MDPI

Article

Attention Structural Change During CLIP Fine-Tuning: A Comparative Study of Full Fine-Tuning, LoRA, and Regularization

Ruize Xia [1]

[1] Nanjing Foreign Language School, Nanjing, Jiangsu, China; xiaruize0911@gmail.com

Abstract

Fine-tuning contrastive vision–language models such as CLIP is an essential step for downstream task specialization, yet the associated changes in the internal attention geometry remain insufficiently characterized. This study investigates the CLIP ViT-B/32 visual encoder across 21 experimental configurations on EuroSAT and Oxford-IIIT Pets, comparing full fine-tuning, low-rank adaptation (LoRA), and explicit attention-based regularization. We employ a multi-faceted metric suite—including CLS-to-patch attention entropy, effective receptive field (ERF), Gini concentration, and head diversity—complemented by attention rollout, CKA representational analysis, and subset-sensitivity tests. Our results demonstrate that full fine-tuning generally induces a contraction of attention support, whereas LoRA configurations maintain or even broaden the spatial distribution relative to the pretrained baseline. Statistical analysis via exact permutation tests identifies the learning rate as a critical determinant of these structural shifts. Furthermore, zero-shot reevaluation confirms that while LoRA incurs minor transfer degradation, it remains significantly more conservative than full fine-tuning. These findings provide a new structural lens for understanding the dynamics of transformer adaptation and suggest that structural preservation is a key mechanism for maintaining the generalizability of foundation models.

Keywords: CLIP; vision–language models; foundation models; transformer adaptation; LoRA; attention analysis; structural collapse; computer vision

1. Introduction

Contrastive vision–language pretraining has established CLIP as a foundational architecture for cross-modal retrieval and zero-shot visual recognition [1,2]. Despite its impressive zero-shot capabilities, practical deployment often requires downstream adaptation to achieve competitive performance on specialized domains. However, recent evidence suggests that such fine-tuning can distort the well-calibrated pretrained feature space, leading to significant performance drops in out-of-distribution (OOD) scenarios [3,4]. While the functional impact of this distortion is well-documented, the internal structural evolution of the self-attention mechanism during adaptation—specifically for CLIP—remains a subject of active research.

Self-attention is the central mechanism for long-range dependency modeling in Vision Transformers (ViT) [5,6]. Prior work has shown that pretrained ViTs develop specialized attention patterns, such as global context aggregation and local feature extraction, which are critical for robust representation learning [7,8]. Understanding whether these patterns are preserved or overwritten during fine-tuning is vital for developing safer and more

Received:
Accepted:
Published:

interpretable adaptation strategies. In this paper, we focus on the concept of *attention structural change*, exploring how different optimization regimes reshape the spatial allocation of attention mass.

We adopt a descriptive framework for attention analysis, treating measurements as indicators of model state rather than direct causal drivers of inference [9,10]. This distinction is necessary because the mapping from attention weights to model decisions is complex and often non-linear [11]. By focusing on structural descriptors—such as entropy, Gini concentration, and effective receptive field—we quantify the "plasticity" of the CLIP visual encoder under different constraints.

The study addresses four primary research questions: (i) the extent to which full fine-tuning alters the spatial concentration of attention; (ii) how parameter-efficient methods like LoRA differ from full fine-tuning in their structural impact; (iii) the role of hyperparameter selection, particularly the learning rate, in driving these changes; and (iv) the feasibility of mitigating structural drift through explicit regularization. Our contributions include:

- an empirical characterization of structural evolution in CLIP under varying adaptation regimes;
- a comprehensive evaluation of LoRA's structural conservatism and its implications for transfer learning;
- the identification of learning rate as a dominant control variable for attention concentration; and
- a systematic evaluation of various attention-based regularization techniques.

2. Related Work

The landscape of large-scale vision–language pretraining and downstream adaptation has evolved rapidly, centered on the dual challenges of performance and robustness. Therefore, we review the relevant literature across four interconnected domains: Vision Transformers (ViT), contrastive vision–language models (CLIP), fine-tuning dynamics and feature distortion, and parameter-efficient fine-tuning methods such as LoRA.

