ATTENTION SENSITIVITY IN VISION TRANSFORMERS: ANALYSIS AND ROBUST SMALL-DATA TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers (ViTs) have achieved strong performance in large-scale computer vision tasks but often generalize poorly when trained on small datasets. Recent studies suggest that flat attention is associated with poor generalization, yet how attention patterns affect local sensitivity to perturbations remains insufficiently understood. To study this relationship, we analyze how bounded input shifts induce perturbations in attention logits and use this formulation to characterize attention-level sensitivity. Empirically, models trained on small datasets exhibit higher sensitivity to such perturbations and flatter attention patterns than models pre-trained on large-scale datasets. We further provide a local sensitivity analysis showing that the softmax Jacobian vanishes as attention distributions approach a sharp, one-hot pattern, thereby attenuating the first-order loss variation propagated through attention. This analysis also clarifies that temperature-based sharpening alone does not necessarily reduce sensitivity. Building on this local analysis and the empirical sensitivity gap, we propose Gradient-Guided Robust Attention Sharpening (GRASP), a training-time regularization framework that enforces prediction consistency under gradient-guided attention perturbations and implicitly encourages sharper attention patterns. Extensive experiments demonstrate that GRASP consistently improves ViT generalization in small-data training and out-of-distribution evaluation, while providing additional gains in target-domain fine-tuning.

1 INTRODUCTION

Vision Transformers (ViTs) (Dosovitskiy et al., 2021; Liu et al., 2021c; Touvron et al., 2021) have shown outstanding performance in various computer vision tasks (Seo et al., 2023; Yun et al., 2023; Hwang et al., 2026), including image classification, object detection, and instance segmentation. Leveraging the Transformer architecture (Vaswani et al., 2017), ViTs divide an image into patches and use an attention mechanism to model long-range dependencies among patches. The success of attention and ViTs drives an emerging trend to apply ViTs to various data-limited settings (Liu et al., 2021c; Touvron et al., 2021; Fan et al., 2021; Wang et al., 2021; Heo et al., 2021; Wu et al., 2021; Yuan et al., 2021a;b) where practical tasks, such as medical imaging (Bai et al., 2025; Song et al., 2025) and industrial inspection (Asad et al., 2025), are generally associated with limited data.

Training a ViT model on a small dataset is generally non-trivial, as the model may suffer from overfitting (Liu et al., 2021b; Huan et al., 2023; Das et al., 2024; Lee et al., 2021; Chen et al., 2023; Zhang et al., 2025; Verma et al., 2025). Specifically, unlike CNNs, which structurally encode inherent inductive biases such as locality, translation invariance, and hierarchical feature learning (Wang & Wu, 2023; Cohen & Shashua, 2016; Worrall et al., 2017; Kayhan & Gemert, 2020), ViTs learn such priors directly from data, which typically requires large-scale pre-training, e.g., on JFT-300M (Sun et al., 2017), to achieve strong performance on ImageNet-1K (Russakovsky et al., 2015). However, training a ViT from scratch with limited data often leads to poor test accuracy, as the model may fail to learn and generalize long-range dependencies among patches. To mitigate this issue, prior works introduce architectural inductive biases (Liu et al., 2021c; Touvron et al., 2021; Fan et al., 2021; Wang et al., 2021; Heo et al., 2021; Wu et al., 2021; Yuan et al., 2021a;b) or employ specialized variants and self-supervised objectives (Liu et al., 2021b; Huan et al., 2023; Das et al., 2024; Zhang et al., 2025; Verma et al., 2025).

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.

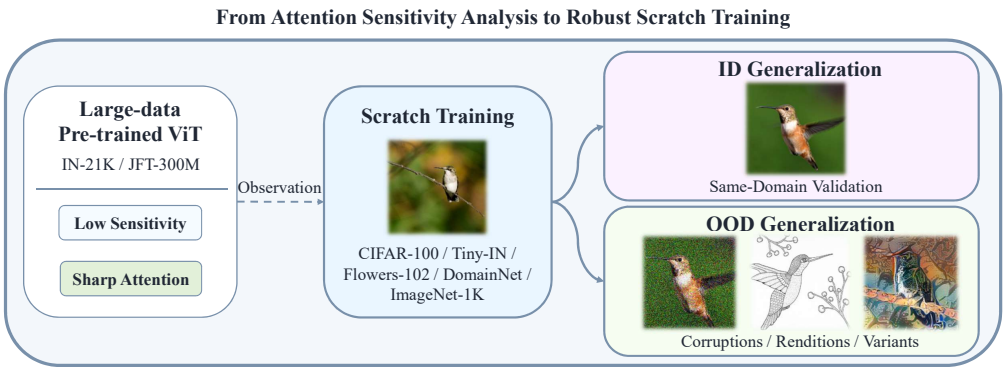


Figure 1: Motivation and scope of this work. By analyzing large-data pre-trained ViTs, we identify attention characteristics associated with strong generalization. Motivated by this observation, we study attention-level sensitivity in scratch-trained ViTs and examine its relationship with both in-distribution (ID) and out-of-distribution (OOD) generalization.

Recent efforts (Lee et al., 2021; Chen et al., 2023; Zhou et al., 2021) have specifically focused on attention, a fundamental component that distinguishes ViTs from CNNs. In particular, ViTs trained on small datasets have been reported to exhibit flat attention, where attention maps become overly uniform and fail to emphasize discriminative regions (Lee et al., 2021; Zhou et al., 2021). Accordingly, several studies (Lee et al., 2021; Zou et al., 2024; Veličković et al., 2025) suggest adjusting the softmax temperature to control the sharpness of attention distributions, thereby improving model generalization and test accuracy. While attention sharpening appears promising for improving ViT training on small datasets, how attention sharpness affects local attention sensitivity, and how this sensitivity relates to generalization, remain insufficiently understood.

This work presents GRASP, an effective approach for training ViTs on small datasets, to address these shortcomings. As summarized in Figure 1, we analyze large-data pre-trained ViTs as a reference for low-sensitivity attention and use the resulting insights to improve the generalization of scratch-trained ViTs. First, we establish an attention-logit perturbation model to characterize how bounded input changes induce perturbations in attention logits and post-softmax attention distributions. Based on this formulation, we derive the first-order worst-case direction under a fixed softmax-effective perturbation magnitude and construct a gradient-guided perturbed attention distribution through attention scaling. Our local sensitivity analysis shows that as the attention distribution becomes sharper and approaches a one-hot distribution, the softmax Jacobian approaches zero, attenuating the first-order loss variation propagated through attention. This leads to two key observations: (1) sharper attention can reduce local loss sensitivity through attention, and (2) temperature scaling does not necessarily reduce this sensitivity. Motivated by these findings, GRASP enforces consistency between the original model output and the perturbed output under the derived gradient-guided attention perturbation.

Our main contributions are summarized as follows:

- **Theoretical Analysis:** We analyze how bounded input shifts propagate to attention logits and formulate the resulting changes as bounded attention-logit perturbations to characterize the resulting local loss variation through attention.

- **Principled Mitigation**: Motivated by this analysis, we introduce GRASP, a consistency regularization method based on gradient-guided attention perturbations. Unlike direct temperature-based sharpening, GRASP enforces prediction stability under the derived perturbed attention, thereby encouraging sharper attention patterns and reducing local attention sensitivity.

- **Empirical Validation**: We validate GRASP on small-data ViT training and out-of-distribution evaluation, with an additional extension to target-domain fine-tuning.

2 Related Work

In this section, we review the evolution of Vision Transformers, the specific challenges of training ViTs on small datasets, and general regularization techniques.

Vision Transformers on ImageNet-1K. ViT (Dosovitskiy et al., 2021) effectively captures long-range dependencies but requires large-scale datasets such as JFT-300M (Sun et al., 2017) to outperform CNNs. To improve data efficiency on ImageNet-1K (Russakovsky et al., 2015), various works have introduced hierarchical structures (Liu et al., 2021c; Wang et al., 2021; Heo et al., 2021; Yuan et al., 2021a;b; Wang et al., 2022; Han et al., 2021), inductive biases (Touvron et al., 2021; Wu et al., 2021; Park & Kim, 2022), or local self-attention (Liu et al., 2021c; Zhang et al., 2022; Yuan et al., 2023). Despite these improvements, ViTs remain less effective when trained on smaller datasets from scratch (Liu et al., 2021b; Das et al., 2024; Lee et al., 2021).

Vision Transformers on Small Datasets. Training ViTs from scratch on small datasets is crucial for domains where pre-training is ineffective or computationally prohibitive, such as medical or industrial imaging (Bai et al., 2025; Liu et al., 2021b; Yu et al., 2024). As a result, there has been an increasing amount of research focused on training ViTs from scratch on small datasets. This research can be broadly categorized into studies leveraging self-supervised learning (Liu et al., 2021b; Huan et al., 2023; Das et al., 2024; Li et al., 2022) and those on model architecture (Lee et al., 2021; Chen et al., 2023; Zhang et al., 2025; Hassani et al., 2021; Akkaya et al., 2024). Some studies recognize the issue of flat attention (Lee et al., 2021; Chen et al., 2023), and recently PAEViT (Verma et al., 2025) proposes a patch attention excitation mechanism to rescale attention scores. However, these approaches do not explicitly characterize the relationship between attention patterns and sensitivity. Moreover, as shown in our experiments, substantial sensitivity to attention perturbations can remain after applying these methods.

