAttack at $\epsilon = 1/255$. Both variants use the same module grouping and eligible weights, with matched total correction norm. Avg. is the mean of clean and robust accuracy.

Constraint	Acc.	Rob.	Avg.
Raw spectral	42.92	23.16	33.04
Source-relative (ours)	**44.80**	**26.12**	**35.46**

Choice of robustness constraint. Table 18 compares raw-spectral and source-relative constraints with the same module grouping and a matched correction norm. The source-relative constraint raises mean robust accuracy from 23.16% to 26.12% and mean clean accuracy from 42.92% to 44.80%, improving their average from 33.04% to 35.46%.

Computational Efficiency. PACOMERGE is applied once after model merging and directly updates the model weights. As shown in Table 19, correcting ViT-B/32 and ViT-L/14 takes only 26.2 and 39.1 seconds, respectively, on a single NVIDIA A800-SXM4-80GB GPU. The additional peak GPU memory is only 130.2 MiB for ViT-B/32 and 224.2 MiB for ViT-L/14. Since PACOMERGE does not introduce additional modules or parameters, the corrected model has the same architecture and inference cost as the base merged model.

Table 19: **Computational efficiency of PACOMERGE.** The offline correction requires modest time and memory and introduces no inference-time overhead.

	ViT-B/32	ViT-L/14
Correction time (s)	26.2 ± 2.0	39.1 ± 0.6
Extra GPU memory (MiB)	130.2	224.2
Inference overhead	None	None

C PROOFS

We provide the derivation of the relative perturbation amplification form in Eq. (6), derive the perturbation-propagation bound used in Sec. 4.5, and prove Theorem 1.

C.1 DERIVATION OF RELATIVE PERTURBATION AMPLIFICATION

Recall the regularized relative perturbation amplification

$$R_W(t) = \sup_{\boldsymbol{u}\neq\mathbf{0}} \frac{\|W(t)\boldsymbol{u}\|_2^2 + \eta\|\boldsymbol{u}\|_2^2}{\|W_A\boldsymbol{u}\|_2^2 + \eta\|\boldsymbol{u}\|_2^2}. \tag{24}$$

Define

$$C_A = W_A^\top W_A + \eta I, \qquad C_t = W(t)^\top W(t) + \eta I. \tag{25}$$

Then Eq. (24) can be written as the generalized Rayleigh quotient

$$R_W(t) = \sup_{\boldsymbol{u}\neq\mathbf{0}} \frac{\boldsymbol{u}^\top C_t \boldsymbol{u}}{\boldsymbol{u}^\top C_A \boldsymbol{u}}. \tag{26}$$

Since $\eta > 0$, $C_A \succ 0$ and its symmetric inverse square root is well defined. Let

$$\boldsymbol{v} = C_A^{1/2}\boldsymbol{u}, \qquad \boldsymbol{u} = C_A^{-1/2}\boldsymbol{v}. \tag{27}$$

Substituting into Eq. (26) yields

$$\begin{aligned} R_W(t) &= \sup_{\boldsymbol{v}\neq\mathbf{0}} \frac{\boldsymbol{v}^\top C_A^{-1/2} C_t C_A^{-1/2} \boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{v}} \\ &= \lambda_{\max}\left(C_A^{-1/2} C_t C_A^{-1/2}\right), \end{aligned} \tag{28}$$

where the last equality follows from the Rayleigh–Ritz characterization of the largest eigenvalue. This recovers Eq. (6).

C.2 PERTURBATION PROPAGATION BOUND

We derive the perturbation-propagation bound used in Sec. 4.5. Suppose that

$$R_W(t) \le 1 + \rho. \tag{29}$$

By Eq. (6), Eq. (29) is equivalent to

$$\lambda_{\max}\left(C_A^{-1/2} C_t C_A^{-1/2}\right) \le 1 + \rho. \tag{30}$$

Since $C_A^{-1/2} C_t C_A^{-1/2}$ is symmetric, this is equivalent to

$$C_A^{-1/2} C_t C_A^{-1/2} \preceq (1+\rho) I. \tag{31}$$

Applying the congruence transformation induced by $C_A^{1/2}$ gives

$$C_t \preceq (1+\rho) C_A. \tag{32}$$

Therefore, for any perturbation $\boldsymbol{\delta}$,

$$\boldsymbol{\delta}^\top C_t \boldsymbol{\delta} \leq (1+\rho)\boldsymbol{\delta}^\top C_A \boldsymbol{\delta}. \tag{33}$$

Expanding C_t and C_A gives

$$\|W(t)\boldsymbol{\delta}\|_2^2 + \eta\|\boldsymbol{\delta}\|_2^2 \leq (1+\rho)\left(\|W_A\boldsymbol{\delta}\|_2^2 + \eta\|\boldsymbol{\delta}\|_2^2\right). \tag{34}$$

Rearranging yields

$$\|W(t)\boldsymbol{\delta}\|_2^2 \leq (1+\rho)\|W_A\boldsymbol{\delta}\|_2^2 + \rho\eta\|\boldsymbol{\delta}\|_2^2, \tag{35}$$

which is the perturbation-propagation bound stated in Sec. 4.5.