2.1. Vision Transformers (ViT)

The Vision Transformer (ViT) architecture [6] shifted the computer vision paradigm by applying the self-attention mechanism [5] to image patches. Unlike Convolutional Neural Networks (CNNs), ViTs lack inductive biases like translation invariance, relying instead on massive datasets to learn spatial hierarchies. Studies on ViT internal representations reveal that they develop distinct "attention heads" for global context and local feature extraction [7], and exhibit emergent properties such as object segmentation when trained with self-supervision [8].

2.2. Contrastive Vision–Language Models (CLIP)

CLIP (Contrastive Language–Image Pretraining) [1] scales ViT-based visual encoders by aligning them with text encoders using a contrastive loss on millions of image-text pairs. This alignment enables zero-shot classification across diverse datasets. However, CLIP's general-purpose features are often suboptimal for specialized domains, necessitating downstream fine-tuning. Benchmarking studies like [2] highlight that while CLIP is a robust foundation, its performance varies significantly across task-specific distributions.

2.3. Fine-Tuning and Feature Distortion

Standard full fine-tuning (Full FT) updates all model parameters to maximize in-domain accuracy. While effective, this process can lead to "feature distortion," where the well-distributed pretrained embeddings collapse into narrow, task-specific manifolds [3].

Wortsman et al. [4] observed that fine-tuning often degrades the model's out-of-distribution (OOD) robustness, prompting the development of weight-ensembling and constrained optimization techniques to preserve the pretrained knowledge.

2.4. Low-Rank Adaptation (LoRA)

To mitigate the computational cost and "forgetting" issues of full fine-tuning, parameter-efficient fine-tuning (PEFT) methods have gained prominence. LoRA (Low-Rank Adaptation) [12] freezes the pretrained weights and injects trainable low-rank matrices into the attention projections. This significantly reduces the degrees of freedom during optimization. While initially developed for Large Language Models (LLMs), LoRA has become a standard for adapting vision models, though its impact on the internal attention structure of multimodal encoders like CLIP remains less explored.

2.5. Attention Analysis and Interpretability

The interpretability of self-attention distributions is a subject of ongoing debate. While attention maps are often used to identify "important" regions of an image, Jain and Wallace [11] argued that they do not always provide a faithful explanation of model decisions. Conversely, Wiegreffe and Pinter [9] and Chefer et al. [10] emphasized that attention provides a valuable structural roadmap of information flow, especially when aggregated via methods like attention rollout [13]. We follow this structural perspective to measure the geometric shifts occurring during CLIP adaptation.

3. Materials and Methods

3.1. Problem Formulation

Let f_0 represent the weights of a pretrained CLIP vision encoder. During adaptation, we minimize a cross-entropy loss $\mathcal{L}_{\text{task}}$ to obtain updated weights θ. We monitor the CLS-to-patch attention distributions A across all layers $\ell \in \{1, \ldots, 12\}$. Since the CLS token serves as the primary conduit for global feature synthesis in the CLIP ViT-B/32 architecture, its attention profile provides a direct measure of how information is sampled from the spatial tokens.

3.2. Structural Metrics

We define four primary metrics to quantify the spatial geometry of the CLS-to-patch attention vectors $a \in \mathbb{R}^{49}$.

3.2.1. Attention Entropy

We quantify the uncertainty or "spread" of the distribution using Shannon entropy:

$$H(a) = -\sum_{j=1}^{49} a_j \ln a_j. \tag{1}$$

Consistent with standard practice in transformer analysis, we report entropy in nats. A decrease in entropy indicates a more concentrated attention focus.

3.2.2. Effective Receptive Field (ERF@0.95)

To measure the spatial fraction of the image required to account for the majority of the attention mass, we define:

$$\text{ERF}_{0.95}(a) = \frac{1}{49} \min \left\{ k : \sum_{j=1}^{k} a_{(j)} \geq 0.95 \right\}, \tag{2}$$

where $a_{(j)}$ are the sorted attention weights. This metric is particularly useful for identifying models that rely on a very small subset of tokens.

3.2.3. Gini Concentration

The Gini coefficient measures the inequality within the attention distribution. A Gini value of 0 indicates a perfectly uniform distribution, while a value approaching 1 indicates extreme sparsity where nearly all mass is allocated to a single token.