Regularization. Various regularization techniques have been studied to prevent overfitting and improve generalization. At the input level, data augmentation (Zhong et al., 2020; DeVries & Taylor, 2017) and consistency losses (Hendrycks et al., 2019; Hwang et al., 2024; Fang et al., 2023) are studied for improving generalization. While AdvProp (Xie et al., 2020) improves generalization by leveraging adversarial samples, its reliance on dual batch normalization (Ioffe & Szegedy, 2015) is incompatible with Transformer architectures. At the weight level, Sharpness-Aware Minimization (SAM) (Foret et al., 2021) explicitly seeks flatter minima in the weight space to enhance robustness. Feature-level methods such as TENET (Liu et al., 2021a) suppress discriminative activations in CNNs to encourage auxiliary feature discovery. In contrast, our work focuses on the attention mechanism, characterizing attention sensitivity and its relationship with model generalization for ViTs.

3 Attention Sensitivity Analysis

This section analyzes attention sensitivity in ViTs trained on limited data. We model input shifts as bounded attention-logit perturbations and derive a gradient-guided worst-case direction based on the resulting local loss variation. We then relate the sensitivity gap between scratch-trained and pre-trained ViTs to attention sharpness through the softmax Jacobian and discuss why temperature-based sharpening does not necessarily reduce sensitivity.

3.1 From Input Shifts to Attention-Logit Perturbations

For a single attention head, let $z(x) \in \mathbb{R}^N$ denote the attention logits for input x, and let $A = \mathrm{softmax}(z(x))$ be the corresponding attention distribution. We analyze how input-level shifts induce perturbations in $z(x)$. Let $x \in \mathbb{R}^d$ be a reference input and consider a label-preserving perturbation

$$x' = x + \delta x, \qquad \|\delta x\|_p \le \rho, \tag{1}$$

where $p \in [1, \infty]$ specifies the input perturbation geometry. Such bounded perturbation views are widely used to study adversarial robustness (Goodfellow et al., 2014; Madry et al., 2018) and have also been linked to robustness under common corruptions and natural shifts (Xie et al., 2020; Rusak et al., 2020). Here, we use this local robustness perspective to analyze attention sensitivity.

The induced attention-logit perturbation is

$$\Delta z := z(x + \delta x) - z(x). \tag{2}$$

Assuming that z is continuously differentiable along the segment between x and $x + \delta x$, we have

$$\Delta z = \int_0^1 J_z(x + t\delta x)\delta x \, dt, \tag{3}$$

where $J_z(x) = \partial z(x)/\partial x \in \mathbb{R}^{N\times d}$. If the Jacobian is locally bounded in the induced operator norm by $L_{p,q} > 0$ as

$$\|J_z(x + t\delta x)\|_{p\to q} \le L_{p,q} \qquad \text{for all } t \in [0, 1], \tag{4}$$

then

$$\|\Delta z\|_q \le \int_0^1 \|J_z(x + t\delta x)\|_{p\to q}\|\delta x\|_p \, dt \le L_{p,q}\rho. \tag{5}$$

Thus, bounded input perturbations induce bounded attention-logit perturbations.

Definition 3.1 (Attention-logit perturbation model)**.** Given attention logits $z \in \mathbb{R}^N$, the perturbed attention distribution is

$$\tilde{A} = \mathrm{softmax}(z + \Delta z), \qquad \|\Delta z\|_q \le \epsilon. \tag{6}$$

Here, q specifies the perturbation geometry of the induced attention-logit perturbation, and ϵ controls the magnitude. Under the local Lipschitz condition above, $\epsilon = L_{p,q}\rho$.

3.2 Gradient-Guided Attention Perturbation

Given the attention-logit perturbation model above, we now derive the first-order direction that maximizes the loss variation. This direction characterizes the most sensitive local change in the attention logits. Let $G \in \mathbb{R}^N$ denote the gradient of the loss with respect to the attention distribution, where $G_i = \partial\mathcal{L}/\partial A_i$. For a sufficiently small attention-logit perturbation Δz, the induced attention variation satisfies

$$\Delta A \approx J\Delta z, \qquad J := \frac{\partial A}{\partial z} = \mathrm{diag}(A) - AA^\top. \tag{7}$$

Accordingly, under a first-order Taylor expansion, the corresponding loss variation is

$$\Delta\mathcal{L}(\Delta z) \approx G^\top \Delta A \approx G^\top J\Delta z. \tag{8}$$

Since softmax is invariant to constant logit shifts, we measure the effective perturbation magnitude after removing this shift-invariant component, as detailed in Appendix A.1, while the L_q bound defines the admissible perturbation region. This leads to the following result.

Proposition 3.2 (Gradient-Guided Worst-Case Direction)**.** *Under a first-order Taylor expansion and a fixed effective perturbation magnitude, the worst-case attention-logit perturbation direction can be represented as*

$$\Delta z^* \propto G, \qquad G = \nabla_A\mathcal{L}. \tag{9}$$

Proof. The proof is provided in Appendix A.1. □

Since the gradient G is defined in terms of the attention distribution A, the additive perturbation in Equation (6) is reformulated as a multiplicative re-weighting of the attention distribution, which we define as attention scaling:

$$\tilde{A}(\Delta z) = \mathrm{Normalize}(A \odot M(\Delta z)), \tag{10}$$

where the scaling factor $M(\Delta z) = \mathrm{softmax}(\Delta z)$ and $\odot$ denotes element-wise multiplication. While mathematically equivalent to the logit perturbation, this formulation is more practical as it allows for direct manipulation of the computed attention distribution without accessing the logits. Detailed derivations are provided in Appendix A.2. Finally, substituting the perturbation direction $\Delta z^* \propto G$ from Proposition 3.2 into Equation (10) yields an explicit construction of the worst-case perturbed attention distribution $\tilde{A}^*$ via attention scaling:

$$\tilde{A}^* = \mathrm{Normalize}(A \odot M(\Delta z^*)) = \mathrm{Normalize}\left(A \odot M\left(\frac{G}{T_G}\right)\right). \tag{11}$$

where T_G is a temperature to control the scale of the perturbation magnitude. This provides a practical construction of the first-order worst-case attention perturbation for a given attention map.

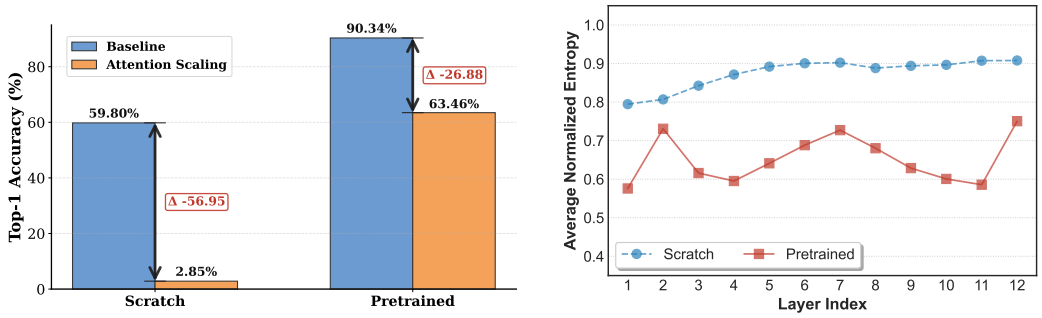


Figure 2: Attention sensitivity and entropy analysis on CIFAR100 with ViT-B. Pre-trained ViTs exhibit smaller accuracy changes under gradient-guided attention scaling and lower attention entropy, indicating sharper and less sensitive attention distributions.

3.3 Sensitivity Gap between Pre-trained and Scratch-trained ViTs

Figure 2 (left) shows that large-scale pre-trained ViTs are less affected by Equation (11) than scratch-trained ViTs. To relate this sensitivity gap to attention flatness, we use the normalized attention entropy:

$$H(A) = -\frac{1}{\log N} \sum_{j=1}^{N} A_j \log A_j. \tag{12}$$

A higher $H(A)$ indicates a more uniform distribution, while a lower value reflects a sharper focus. Consistent with previous works (Lee et al., 2021; Chen et al., 2023; Zhou et al., 2021), models trained from scratch on small datasets exhibit more uniform attention distributions than those pre-trained on large-scale datasets, as shown in Figure 2 (right).

3.4 Effect of Attention Sharpness on Local Sensitivity

To connect this empirical observation to the loss sensitivity characterized in Equation (8), we examine how the shape of the attention distribution affects the softmax Jacobian J.

Proposition 3.3 (Local Sensitivity of Sharp Attention)**.** *Let $J = \mathrm{diag}(A) - AA^\top$ be the Jacobian of the softmax map. As the attention distribution becomes sharper and approaches a one-hot distribution, the Jacobian satisfies $J \to 0$.*

Proof. The proof is provided in Appendix A.3. □

Proposition 3.3 shows that, for locally bounded G and Δz, the loss variation $G^\top J \Delta z$ vanishes as the attention distribution approaches a one-hot pattern. Thus, sharper attention reduces the local propagation of attention-logit perturbations to the loss. This provides an explanation for the observations in Figure 2, where sharper attention patterns are associated with reduced sensitivity to gradient-guided attention perturbations.