C.3 Proof of Theorem 1

Proof. Fix a weight matrix W and write

$$\Delta W = W_M - W_A, \qquad W(t) = W_A + t\Delta W, \qquad t \in [0,1]. \tag{36}$$

For any fixed $\boldsymbol{u} \neq \mathbf{0}$, define

$$r_{\boldsymbol{u}}(t) = \frac{\|W(t)\boldsymbol{u}\|_2^2 + \eta\|\boldsymbol{u}\|_2^2}{\|W_A\boldsymbol{u}\|_2^2 + \eta\|\boldsymbol{u}\|_2^2}. \tag{37}$$

The denominator is strictly positive and independent of t. Using Eq. (36), the numerator expands as

$$\begin{aligned}\|W(t)\boldsymbol{u}\|_2^2 &= \|W_A\boldsymbol{u} + t\Delta W\boldsymbol{u}\|_2^2 \\ &= \|W_A\boldsymbol{u}\|_2^2 + 2t\langle W_A\boldsymbol{u}, \Delta W\boldsymbol{u}\rangle + t^2\|\Delta W\boldsymbol{u}\|_2^2.\end{aligned} \tag{38}$$

The coefficient of t^2 is non-negative, so $r_{\boldsymbol{u}}(t)$ is convex in t.

By definition,

$$R_W(t) = \sup_{\boldsymbol{u}\neq\mathbf{0}} r_{\boldsymbol{u}}(t). \tag{39}$$

A pointwise supremum of convex functions is convex; hence $R_W(t)$ is convex on $[0,1]$. It is also finite on this interval because $\eta > 0$ and $W(t)$ is bounded for $t \in [0,1]$. Furthermore, Eq. (6) expresses $R_W(t)$ as the largest eigenvalue of a symmetric matrix that depends continuously on t, so $R_W(t)$ is continuous.

Consider the feasible set

$$\mathcal{T}_W(\rho) = \{t \in [0,1] : R_W(t) \leq 1+\rho\}. \tag{40}$$

Convexity of R_W implies that $\mathcal{T}_W(\rho)$ is convex, while continuity implies that it is closed. Moreover,

$$R_W(0) = 1 \leq 1+\rho, \tag{41}$$

so $\mathcal{T}_W(\rho)$ is nonempty and contains 0. Because it is a closed subset of the compact interval $[0,1]$, its maximum exists. Define

$$t_W^\star = \max \mathcal{T}_W(\rho). \tag{42}$$

Since $\mathcal{T}_W(\rho)$ is convex and contains both 0 and $t_W^\star$, it contains every point between them. By maximality of $t_W^\star$, it contains no point larger than $t_W^\star$. Therefore,

$$\mathcal{T}_W(\rho) = [0, t_W^\star]. \tag{43}$$

Now consider a module g with weight matrices $\mathcal{W}_g$. A shared step is feasible exactly when it is feasible for every matrix in the module. Hence

$$\begin{aligned}\mathcal{T}_g(\rho) &= \bigcap_{W\in\mathcal{W}_g} \mathcal{T}_W(\rho) \\ &= \bigcap_{W\in\mathcal{W}_g} [0, t_W^\star] \\ &= \left[0, \min_{W\in\mathcal{W}_g} t_W^\star\right].\end{aligned} \tag{44}$$

It follows that

$$t_g^\star = \min_{W \in \mathcal{W}_g} t_W^\star = \max\{t \in [0,1] : R_W(t) \le 1+\rho,\ \forall W \in \mathcal{W}_g\}. \tag{45}$$

It remains to show the minimum-correction property. Along the shared module path, define

$$J_g(t) = \sum_{W \in \mathcal{W}_g} \|W_A + t(W_M - W_A) - W_M\|_F^2. \tag{46}$$

Since

$$W_A + t(W_M - W_A) - W_M = -(1-t)(W_M - W_A), \tag{47}$$

we obtain

$$J_g(t) = (1-t)^2 \sum_{W \in \mathcal{W}_g} \|W_M - W_A\|_F^2. \tag{48}$$

Thus, $J_g(t)$ is non-increasing on $[0,1]$. The jointly feasible steps form the interval $[0, t_g^\star]$, so $J_g(t)$ is minimized at its largest feasible point, $t = t_g^\star$. Therefore,

$$t_g^\star \in \arg\min_{t \in [0,1]} \sum_{W \in \mathcal{W}_g} \|W_A + t(W_M - W_A) - W_M\|_F^2 \tag{49}$$

subject to

$$R_W(t) \le 1+\rho, \qquad \forall W \in \mathcal{W}_g. \tag{50}$$

If at least one matrix in the module satisfies $W_M \neq W_A$, the objective in Eq. (48) is strictly decreasing and the minimizer is unique. This proves Theorem 1. □

D Limitation and Future Work

E Limitation and Future Work

Our empirical study primarily focuses on CLIP-based vision encoders and their integration into downstream vision-language systems. While the experiments cover multiple backbone architectures, source-model pairs, base merging methods, and downstream tasks, the evaluated settings are still largely centered on CLIP-style visual representations. The extent to which the proposed path-constrained correction generalizes to substantially different vision-language architectures and representation structures therefore remains an open question. In addition, our robustness constraint is defined at the level of individual linear transformations. While it provides an explicit bound on local perturbation propagation and consistently correlates with improved adversarial robustness in our experiments, it does not constitute a formal guarantee of end-to-end robustness for the full nonlinear model. Establishing tighter connections between weight-space amplification and network-level adversarial robustness is an interesting direction for future work. A natural direction for future work is to evaluate the method on a broader range of model families, modalities, and application domains, including more recent multimodal architectures and settings beyond image–text understanding. It would also be interesting to study whether similar robustness-preserving constraints can be extended from the visual encoder to other components of multimodal systems. These extensions would help better characterize the scope and generality of the proposed approach.