3.2.4. Head Diversity

Transformer layers typically employ multiple heads to capture diverse feature relations [5]. We define head diversity as $1 - \bar{\rho}$, where $\bar{\rho}$ is the mean pairwise cosine similarity between the CLS-to-patch attention maps of all heads within a layer. Lower diversity suggests a collapse of head-wise specialization.

3.3. Additional Validation Metrics

To ensure the robustness of our findings, we incorporate auxiliary analyses. Attention rollout [13] provides a multi-layer view of routing. Layerwise centered kernel alignment (CKA) is used to measure the similarity between pretrained and fine-tuned feature representations. Subset-sensitivity analysis serves as a control for dataset-specific artifacts, ensuring that the observed structural shifts are representative of the model's behavior rather than noise in a specific evaluation batch.

3.4. Experimental Setup

Our experiments utilize the `openai/clip-vit-base-patch32` model. We evaluate adaptation on EuroSAT [14] (for remote sensing) and Oxford-IIIT Pets [15] (for fine-grained classification). These datasets represent distinct visual domains, allowing us to assess the consistency of structural changes.

For the optimization, we use AdamW [16] with its standard implementation [17] and a cosine-decay schedule. In full fine-tuning (Full FT), all vision encoder parameters are updated. In LoRA, we insert low-rank adapters into the query and value projections (`q_proj`, `v_proj`) with rank $r = 8$ and scaling $\alpha = 16$. Experimental runs were standardized using NVIDIA A40 hardware to ensure reproducibility.

3.5. Statistical Analysis

Because per-layer measurements within a single model are not independent, per-layer t-tests are treated as exploratory rather than confirmatory. For the five-point learning-rate analysis, approximate small-sample p-values are therefore replaced with exact permutation tests. For the family of regularizer comparisons, Holm-Bonferroni-adjusted p-values are reported. The available evidence is nonetheless limited by the absence of multi-seed replications for the main training configurations.

4. Results

4.1. Overall Comparison of Adaptation Regimes

Table 1 summarizes the central comparisons. Full fine-tuning reaches the highest EuroSAT validation accuracy (99.30%) but slightly reduces average attention entropy (-0.14%) and ERF (-1.73%) relative to the pretrained baseline. On Pets, full fine-tuning produces a larger entropy decrease (-1.82%), a larger ERF contraction (-4.23%), and a substantial Gini increase ($+10.54\%$), indicating stronger concentration. By contrast, the EuroSAT LoRA configurations consistently increase entropy and reduce Gini relative to the baseline while keeping task accuracy above 98.6%.

Table 1. Main experiment results. Positive ΔEntropy and ΔERF indicate broader attention support relative to the corresponding pretrained baseline; positive ΔGini indicates more concentrated attention.

Method	Dataset	Task Acc. (%)	ΔEntropy (%)	ΔERF (%)	ΔGini (%)	ΔHeadDiv (%)	CIFAR-100 ZS (%)	Flowers102 ZS (%)
Full FT	EuroSAT	99.30	-0.14	-1.73	-3.54	-0.62	12.71	1.37
Full FT	Oxford-IIIT Pets	90.79	-1.82	-4.23	+10.54	-35.91	5.89	0.88
LoRA r=4	EuroSAT	98.67	+0.98	+1.09	-9.94	+3.86	60.22	0.20
LoRA r=8	EuroSAT	98.96	+0.58	+0.37	-6.48	+7.17	60.22	0.20
LoRA r=16	EuroSAT	98.93	+0.44	+0.03	-5.43	+5.14	60.22	0.20
LoRA r=8	Oxford-IIIT Pets	92.07	-0.73	-2.64	+3.20	-15.00	—	—

Figure 1 shows that the pretrained model already exhibits heterogeneous attention structure across depth, so the appropriate question is not whether fine-tuning creates structure from nothing, but whether it pushes that structure toward greater concentration. Figure 2 then shows that full fine-tuning and LoRA trace visibly different trajectories in entropy, ERF, Gini, and head diversity.