3.5 Relation to Temperature-based Attention Sharpening

The analysis above shows that sharper attention distributions can reduce local sensitivity by suppressing the softmax Jacobian. A common way to control attention sharpness is to modify the softmax temperature. Several studies (Lee et al., 2021; Zou et al., 2024; Veličković et al., 2025) adjust the temperature T to obtain sharper attention distributions, $A_T = \mathrm{softmax}(z/T)$. Although decreasing T can make the resulting attention distribution sharper, temperature scaling does not directly optimize robustness to attention-logit perturbations. The local Jacobian with respect to the original logits is

$$J_T = \frac{\partial A_T}{\partial z} = \frac{1}{T}\left(\mathrm{diag}(A_T) - A_T A_T^\top\right), \tag{13}$$

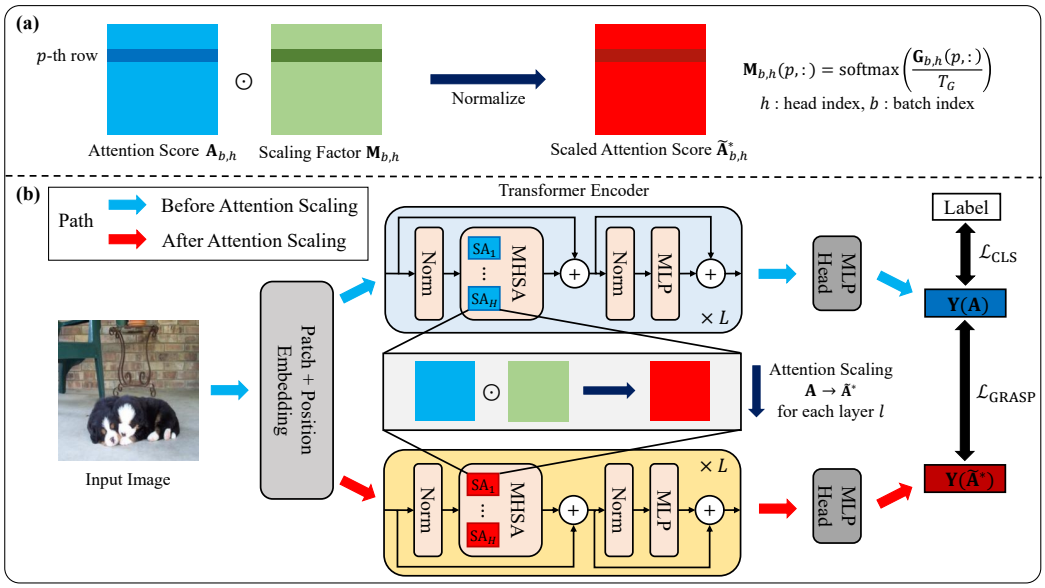


Figure 3: Illustration of the proposed techniques. (a) Gradient-guided attention scaling constructs a perturbed attention distribution by reweighting the original attention distribution with a gradient-based scaling factor followed by normalization. (b) During training, GRASP computes the original and perturbed outputs and minimizes a Jensen-Shannon divergence loss together with the standard classification objective.

where both the explicit factor $1/T$ and the attention distribution A_T vary with T. Therefore, lowering T has two competing effects: it may reduce the distribution-dependent Jacobian term by making A_T sharper, while also amplifying the overall sensitivity through the explicit $1/T$ factor. As a result, sharper attention induced by temperature scaling does not necessarily imply lower local sensitivity to logit perturbations.

4 GRASP: Gradient-Guided Robust Attention Sharpening

Following the sensitivity analysis in Section 3, we introduce Gradient-Guided Robust Attention Sharpening (GRASP) as a training-time regularizer. GRASP leverages the robustness of large-scale pre-trained ViTs to the gradient-guided attention perturbation formulated in Equation (11) during training. Figure 3 illustrates this procedure. Given the perturbed attention distribution $\tilde{A}^*$, the model computes the original output $Y(A)$ and the perturbed output $Y(\tilde{A}^*)$. We define the GRASP loss as the Jensen-Shannon divergence (JSD) (Hendrycks et al., 2019; Hwang et al., 2024):

$$\mathcal{L}_{\text{GRASP}} = \text{JSD}\left(Y(A) \,\|\, Y(\tilde{A}^*)\right). \tag{14}$$

The final training objective is a weighted sum of the cross-entropy loss $\mathcal{L}_{\text{CLS}}$ and the GRASP loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CLS}} + \lambda \mathcal{L}_{\text{GRASP}}, \tag{15}$$

where λ controls the strength of the GRASP regularization. The perturbed attention path is used only during training and therefore introduces no additional inference cost. By minimizing the prediction discrepancy caused by the gradient-guided attention perturbation, GRASP encourages the model to remain stable under attention-level perturbations. As shown in Appendix A.4, this consistency objective upper-bounds the corresponding cross-entropy loss deviation for the perturbations considered in our analysis. Unlike simple sharpening methods that directly reduce attention entropy, GRASP does not optimize attention entropy itself. Instead, sharper attention is implicitly induced by its sensitivity-aware regularization, as formalized below.

Proposition 4.1 (Implicit Sharpness Induction). *Under the local first-order approximation, GRASP encourages sharper attention distributions when the token-wise gradients are distinct and the scaling factor T_G is finite.*



Figure 4: Visualization of three Tiny-ImageNet categories: monarch butterfly, bowtie, and fur coat. GRASP yields sharper class-token attention on the target object.

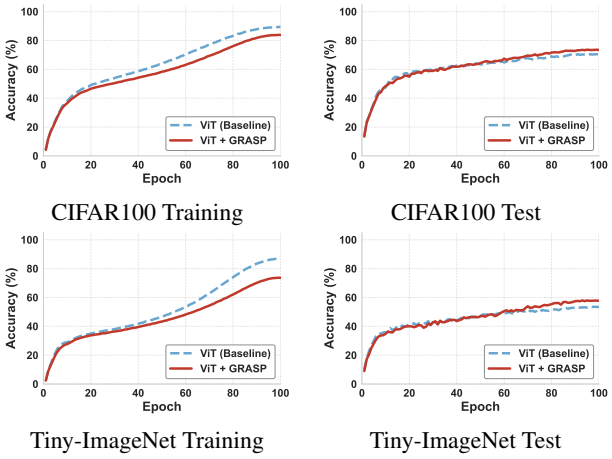


Figure 5: Comparison of training accuracy (left) and test accuracy (right) for CIFAR100 and Tiny-ImageNet with the ViT model. GRASP shows lower training accuracy but achieves higher test accuracy than the baseline, mitigating overfitting and improving generalization.

Proof. The proof is provided in Appendix A.5. □

5 Experiments

We evaluate GRASP across diverse training and evaluation settings. After describing the experimental setups, we present the main empirical results and analyze key design choices through ablation studies.

5.1 Training Scenarios and Datasets

To evaluate the effectiveness of GRASP, experiments are conducted across three scenarios. The two primary scenarios are: (1) **training on small datasets**, i.e., datasets smaller than ImageNet-1K (Russakovsky et al., 2015), including CIFAR-100 (Krizhevsky et al., 2009), Tiny-ImageNet (Le & Yang, 2015), Flowers102 (Nilsback & Zisserman, 2008), and the ClipArt, Sketch, and Painting subsets of DomainNet (Peng et al., 2019), where overfitting is commonly observed; and (2) **ImageNet-scale testing and out-of-distribution evaluation**, using ImageNet-1K to assess scalability and ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-V2 (Recht et al., 2019), and ImageNet-Stylized (Geirhos et al., 2018) to evaluate robustness under significant out-of-distribution (OOD) shifts. As an additional scenario, we also consider (3) **fine-tuning under domain shift** to adapt pre-trained ImageNet-1K models to specific target datasets. The detailed statistics, including the size, resolution, and number of classes for each dataset, are described in Table 1. Appendices B.1, B.2, and B.3 provide experimental setups for each scenario, including the baseline training recipe with standard augmentation and regularization.

Table 1: Summary of dataset statistics, including training and test sizes, image resolutions, and the number of classes.

Dataset	Train Size	Test Size	Resolution	Classes
CIFAR100	50,000	10,000	32×32	100
Tiny-ImageNet	100,000	10,000	64×64	200
Flowers102	2,040	6,149	224×224	102
ClipArt	33,525	14,604		
Sketch	48,212	20,916	224×224	345
Painting	50,416	21,850		
ImageNet-1K	1,281,167	50,000	224×224	1000

5.2 Main Results

Experimental results for the three scenarios are presented below.

Table 2: Top-1 Accuracy (%) on various datasets and ViTs when trained from scratch. By applying GRASP, we observe consistent performance improvements across different datasets and ViTs. RegNetX is included for comparison with CNN models.

Model	Params(M)	Method	CIFAR100	T-ImageNet	Flowers102	ClipArt	Sketch	Painting
ViT	2.7	Baseline	71.05±0.50	53.40±0.39	61.26±0.56	49.90±0.31	42.53±0.44	38.88±0.37
		+GRASP	**73.86**±0.39	**57.27**±0.59	**63.90**±0.41	**51.79**±0.28	**45.62**±0.39	**42.94**±0.34
PiT	7.0	Baseline	68.89±0.53	53.08±0.25	51.35±0.38	50.47±0.36	44.81±0.41	36.64±0.48
		+GRASP	**72.68**±0.52	**57.81**±0.18	**53.77**±0.62	**53.10**±0.29	**46.68**±0.33	**40.27**±0.42
Swin	7.1	Baseline	75.19±0.07	55.61±0.42	73.70±0.42	61.51±0.24	57.34±0.30	51.09±0.27
		+GRASP	**76.83**±0.25	**59.21**±0.34	**75.77**±0.61	**64.17**±0.22	**59.74**±0.26	**55.03**±0.25
RegNetX-200MF	2.4	-	57.97	47.88	68.47	62.54	56.32	52.38
RegNetX-400MF	4.8	-	60.89	50.11	70.39	64.50	58.79	55.04
RegNetX-800MF	6.7	-	64.90	52.69	76.63	66.69	60.70	57.38

5.2.1 Training on Small Datasets

Performance Comparison. Table 2 evaluates GRASP on ViT, PiT, and Swin, covering global attention, pooling-based hierarchy, and shifted-window attention, respectively. GRASP consistently improves all three architectures, with gains of 3.87 percentage points (%p) and 4.73%p for ViT and PiT on Tiny-ImageNet. As shown in Figure 5, the baseline ViT reaches high training accuracy but exhibits a large generalization gap, whereas GRASP achieves lower training but higher test accuracy, indicating reduced overfitting. Figure 4 further shows that GRASP produces sharper, more localized attention on target objects. Layer-wise analyses in Appendix F show that this sharpening is particularly evident in deeper layers. Together, these results support the association between reduced attention sensitivity and improved generalization on small datasets.

Compatibility with State-of-the-Art. Table 3 demonstrates that GRASP is highly complementary to diverse SOTA methods. This is evident in architectural modifications such as DWConv (Zhang et al., 2025), PAEViT (Verma et al., 2025) and ELSA (Zhou et al., 2023), local attention with temperature scaling such as SL-ViT (Lee et al., 2021), as well as auxiliary regularization like Drloc (Liu et al., 2021b), Trans (Huan et al., 2023) and SSAT (Das et al., 2024). For example, applying GRASP to ViT+Trans improves CIFAR100 accuracy by 2.59%p ($74.01 \rightarrow 76.60\%$), while boosting ELSA-Swin on Tiny-ImageNet by 2.84%p ($61.81 \rightarrow 64.65\%$). As illustrated in Figure 6, even models enhanced with these SOTA techniques still exhibit significant sensitivity. Together with the attention-entropy analysis in Appendix F, these results highlight that directly sharpening or modifying attention does not necessarily reduce attention sensitivity, whereas GRASP explicitly regularizes this sensitivity.

Table 3: Comparison of the proposed GRASP with state-of-the-art approaches.

Model	Method	CIFAR100	T-ImageNet	Flowers102
ViT	Baseline	70.51	53.55	61.54
	+ **GRASP**	**73.61**	**57.96**	**63.52**
	ViT+Drloc	71.32	53.88	65.83
	+ **GRASP**	**74.57**	**58.02**	**67.34**
	ViT+Trans	74.01	58.30	65.36
	+ **GRASP**	**76.60**	**61.32**	**68.12**
	ViT+SSAT	72.75	54.35	63.39
	+ **GRASP**	**75.17**	**59.22**	**65.34**
	SL-ViT	75.29	56.35	62.34
	+ **GRASP**	**76.10**	**59.26**	**64.42**
	ViT+MOHSA	71.63	53.43	60.77
	+ **GRASP**	**74.42**	**57.35**	**63.62**
	ViT+DWConv	74.98	57.06	64.87
	+ **GRASP**	**75.26**	**59.24**	**66.04**
	PAEViT	74.05	56.28	60.92
	+ **GRASP**	**75.39**	**59.48**	**64.89**
Swin	Baseline	75.29	55.08	73.12
	+ **GRASP**	**76.58**	**59.07**	**75.18**
	Swin+Drloc	75.52	55.94	77.20
	+ **GRASP**	**77.74**	**60.18**	**78.48**
	Swin+Trans	77.35	59.52	79.40
	+ **GRASP**	**78.61**	**61.61**	**80.58**
	SL-Swin	78.49	60.79	76.31
	+ **GRASP**	**79.74**	**63.32**	**77.65**
	ELSA-Swin	79.16	61.81	75.61
	+ **GRASP**	**80.42**	**64.65**	**79.31**
	Swin+MOHSA	74.72	56.08	74.09
	+ **GRASP**	**76.35**	**59.51**	**76.47**
	Swin+DWConv	78.15	60.14	80.64
	+ **GRASP**	**79.42**	**62.71**	**81.92**

5.2.2 ImageNet-Scale Testing and Out-of-Distribution Evaluation

As shown in Table 4, GRASP demonstrates scalability by improving clean accuracy on ImageNet-1K from 77.4% to 78.4%. The performance gains are even more pronounced in scenarios involving large distribution shifts. For instance, GRASP reduces the mean corruption error (mCE) on ImageNet-C by 1.5%p and improves accuracy on ImageNet-A and ImageNet-Stylized by 2.4%p and 2.9%p,

Table 4: ImageNet-1K performance of ViT-B/16 (86M) with and without GRASP. GRASP improves clean accuracy and robustness under distribution shifts.

Method	Clean (Accuracy ↑)	Corruption (mCE ↓)	Out-of-Distribution (Accuracy ↑)			
	ImageNet	**IN-C**	**IN-A**	**IN-R**	**IN-V2**	**IN-Stylized**
ViT-B	77.4	61.2	12.5	38.0	64.6	13.4
+GRASP	**78.4**	**59.7**	**14.9**	**39.2**	**65.9**	**16.3**

Table 5: ViT-S (22M) performance on multiple datasets with and without ImageNet-1K pre-training and GRASP.

Method	**CIFAR100**	**ClipArt**	**Sketch**
Scratch	65.25	46.74	39.49
+GRASP	**67.13**	**48.19**	**40.50**
IN-1K pre-trained	86.29	74.15	65.21
+GRASP	**87.07**	**75.73**	**66.50**

respectively. These improvements across diverse benchmarks demonstrate that GRASP provides robust generalization against a broad range of distribution shifts.

5.2.3 Fine-tuning with Domain-shift Datasets

Table 5 shows that GRASP provides consistent gains for both training from scratch and fine-tuning scenarios. For DomainNet datasets like ClipArt and Sketch, which exhibit large domain shifts from ImageNet, GRASP improves fine-tuning accuracy by 1.58%p ($74.15 \rightarrow 75.73\%$) and 1.29%p ($65.21 \rightarrow 66.50\%$). These improvements are comparable to the gains observed when training from scratch (+1.45%p and +1.01%p). In contrast, for CIFAR100, the gain during fine-tuning is 0.78%p ($86.29 \rightarrow 87.07\%$), which is smaller than the 1.88%p improvement seen when training from scratch. This indicates that GRASP remains beneficial after pre-training, with larger gains under substantial domain shift. Appendix G further analyzes the residual attention sensitivity and layer-wise entropy of pre-trained models.

5.3 Ablation Study

We first verify that the baseline saturates by 100 epochs and analyze λ and T_G. For a fixed gradient, a smaller T_G produces a sharper scaling factor and thus a stronger perturbation. Across ViT and Swin, GRASP remains robust to a range of hyperparameter choices, while careful tuning can further improve performance. The full epoch and hyperparameter ablations are provided in Appendix D.

We also compare two designs for the GRASP objective on CIFAR100 with ViT. Using the target-label cross-entropy on the perturbed-attention prediction, $\mathcal{L}_{\mathrm{CLS}}(Y(\tilde{A}^*), C)$, improves accuracy from 70.51% to 72.22%. In contrast, the JSD consistency objective, $\mathcal{L}_{\mathrm{JSD}}(Y(A), Y(\tilde{A}^*))$, achieves 73.61%. While $\mathcal{L}_{\mathrm{CLS}}$ primarily focuses on increasing the confidence of the target class, $\mathcal{L}_{\mathrm{JSD}}$ enforces distribution consistency across all classes, including those with low confidence. This comprehensive regularization leads to superior performance gains. Consequently, we adopt the JSD loss for GRASP as it effectively addresses sensitivity of the model regardless of the target class confidence.

6 Conclusion

This paper establishes a theoretical framework for analyzing input perturbations through their induced effects on the attention mechanism. By formulating shift-induced attention variation as bounded perturbations to attention logits and deriving the worst-case attention perturbation, we show empirically that models trained on small datasets exhibit substantially higher sensitivity than those pre-trained on large-scale datasets. Furthermore, a local sensitivity analysis shows that the softmax Jacobian vanishes as attention becomes sharp, thereby attenuating the first-order loss variation propagated through attention under shift-induced attention-logit perturbations. To bridge the empirical observations and this local sensitivity analysis, GRASP is proposed as an optimization framework that enforces prediction consistency under the derived worst-case attention perturbations. The analysis further indicates that, under cross-entropy training, robustifying models against such perturbations implicitly induces sharper attention distributions, thereby reducing sensitivity to shift-induced attention perturbations. Extensive experiments across diverse scenarios demonstrate that GRASP significantly improves robustness and generalization. Overall, this work provides a framework for analyzing attention sensitivity and a practical approach for improving ViT robustness and generalization across challenging environments.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, including improving grammar, wording, clarity, and readability. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ibrahim Batuhan Akkaya, Senthilkumar S Kathiresan, Elahe Arani, and Bahram Zonooz. Enhancing performance of vision transformers on small datasets through local inductive bias incorporation. *Pattern Recognition*, 153:110510, 2024.

Mujtaba Asad, Waqar Azeem, He Jiang, Hafiz Tayyab Mustafa, Jie Yang, and Wei Liu. 2m3df: Advancing 3d industrial defect detection with multi perspective multimodal fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Yu Bai, Liang Bai, Xian Yang, and Jiye Liang. Label-semantic-based prompt tuning for vision transformer adaptation in medical image analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3984–3992, 2023.

Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. *arXiv preprint arXiv:1605.06743*, 2016.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Srijan Das, Tanmay Jain, Dominick Reilly, Pranav Balaji, Soumyajit Karmakar, Shyam Marjit, Xiang Li, Abhijit Das, and Michael S Ryoo. Limited data, unlimited potential: A study on vits augmented by masked autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6878–6888, 2024.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

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. *ICLR*, 2021.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

Xiuwen Fang, Mang Ye, and Xiyuan Yang. Robust heterogeneous federated learning under data corruption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5020–5030, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021.

Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11936–11945, 2021.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8126–8135. IEEE, 2020.

Chen Huan, Wentao Wei, and Ping Yao. Train vit on small dataset with translation perceptibility. In *British Machine Vision Conference*, 2023.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 646–661. Springer, 2016.

Hyunha Hwang, Se-Hun Kim, Kyujoong Lee, and Hyuk-Jae Lee. Improving model robustness with frequency component modification and mixing. *IEEE Access*, 12:182171–182189, 2024.

Hyunha Hwang, Xuan Truong Nguyen, and Hyuk-Jae Lee. Ls-vit: Least-squares hessian based block reconstruction for low-bit post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33588–33597, 2026.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14285, 2020.