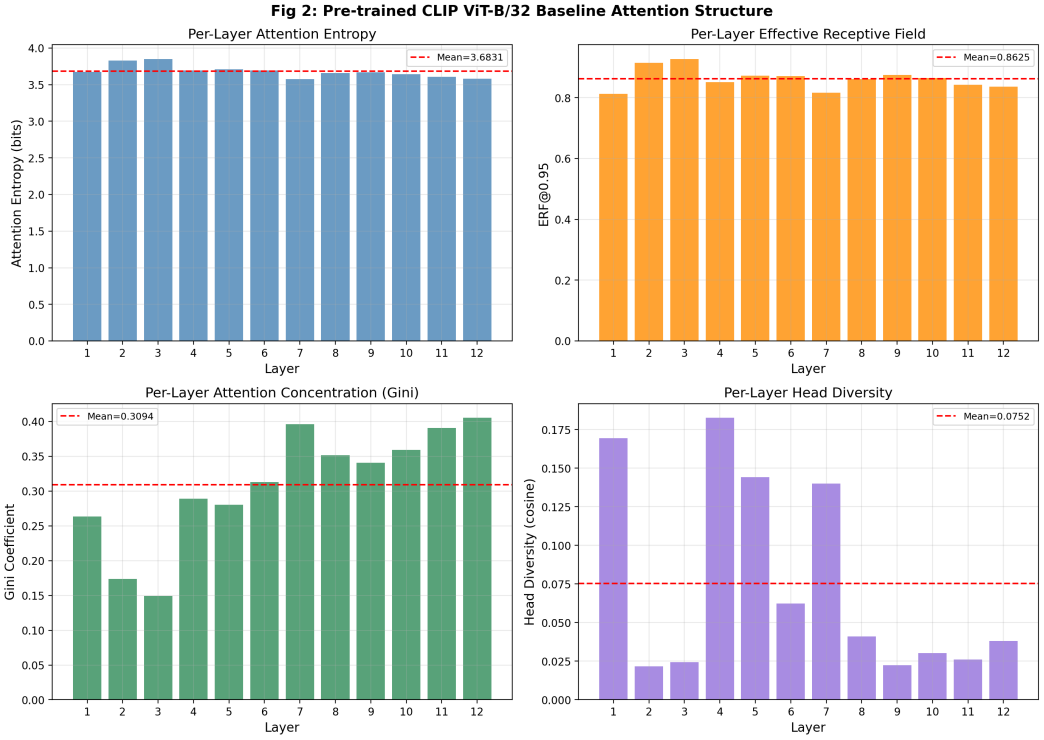

Figure 1. Baseline attention structure of pretrained CLIP ViT-B/32 across layers. The model is already depth-heterogeneous before any downstream adaptation.

Table 2. Corrected inferential statistics. Exact permutation tests are used for the five-point learning-rate analysis. Holm–Bonferroni correction is applied across the regularizer family; no regularizer configuration survives this correction.

Comparison	Test	Statistic	p-value	Note
LR vs. ΔEntropy	Exact permutation Pearson r	-0.893	0.0333	$n = 5$ learning rates
LR vs. ΔEntropy	Exact permutation Spearman ρ	-1.000	0.0167	Small-n exact test
APR $\lambda = 0.01$ vs. baseline	Paired per-layer t-test	2.132	0.2194	Holm–Bonferroni adjusted
Entropy floor $\lambda = 0.1$ vs. baseline	Paired per-layer t-test	2.638	0.1154	No regularizer survives Holm correction