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

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *European Conference on Computer Vision*, pp. 110–127. Springer, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haozhe Liu, Haoqian Wu, Weicheng Xie, Feng Liu, and Linlin Shen. Group-wise inhibition based feature regularization for robust classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 478–486, 2021a.

Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34: 23818–23830, 2021b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European conference on computer vision*, pp. 53–69. Springer, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Minseok Seo, Hyuk-Jae Lee, and Xuan Truong Nguyen. Vit-p3de*: Vision transformer based multi-camera instance association with pseudo 3d position embeddings. In *IJCAI*, pp. 1340–1350, 2023.

Jincai Song, Haipeng Chen, Yingda Lyu, Weizhi Nie, and An-An Liu. Causality-inspired unsupervised domain adaptation with target style imitation for medical image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14454–14463, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *International conference on machine learning*. PMLR, 2025.

Akash Verma, Shiv Ram Dubey, and Satish Kumar Singh. Patch attention excitation based vision transformer for small-sized datasets. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

Zihao Wang and Lei Wu. Theoretical analysis of the inductive biases in deep convolutional networks. *Advances in Neural Information Processing Systems*, 36:74289–74338, 2023.

Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5028–5037, 2017.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22–31, 2021.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 819–828, 2020.

Qianzi Yu, Kai Zhu, Yang Cao, Feijie Xia, and Yu Kang. Tf^2: Few-shot text-free training-free defect image generation for industrial anomaly inspection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11825–11837, 2024.

Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 579–588, 2021a.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021b.

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):6575–6586, 2023.

Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6101–6112, 2023.

Tianxiao Zhang, Bo Luo, and Guanghui Wang. Improving vision transformers by overlapping heads in multi-head self-attention. *arXiv preprint arXiv:2410.14874*, 2024.

Tianxiao Zhang, Wenju Xu, Bo Luo, and Guanghui Wang. Depth-wise convolutions in vision transformers for efficient training on small datasets. *Neurocomputing*, 617:128998, 2025.

Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3417–3425, 2022.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Daquan Zhou, Yujun Shi, Bingyi Kang, Weihao Yu, Zihang Jiang, Yuan Li, Xiaojie Jin, Qibin Hou, and Jiashi Feng. Refiner: Refining self-attention for vision transformers. *arXiv preprint arXiv:2106.03714*, 2021.

Jingkai Zhou, Pichao Wang, Jiasheng Tang, Fan Wang, Qiong Liu, Hao Li, and Rong Jin. What limits the performance of local self-attention? *International Journal of Computer Vision*, 131(10): 2516–2528, 2023.

Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:116332–116354, 2024.

A Main Proofs

A.1 Proof of Proposition 3.2

Here, we derive the gradient-guided perturbation direction by maximizing the first-order loss variation under a fixed softmax-effective perturbation magnitude.

Step 1: Linearization of Scaling. Let $z \in \mathbb{R}^N$ be the input logits such that $A = \mathrm{softmax}(z)$. We define the perturbed attention $\tilde{A}$ by adding a perturbation vector $\Delta z \in \mathbb{R}^N$ to the logits. The i-th element of $\tilde{A}(\Delta z)$ is:

$$\tilde{A}(\Delta z)_i = \frac{\exp(z_i + \Delta z_i)}{\sum_{j=1}^{N} \exp(z_j + \Delta z_j)}. \tag{16}$$

By the property of the exponential function, $\exp(z_i + \Delta z_i) = \exp(z_i)\exp(\Delta z_i)$. Substituting this into Equation (16), we obtain:

$$\tilde{A}(\Delta z)_i = \frac{\exp(z_i)\exp(\Delta z_i)}{\sum_{j=1}^{N} \exp(z_j)\exp(\Delta z_j)}. \tag{17}$$

To express this in terms of the original attention A_i, we divide both the numerator and the denominator by the original normalization constant $Z = \sum_{k=1}^{N} \exp(z_k)$, noting that $A_i = \exp(z_i)/Z$:

$$\tilde{A}(\Delta z)_i = \frac{\frac{\exp(z_i)}{Z}\exp(\Delta z_i)}{\sum_{j=1}^{N} \frac{\exp(z_j)}{Z}\exp(\Delta z_j)} = \frac{A_i \exp(\Delta z_i)}{\sum_{j=1}^{N} A_j \exp(\Delta z_j)}. \tag{18}$$

We apply the first-order Taylor expansion $\exp(x) \approx 1 + x$ for small x:

$$\tilde{A}_i \approx \frac{A_i(1 + \Delta z_i)}{\sum_{j=1}^{N} A_j(1 + \Delta z_j)} = \frac{A_i(1 + \Delta z_i)}{1 + \mathbb{E}_A[\Delta z]}, \tag{19}$$

where $\mathbb{E}_A[\Delta z] = \sum_j A_j \Delta z_j$ and we use the fact that $\sum_j A_j = 1$.

Step 2: Approximation of Attention Variation. Using the approximation $(1 + x)^{-1} \approx 1 - x$, we expand the denominator:

$$\begin{aligned} \tilde{A}_i &\approx A_i(1 + \Delta z_i)(1 - \mathbb{E}_A[\Delta z]) \\ &\approx A_i + A_i(\Delta z_i - \mathbb{E}_A[\Delta z]) \quad (\text{ignoring } \mathcal{O}(\|\Delta z\|^2)). \end{aligned} \tag{20}$$

Thus, the variation in attention $\Delta A_i = \tilde{A}_i - A_i$ is:

$$\Delta A_i \approx A_i(\Delta z_i - \mathbb{E}_A[\Delta z]). \tag{21}$$

Step 3: Loss Variation and Covariance. The change in perturbation loss is approximated by the dot product with the gradient $G = \nabla_A \mathcal{L}$:

$$\begin{aligned} \Delta\mathcal{L}(\Delta z) &\approx \sum_{i=1}^{N} G_i \Delta A_i \\ &= \sum_{i=1}^{N} G_i \left[A_i(\Delta z_i - \mathbb{E}_A[\Delta z])\right] \\ &= \mathbb{E}_A[G \cdot \Delta z] - \mathbb{E}_A[\Delta z]\mathbb{E}_A[G] \\ &= \mathrm{Cov}_A(G, \Delta z). \end{aligned} \tag{22}$$

Step 4: Optimality under Fixed Effective Magnitude. From Equation (22), we can rewrite the first-order loss variation as

$$\Delta\mathcal{L}(\Delta z) = \mathrm{Cov}_A(G, \Delta z) = \mathbb{E}_A[(G - \mu_G)(\Delta z - \mu_{\Delta z})], \tag{23}$$

where $\mu_G = \mathbb{E}_A[G]$ and $\mu_{\Delta z} = \mathbb{E}_A[\Delta z]$.

Since the softmax function is invariant to additive constant shifts in logits, the constant component of Δz does not affect the perturbed attention distribution. Thus, the effective magnitude of a logit perturbation is naturally measured by its centered variation,

$$\mathrm{Var}_A(\Delta z) = \mathbb{E}_A\left[(\Delta z - \mathbb{E}_A[\Delta z])^2\right]. \tag{24}$$

This effective magnitude is controlled by the ℓ_q perturbation budget:

$$\mathrm{Var}_A(\Delta z) \le \mathbb{E}_A[\Delta z^2] \le \|\Delta z\|_\infty^2 \le \|\Delta z\|_q^2. \tag{25}$$

Therefore, if $\|\Delta z\|_q \le \epsilon$, then

$$\sqrt{\mathrm{Var}_A(\Delta z)} \le \epsilon. \tag{26}$$

We now compare perturbation directions with the same effective magnitude. By the Cauchy–Schwarz inequality,

$$|\Delta\mathcal{L}(\Delta z)| \le \sqrt{\mathrm{Var}_A(G)}\sqrt{\mathrm{Var}_A(\Delta z)}. \tag{27}$$

For a fixed value of $\mathrm{Var}_A(\Delta z)$, the loss-increasing worst-case direction attains this bound when

$$\Delta z - \mu_{\Delta z}\mathbf{1} = k\,(G - \mu_G\mathbf{1}), \qquad k > 0. \tag{28}$$

Equivalently,

$$\Delta z = kG + c\mathbf{1}, \tag{29}$$

where $c = \mu_{\Delta z} - k\mu_G$ is a constant. Since the additive constant $c\mathbf{1}$ does not change the softmax output, we may choose the representative with $c = 0$, yielding

$$\Delta z^* \propto G. \tag{30}$$

A.2 Derivation of Equation (10)

Proof. Let $M = \mathrm{softmax}(\Delta z)$ be the normalized perturbation vector, where $M_i = \exp(\Delta z_i)/Z_{\Delta z}$ and $Z_{\Delta z} = \sum_k \exp(\Delta z_k)$. By substituting $\exp(\Delta z_i) = M_i Z_{\Delta z}$ into Equation (18), the constant $Z_{\Delta z}$ cancels out:

$$\tilde{A}(\Delta z)_i = \frac{A_i(M_i Z_{\Delta z})}{\sum_{j=1}^N A_j(M_j Z_{\Delta z})} = \frac{A_i M_i}{\sum_{j=1}^N A_j M_j}. \tag{31}$$

This demonstrates that an additive perturbation Δz in the logit space is mathematically equivalent to a multiplicative re-weighting of the attention scores by a normalized factor $M = \mathrm{softmax}(\Delta z)$, followed by re-normalization: $\tilde{A} = \mathrm{Normalize}(A \odot M)$.

□

A.3 Proof of Proposition 3.3

Proof. For $A = \mathrm{softmax}(z)$, the local sensitivity Jacobian is

$$J = \mathrm{diag}(A) - AA^\top. \tag{32}$$

For any vector $v \in \mathbb{R}^N$, we have

$$v^\top J v = \sum_i A_i v_i^2 - \left(\sum_i A_i v_i\right)^2 = \mathrm{Var}_A(v) \ge 0. \tag{33}$$

Thus, J is positive semidefinite and measures the local variance of logit directions under the attention distribution.

If the attention is one-hot, i.e., $A = e_k$ for some k, then

$$\mathrm{diag}(A) = e_k e_k^\top \qquad \text{and} \qquad AA^\top = e_k e_k^\top. \tag{34}$$

Therefore,

$$J = \mathrm{diag}(A) - AA^\top = 0. \tag{35}$$

Hence, for any infinitesimal logit perturbation Δz, the first-order attention variation satisfies

$$\Delta A \approx J\Delta z = 0. \tag{36}$$

More generally, if the attention distribution approaches a one-hot distribution, $A \to e_k$, then by continuity of

$$J(A) = \mathrm{diag}(A) - AA^\top, \tag{37}$$

we have

$$J(A) \to J(e_k) = 0. \tag{38}$$

Therefore, for locally bounded G and Δz, $G^\top J\Delta z \to 0$ as the attention distribution approaches a one-hot distribution. □

A.4 Bounding Local Loss Deviation with GRASP

Let $p = Y(A)$ and $q = Y(\tilde{A}^*)$ denote the predictions before and after the worst-case attention perturbation, respectively. We connect the JSD consistency loss used in GRASP to the local cross-entropy loss deviation. For the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}(p) = -\log p_y$, assume that the target-class probabilities remain locally bounded away from zero, i.e., $p_y, q_y \geq m > 0$. This is a local regularity condition aligned with cross-entropy training, since $\mathcal{L}_{\mathrm{CE}}(p)$ and $|\partial\mathcal{L}_{\mathrm{CE}}/\partial p_y|$ become unbounded as $p_y \to 0^+$. Since $-\log t$ is $1/m$-Lipschitz on $[m, 1]$, we have

$$\Delta\mathcal{L}(\Delta z^*) = |\mathcal{L}_{\mathrm{CE}}(q) - \mathcal{L}_{\mathrm{CE}}(p)| = |-\log q_y + \log p_y| \leq \frac{1}{m}|q_y - p_y| \leq \frac{1}{m}\|q - p\|_1. \tag{39}$$

By Pinsker's inequality and the definition of the Jensen–Shannon divergence,

$$\|q - p\|_1 \leq \sqrt{8\,\mathrm{JSD}(p\|q)}. \tag{40}$$

Therefore,

$$\Delta\mathcal{L}(\Delta z^*) = |\mathcal{L}_{\mathrm{CE}}(q) - \mathcal{L}_{\mathrm{CE}}(p)| \leq \frac{\sqrt{8}}{m}\sqrt{\mathrm{JSD}(p\|q)} = \frac{\sqrt{8}}{m}\sqrt{\mathcal{L}_{\mathrm{GRASP}}}. \tag{41}$$

Thus, minimizing $\mathcal{L}_{\mathrm{GRASP}}$ reduces an upper bound on the local cross-entropy loss deviation induced by the worst-case attention perturbation.

A.5 Proof of Proposition 4.1

Proof. Substituting the perturbation $\Delta z^* = G/T_G$ into the covariance formula Equation (22) yields:

$$\Delta\mathcal{L}(\Delta z^*) \approx \mathrm{Cov}_A\left(G, \frac{G}{T_G}\right) = \frac{1}{T_G}\,\mathrm{Var}_A(G). \tag{42}$$

Using the pairwise form of the weighted variance,

$$\mathrm{Var}_A(G) = \frac{1}{2}\sum_{i,j} A_i A_j (G_i - G_j)^2. \tag{43}$$

Let

$$\delta_G = \min_{i \neq j} |G_i - G_j|. \tag{44}$$

When the token-wise gradients are pairwise distinct, $\delta_G > 0$. Therefore,

$$\Delta\mathcal{L}(\Delta z^*) \approx \frac{1}{2T_G}\sum_{i,j} A_i A_j (G_i - G_j)^2 \tag{45}$$

$$\geq \frac{\delta_G^2}{2T_G}\sum_{i \neq j} A_i A_j \tag{46}$$

$$= \frac{\delta_G^2}{2T_G}\left(1 - \sum_i A_i^2\right) \tag{47}$$

$$\geq \frac{\delta_G^2}{2T_G}\left(1 - \max_i A_i\right). \tag{48}$$

Thus, for finite T_G, reducing the perturbation-induced loss variation encourages the attention distribution to become more concentrated, with $\sum_i A_i^2$ and $\max_i A_i$ approaching one. This corresponds to a sharper attention distribution.

The strength of this implication depends on the separation between token-wise gradients. When the gradients become similar, δ_G becomes small and the concentration bound becomes weaker. Moreover, $G = \nabla_A \mathcal{L}(A)$ denotes the current gradient at each optimization step. After each parameter update, both A and G may change, and the same local argument applies to their updated values.

Finally, the finite-T_G condition excludes the trivial case $T_G \to \infty$, for which $\Delta z^* \to 0$ and the perturbation vanishes. Hence, under the local first-order approximation and pairwise-distinct token-wise gradients, GRASP induces a local pressure toward sharper attention distributions. □

B Detailed Experimental Setup

B.1 Experimental Setup for Training on Small Datasets

Following prior work (Lee et al., 2021; Huan et al., 2023), three models, ViT (Dosovitskiy et al., 2021), PiT (Heo et al., 2021) and Swin (Liu et al., 2021c) are utilized for training on a small dataset. For images with a small resolution (e.g., 32×32 and 64×64), the patch sizes and the window sizes of Swin are adjusted based on higher-resolution settings to ensure compatibility (See Table 6 for more details). Data augmentation techniques such as Auto Augment (Cubuk et al., 2019), Repeated Augment (Hoffer et al., 2020), and random erasing (Zhong et al., 2020), as well as regularization methods like label smoothing (Szegedy et al., 2016) and stochastic depth (Huang et al., 2016), are utilized together. We utilize AdamW (Loshchilov & Hutter, 2017) as the optimizer. The models are trained for 100 epochs with a batch size of 128 on a single A100 GPU. The learning rate is set to 0.003 for ViT and CNNs, while it is set to 0.001 for the other models. For compatibility with SOTA methods, we integrate GRASP by either replacing $\mathcal{L}_{\text{CLS}}$ with the total loss (Liu et al., 2021b; Huan et al., 2023; Das et al., 2024) or adopting architectural modifications (Lee et al., 2021; Zhang et al., 2024; 2025; Zhou et al., 2023; Verma et al., 2025). Specifically, we evaluate Drloc (Liu et al., 2021b), Trans (Huan et al., 2023), SL-ViT (Lee et al., 2021), MOHSA (Zhang et al., 2024), and DWConv (Zhang et al., 2025) on both ViT and Swin backbones. Due to code availability or architecture-specific designs, SSAT (Das et al., 2024) and PAEViT (Verma et al., 2025) are compared solely on ViT, while ELSA (Zhou et al., 2023) is evaluated exclusively on the Swin architecture. λ is set to 12 based on previous studies (Hendrycks et al., 2019; Hwang et al., 2024), and T_G is set to 1 as the default value. The analysis of these hyperparameters is in Section D. Specific model settings for these main experiments are summarized in Table 6.

Table 6: Comparison of ViT, PiT, and Swin model configurations

Model	Resolution	Depth	Patch-size	Token Dimension	Heads	MLP-ratio	Window-size
ViT	32×32 64×64 224×224	9	4×4 8×8 16×16	192	12	2	-
PiT	32×32 64×64 224×224	[2,4,6]	2×2 4×4 16×16	96	[2,4,8]	2	-
Swin	32×32 64×64 224×224	[2,6,4]	2×2 4×4 4×4	96	[3,6,4]	2	4×4 4×4 7×7

B.2 Experimental Setup for ImageNet-Scale Testing and Out-of-Distribution Evaluation

We validate the scalability of the proposed method by conducting experiments on ImageNet-1K with a larger model such as ViT-B (Dosovitskiy et al., 2021). We apply a learning rate of 1.25e-4 to ensure stable convergence for this larger model (Steiner et al., 2021; Liu et al., 2021b). We also train the model for 300 epochs, referring to previous studies (Touvron et al., 2021; Heo et al., 2021; Liu et al., 2021c). To further validate that the proposed method improves generalization, we evaluate its

performance on various test sets. This includes ImageNet-C (Hendrycks & Dietterich, 2019), which introduces natural corruption to the original dataset, as well as out-of-distribution datasets such as ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-V2 (Recht et al., 2019), and ImageNet-Stylized (Geirhos et al., 2018), to assess robustness.

B.3 Experimental Setup for Fine-tuning with Domain-shift Datasets

The fine-tuning experiments are conducted with ViT-S (Touvron et al., 2021), where ImageNet-1K pre-training is performed with a learning rate of 1.25e-4, and fine-tuning is done with a learning rate of 1e-4. To ensure consistency with ImageNet-1K, all datasets are resized to 224×224.

C Additional Experiments on Downstream Tasks

Detection and segmentation generally use a pre-trained backbone through a classification dataset, and this study addresses cases with limited data. We also conduct experiments to investigate the effect of fine-tuning with GRASP on downstream tasks such as object detection and instance segmentation.