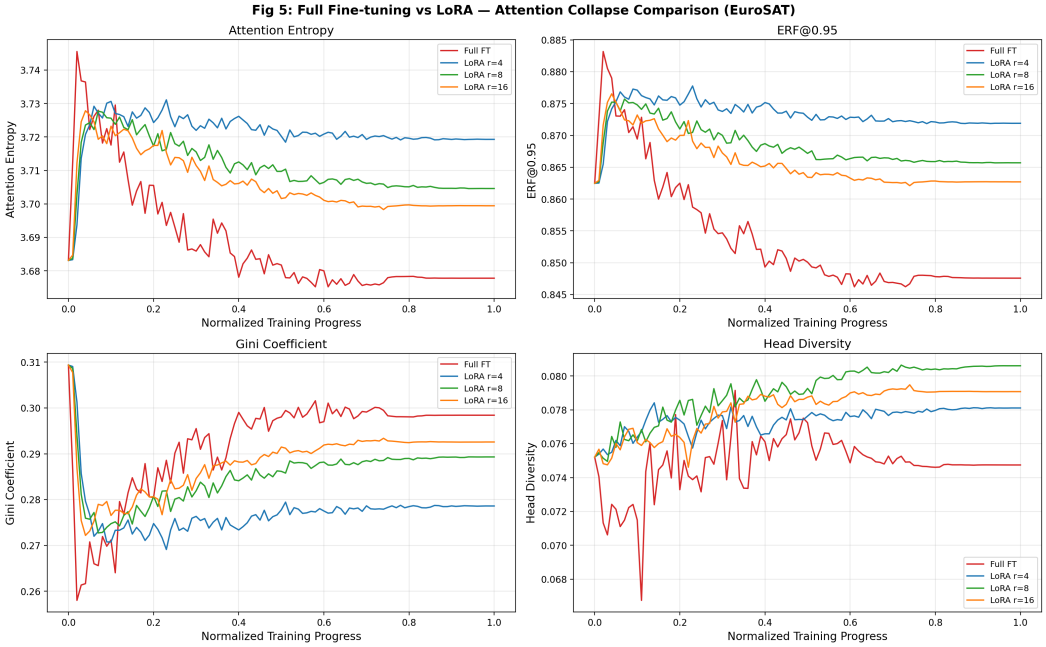

Figure 2. EuroSAT full fine-tuning versus LoRA. LoRA increases entropy and ERF relative to the pretrained baseline, whereas full fine-tuning remains near neutral to mildly contractive.

4.2. Statistical Evidence and Layerwise Patterns

Figure 3 shows that the direction and magnitude of per-layer entropy change depend on both dataset and adaptation regime. However, because layers within a single trained model are not statistically independent, we do not treat these per-layer comparisons as formal evidence of independence. Instead, we interpret the layerwise patterns descriptively and reserve inferential claims for analyses whose sampling assumptions are more defensible.

The clearest corrected inferential result concerns the learning rate. When the five EuroSAT full fine-tuning learning rates are analyzed with exact permutation tests, the negative association between learning rate and Δentropy remains significant (Pearson $p = 0.0333$; Spearman $p = 0.0167$). By contrast, the regularization family no longer contains a configuration that survives Holm–Bonferroni correction, including entropy floor with $\lambda = 0.1$.

Table 3. Learning-rate sweep for EuroSAT full fine-tuning. Entropy decreases monotonically as the learning rate increases across the tested range.

LR	Task Acc. (%)	ΔEntropy (%)	ΔERF (%)	ΔGini (%)	CIFAR-100 ZS (%)	Flowers102 ZS (%)
1e-6	98.85	+1.45	+1.94	-16.60	—	—
5e-6	99.11	+0.33	-0.49	-6.68	—	—
1e-5	99.30	-0.14	-1.73	-3.54	12.71	1.37
5e-5	98.89	-4.19	-8.48	+23.44	2.07	0.88
1e-4	97.85	-10.40	-18.02	+56.94	1.86	0.98

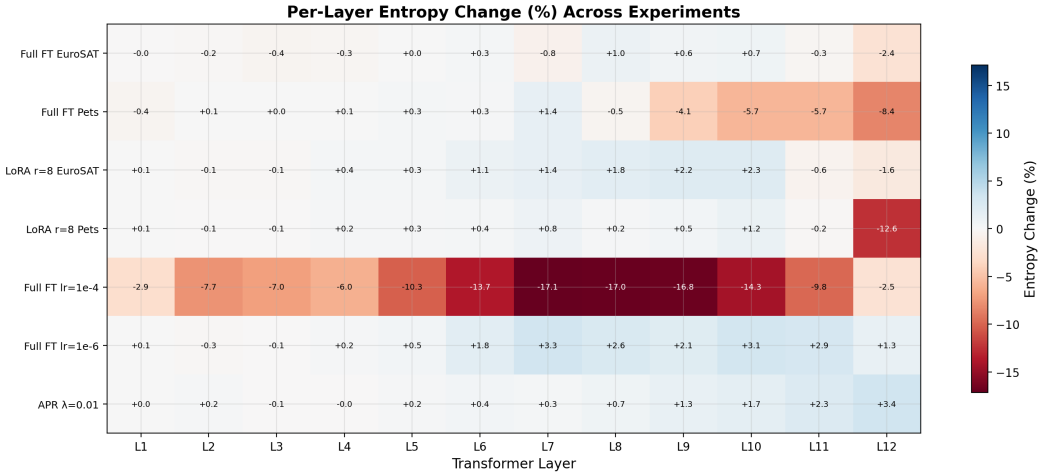

Figure 3. Per-layer entropy change across selected experiments. The sign and magnitude of the change depend on the training regime and the dataset.

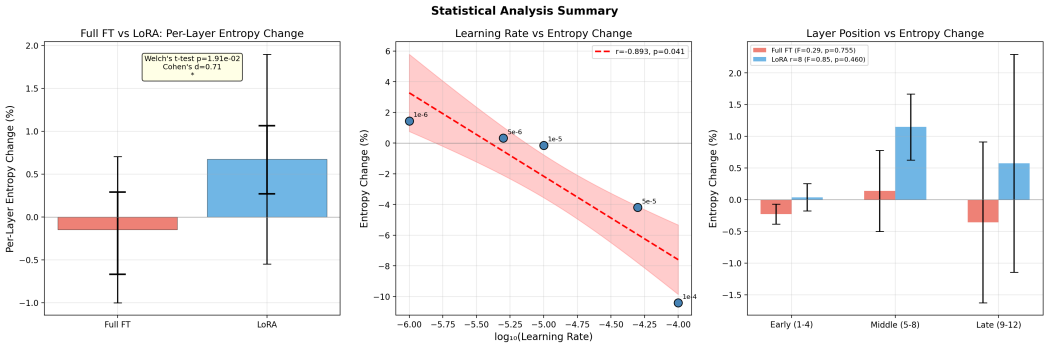

Figure 4. Summary of exploratory per-layer comparisons and corrected statistical analyses, including the learning-rate trend and layer-group summaries.

4.3. Impact of Learning Rate on Structural Dynamics

Among the manipulated hyperparameters, the learning rate appeared to be the dominant control variable for structural change. Table 3 shows a monotonic trend: entropy increases at 10^{-6}, is nearly neutral at 10^{-5}, and decreases sharply at 5×10^{-5} and 10^{-4}. Under exact permutation testing on the five learning-rate settings, the Pearson correlation between $\log_{10}(\text{LR})$ and Δentropy is -0.893 ($p = 0.0333$), while the Spearman coefficient is -1.0 ($p = 0.0167$).

Table 4. Regularization study on EuroSAT full fine-tuning. Among individually tested regularizers, entropy floor with $\lambda = 0.1$ is the only setting with a paired per-layer entropy difference reaching $p < 0.05$ in the current analysis.

Setting	Task Acc. (%)	ΔEntropy (%)	ΔERF (%)	ΔGini (%)
No regularizer	99.30	-0.14	-1.73	-3.54
APR $\lambda = 0.01$	99.26	+0.86	+1.50	-8.83
APR $\lambda = 0.1$	99.19	+0.56	+1.03	-5.36
APR $\lambda = 1.0$	98.96	+0.09	+0.16	-0.75
Entropy floor $\lambda = 0.01$	99.11	+0.58	-0.09	-6.41
Entropy floor $\lambda = 0.1$	99.07	+0.79	+0.30	-8.22
Weight decay 0.0	99.19	-0.18	-1.81	-3.16
Weight decay 0.1	99.15	-0.25	-1.90	-2.38