C.1 Experimental Setup

Following the experimental setup of DRLoc (Liu et al., 2021b), which adheres to the experimental environment of Swin (Liu et al., 2021c), we pre-train the backbone using a small dataset, ImageNet-100 (Tian et al., 2020), for 300 epochs. ImageNet-100 consists of 126,689 training images, 5,000 test images, and 100 classes. We use the Swin-T backbone and compare the performance with and without the application of GRASP during the fine-tuning process. We utilize the Cascade Mask R-CNN (Cai & Vasconcelos, 2018) and Mask R-CNN (He et al., 2017) architectures, incorporating multi-scale training (Carion et al., 2020; Sun et al., 2021). Specifically, we resize the input image such that the shortest side is between 480 and 800 pixels, while ensuring the longest side does not exceed 1333 pixels. The AdamW optimizer (Loshchilov & Hutter, 2017) is used with an initial learning rate of 0.0001, a weight decay of 0.05, and a batch size of 16. We follow a 1x schedule, training for 12 epochs with the learning rate decaying by a factor of 0.1 at epochs 8 and 11. All configurations are based on the standard MMDetection setup (Chen et al., 2019). The COCO 2017 dataset, with 118K training images, 5K validation images, and 20K test-dev images, is used for these experiments (Lin et al., 2014).

Table 7: Results on the COCO object detection and instance segmentation tasks. Models are pre-trained on ImageNet-100 and fine-tuned with/without GRASP.

backbone	**Architecture**	**Method**	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Swin-T	Mask RCNN	Baseline	41.8	60.3	45.1	36.7	57.4	39.4
		+$\mathcal{L}_{\text{GRASP}}$	**42.8**	**61.5**	**46.0**	**37.4**	**58.5**	**40.2**
	Cascade Mask RCNN	Baseline	36.0	58.2	38.6	33.8	55.2	35.9
		+ $\mathcal{L}_{\text{GRASP}}$	**37.5**	**59.8**	**40.5**	**34.8**	**56.4**	**37.0**

C.2 Results

Table 7 shows that fine-tuning with GRASP improves performance on downstream tasks, such as object detection and instance segmentation.

D Additional Ablation Studies

We first examine whether longer training improves the baseline. As shown in Table 8, baseline performance is largely saturated at 100 epochs, with only marginal gains from extending training to 300 epochs. Thus, all methods are trained for 100 epochs in our main experiments.

Table 8: Epoch ablation on Tiny-ImageNet.

Method	Epoch	ViT (%)	PiT (%)	Swin (%)
Baseline	100	53.55	53.20	55.08
Baseline	300	54.18	53.66	55.12
GRASP	100	57.96	58.08	59.07

Tables 9 and 10 report the sensitivity of GRASP to λ and T_G on CIFAR100. Overall, GRASP remains robust across a range of hyperparameter choices, while additional tuning can further improve performance.

Table 9: Hyperparameters λ and T_G for ViT on CIFAR100. **Bold** indicates the best accuracy, and underline refers to the hyperparameters used in the experiments.

T_G \ λ	4	8	12	16
0.5	73.44	73.41	73.35	73.32
1	73.02	73.87	<u>73.61</u>	73.33
1.5	73.92	73.87	**74.26**	73.70
2	73.88	73.41	74.08	73.20

Table 10: Hyperparameters λ and T_G for Swin on CIFAR100. **Bold** indicates the best accuracy, and underline refers to the hyperparameters used in the experiments.

T_G \ λ	4	8	12	16	20
0.5	76.19	76.47	76.28	76.79	77.10
1	76.32	76.67	<u>76.58</u>	76.97	**77.48**
1.5	76.10	76.93	77.08	76.72	77.08
2	76.54	76.45	77.19	76.95	77.40

E Training Cost Analysis

GRASP requires additional computation during training compared to the baseline as shown in Table 11. While this overhead is a limitation, the additional cost is comparable to other popular gradient-based optimization methods. For example, SAM (Foret et al., 2021), a widely adopted optimization method known for its effectiveness in enhancing generalization, takes 2.10h to train for CIFAR100. GRASP, which similarly leverages gradient information to regularize attention sensitivity, requires a comparable 2.25h as shown in Table 12. It is also more efficient than the self-supervised method Trans (Huan et al., 2023), which takes 4.27h due to the computation of its auxiliary loss. Furthermore, unlike standard fine-tuning or transfer learning methods that often necessitate upscaling CIFAR images to ImageNet-level resolution (e.g., 224×224) for architectural compatibility, GRASP trains directly on the original resolution (e.g., 32×32). This makes the 2.25-hour training cost relatively small compared to the massive resource requirements of high-resolution fine-tuning or full-scale pre-training. Importantly, these costs occur only during the training process and are not present during inference. Future work will focus on further reducing this training overhead while maintaining the robustness gains.

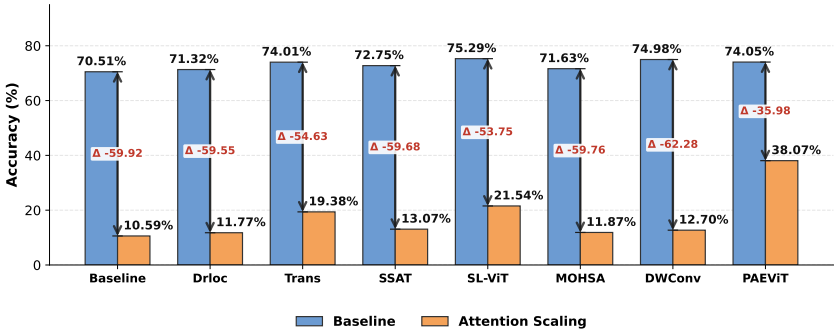
Table 11: Comparison of training time on CIFAR100 using ViT.

Method	**Training Time (h)**
Baseline	1.06
Drloc	1.22
Trans	4.27
SSAT	1.58
SL-ViT	1.23
MOHSA	1.71
DWConv	1.17
PAEViT	1.13
GRASP	2.25

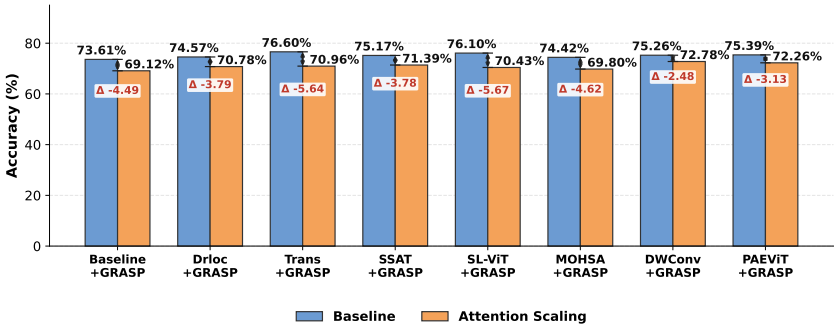
Table 12: Accuracy (%) and training time (h) comparison with SAM.

Method	**ViT**	**PiT**	**Swin**	**Time (h)**
Baseline	70.51	69.55	75.29	1.06
SAM	70.73	70.97	75.46	2.10
GRASP	73.61	72.68	76.58	2.25

F Extended Comparison with State-of-the-Art Baselines



(a) Results before applying GRASP



(b) Results after applying GRASP

Figure 6: Accuracy changes under gradient-guided attention scaling on CIFAR100 with ViT. Integrating GRASP into the state-of-the-art methods reduces sensitivity and enhances accuracy.

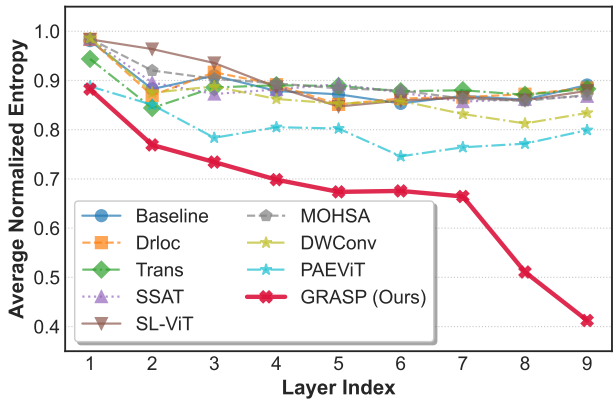


Figure 7: Comparison of attention entropy between GRASP and various state-of-the-art methods. GRASP induces sharper attention patterns, achieving lower normalized entropy across layers compared to existing methods.

We further investigate the impact of various SOTA methods and GRASP on attention characteristics. Figure 7 shows that GRASP achieves the lowest normalized attention entropy across layers. In particular, GRASP enforces sharper attention patterns in deeper layers, while existing baselines maintain relatively high entropy. Together with the reduced accuracy degradation under gradient-guided attention scaling shown in Figure 6, these results indicate that GRASP stabilizes the model while inducing sharper attention patterns. As a result, while existing SOTA methods suffer from

Table 13: Comparison with direct attention sharpening on CIFAR100.

Model	Baseline ($T = 1$)	$T = 2$	$T = 3$	$T = 4$	$T = 1/2$	$T = 1/3$	$T = 1/4$	GRASP
ViT	71.05	70.40	70.52	70.06	70.67	71.38	71.52	**73.86**
PiT	68.89	69.01	69.35	69.60	69.02	68.29	67.77	**72.68**
Swin	75.19	74.32	74.38	73.78	75.02	75.24	75.21	**76.83**

Table 14: Comparison with input-level adversarial training on CIFAR100.

Model	Baseline	$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 4/255$	GRASP
ViT	71.05	70.73	68.22	62.86	**73.86**
PiT	68.89	66.45	64.28	60.03	**72.68**
Swin	75.19	70.82	66.16	59.96	**76.83**

significant accuracy degradation under attention scaling (Figure 6a), the integration of GRASP effectively stabilizes performance (Figure 6b).

F.1 Comparison with Direct Attention Sharpening

Table 13 compares GRASP with direct softmax temperature scaling on CIFAR100. Lowering the temperature does not yield consistent improvements across architectures. For ViT, aggressive sharpening provides only a modest gain from 71.05% to 71.52%, while for PiT it degrades accuracy from 68.89% to 67.77% at $T = 1/4$. Swin also shows little benefit from direct sharpening. In contrast, GRASP consistently improves all three architectures, achieving 73.86%, 72.68%, and 76.83% for ViT, PiT, and Swin, respectively. These results support the analysis in Section 3.5: attention sharpening alone is insufficient to explain the gains of GRASP.