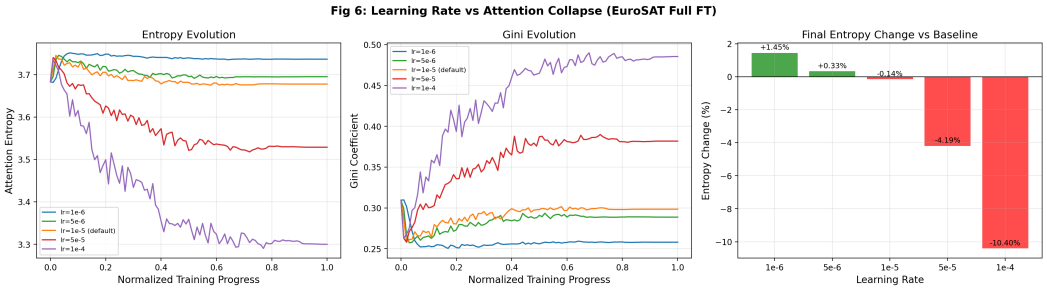

Figure 5. Learning-rate sweep for EuroSAT full fine-tuning. Larger learning rates produce stronger entropy decreases and stronger ERF contraction.

This result matters because it reframes collapse as an optimization-scale phenomenon rather than merely a binary property of full fine-tuning versus parameter-efficient tuning. In the revised analysis, the effect remains when the small-sample correlation test is recomputed exactly rather than through asymptotic approximations.

4.4. Regularization and Training Dynamics

Table 4 summarizes the regularization study. Both APR and entropy-floor regularization shift entropy upward relative to the no-regularizer baseline, while preserving high task accuracy. However, after Holm–Bonferroni correction across the regularizer family, no individual configuration remains statistically significant within the exploratory per-layer comparison framework used here. The regularization results should therefore be interpreted as promising but preliminary.

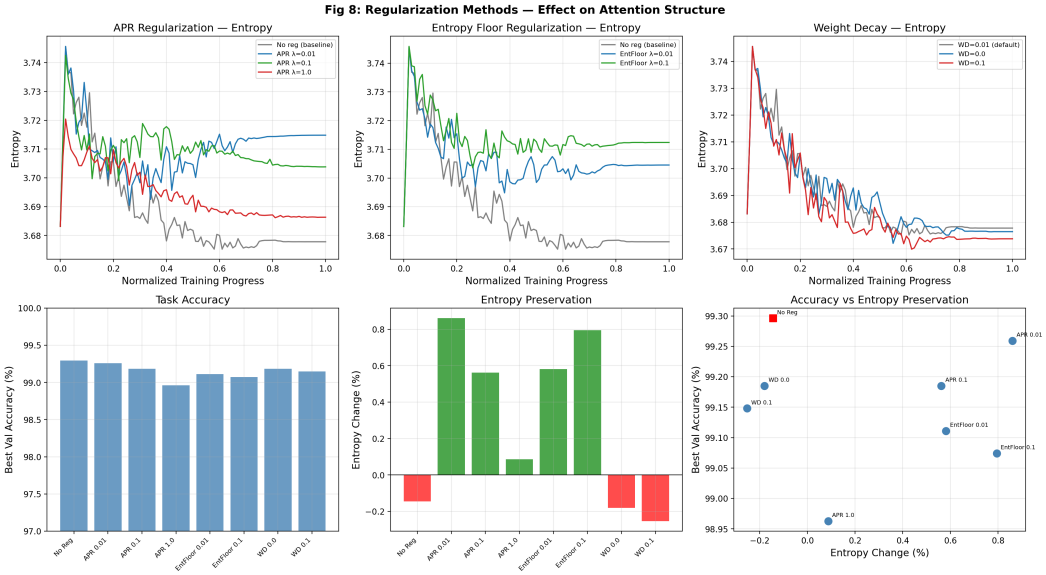

Figure 6. Regularization results. Both APR and entropy floor preserve a broader attention structure than the no-regularizer baseline in descriptive terms. Still, none of the tested settings survives Holm–Bonferroni correction across the regularizer family.

Figure 7 complements the final-state tables by showing training trajectories and collapse-onset markers. The main qualitative lesson is that structural drift appears early when the learning rate is high, whereas LoRA traces a flatter path. Again, this constitutes observational evidence rather than a mechanistic proof.