F.2 Comparison with Input-Level Adversarial Training

Table 14 compares GRASP with FGSM-based adversarial training on CIFAR100. Input-level adversarial training consistently degrades clean accuracy as the perturbation magnitude increases across all three architectures. For example, Swin accuracy decreases from 75.19% to 70.82%, 66.16%, and 59.96% as ϵ increases from $1/255$ to $4/255$. In contrast, GRASP improves the baseline accuracy to 76.83%. Similar trends are observed for ViT and PiT, indicating that directly regularizing attention-level perturbations is more effective for improving generalization in our setting than applying adversarial perturbations at the input level.

G Extended Comparison with ImageNet-1K pre-training

Figures 8 and 9 show the effect of ImageNet-1K pre-training and GRASP on CIFAR100 with the ViT-S model. Figure 9 shows that GRASP sharpens attention regardless of the training regime, reducing normalized attention entropy for both scratch-trained and pre-trained models. Combined with the sensitivity comparison in Figure 8, these results suggest that GRASP improves stability while consistently inducing sharper attention patterns.

H Evaluating Generalization

The results in Figures 5, 11 and 12 present the training and test accuracies for the ViT, PiT, and Swin models on the CIFAR100 and Tiny-ImageNet datasets across epochs. GRASP effectively prevents overfitting and enhances the generalization of these models. However, GRASP has a smaller effect on Swin than on ViT or PiT. As shown in Figure 10, Swin is structurally less sensitive than ViT. This structural difference is consistent with the smaller gains of GRASP observed for Swin.

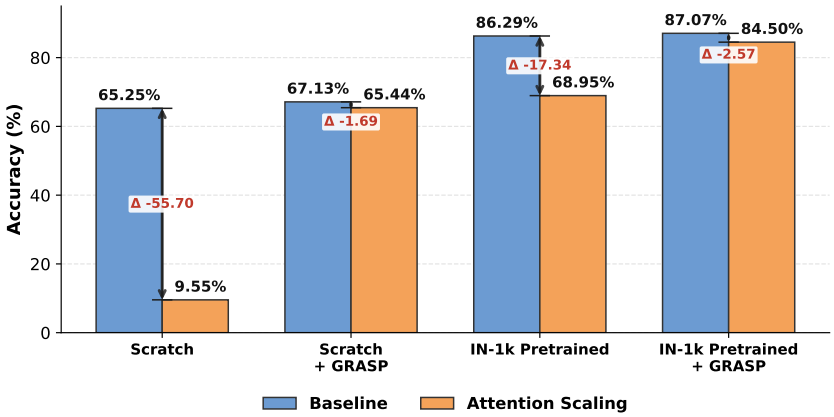


Figure 8: Sensitivity comparison of CIFAR100 with and without ImageNet-1K pre-training and GRASP.

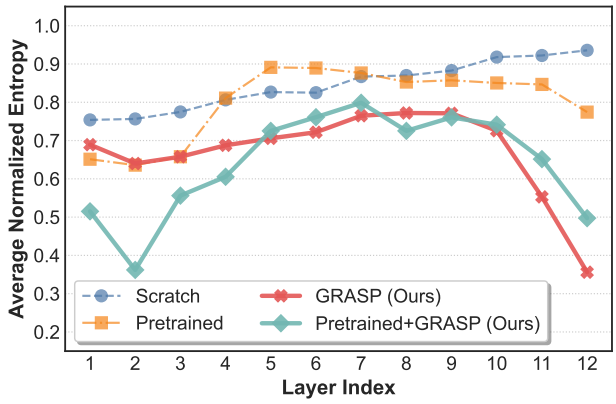


Figure 9: Impact of GRASP on attention entropy across different training regimes (scratch vs. pre-trained). GRASP consistently induces sharper attention distributions when applied to both scratch-trained and pre-trained baselines, particularly in deeper layers.

I LIMITATIONS

This paper focuses on attention-logit sensitivity as a factor underlying the robustness of vision transformers. This scope does not cover all sources of robustness degradation under distribution shift. Our local sensitivity analysis does not establish a formal generalization bound; rather, the relationship between reduced attention sensitivity and improved generalization is supported by our empirical evaluations. The perturbation direction in GRASP is derived from a local sensitivity analysis, following prior work that links local sensitivity to robustness under natural distribution shifts. The goal of GRASP is to reduce attention sensitivity, not to maximize attention sharpness. While the induced sharpening is effective in our setting, it may not be universally desirable across tasks. As provided in Appendix E, GRASP introduces additional computation during training because it requires gradient-guided attention perturbations, but this overhead is comparable to other gradient-based regularization methods such as SAM (Foret et al., 2021) and does not affect inference.

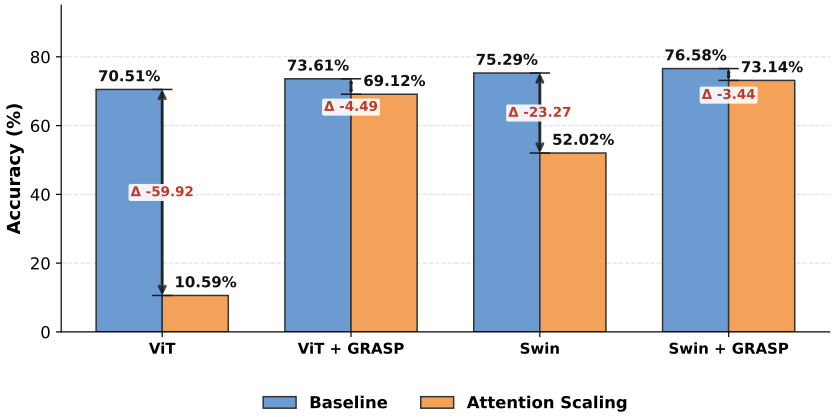
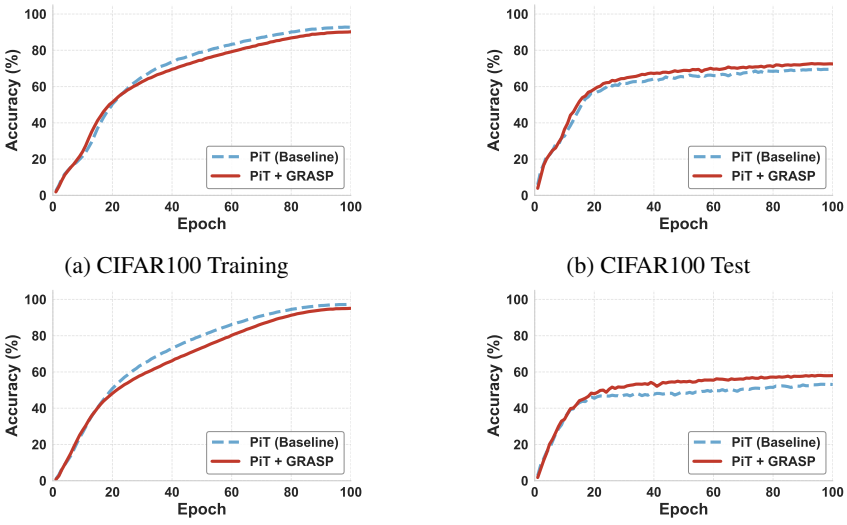


Figure 10: Sensitivity comparison of ViT and Swin on CIFAR100.



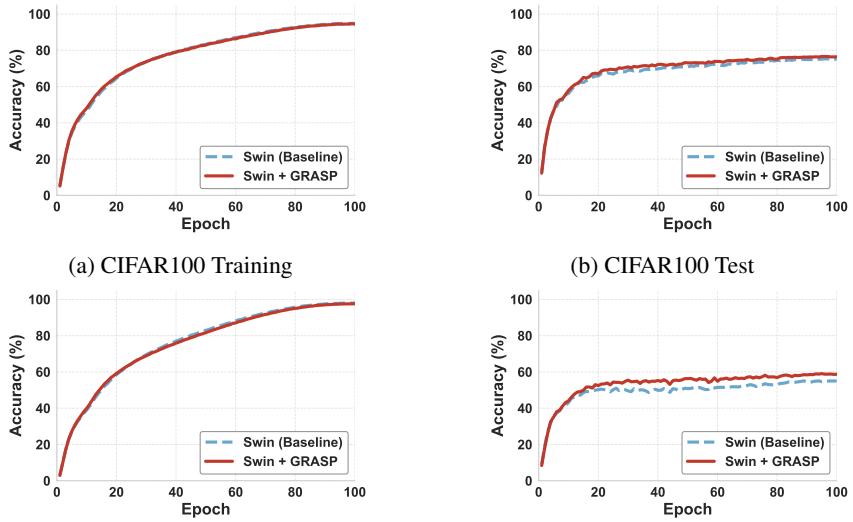
(a) CIFAR100 Training

(b) CIFAR100 Test

(c) Tiny-ImageNet Training

(d) Tiny-ImageNet Test

Figure 11: Comparison of training accuracy (left) and test accuracy (right) for CIFAR100 and Tiny-ImageNet with the PiT model. The proposed GRASP shows a lower training accuracy but achieves higher test accuracy compared to baseline, effectively mitigating overfitting and improving generalization.



(a) CIFAR100 Training

(b) CIFAR100 Test

(c) Tiny-ImageNet Training

(d) Tiny-ImageNet Test

Figure 12: Comparison of training accuracy (left) and test accuracy (right) for CIFAR100 and Tiny-ImageNet with the Swin model. The proposed GRASP shows a lower training accuracy but achieves higher test accuracy compared to baseline, effectively mitigating overfitting and improving generalization.