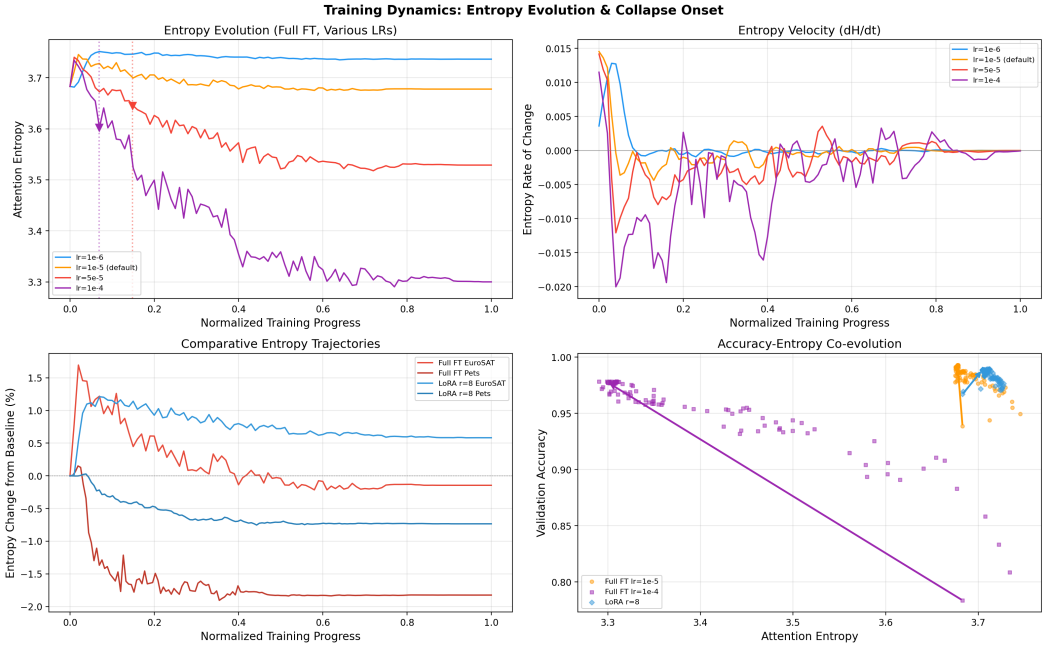

Figure 7. Training dynamics and collapse-onset analysis. High-learning-rate full fine-tuning exhibits the earliest and strongest decrease in entropy.

4.5. Auxiliary Structural Validation

To test whether the main conclusions are specific to CLS-to-patch entropy, Table 5 reports supplementary rollout, patch-to-patch, representational-similarity, and subset-sensitivity analyses for the clearest EuroSAT comparison. LoRA $r = 8$ exhibits higher rollout entropy, larger rollout ERF@0.95, lower rollout Gini, and higher patch-to-patch entropy than full fine-tuning, indicating that the broader-support pattern is not confined to the original

Table 5. Auxiliary validation analyses beyond CLS-to-patch entropy. Rollout and patch-to-patch metrics indicate broader attention support for LoRA than for Full FT on EuroSAT, while layerwise CKA shows higher representational similarity to the pretrained encoder. The subset-sensitivity coefficient of variation (CV) is low for both methods, indicating that the reported metrics are stable across five balanced evaluation subsets.

Model	Rollout Entropy	Rollout ERF@0.95	Rollout Gini	PTP Entropy	Mean CKA	Entropy CV
Full FT EuroSAT	5.593	0.944	0.090	4.721	0.694	0.0006
LoRA $r = 8$ EuroSAT	5.605	0.952	0.064	4.794	0.815	0.0005

Table 6. Zero-shot transfer accuracy (%) on CIFAR-100 and Flowers102 after EuroSAT fine-tuning. LoRA preserves substantially more CIFAR-100 accuracy than full fine-tuning, while Flowers102 values are uniformly low for all models and should be treated as supplemental evidence only.

Model	CIFAR-100 (%)	Flowers102 (%)
Pretrained	60.22	0.20
Full FT (EuroSAT)	9.43	1.37
LoRA $r = 4$ (EuroSAT)	40.51	1.27
LoRA $r = 8$ (EuroSAT)	47.28	0.78
LoRA $r = 16$ (EuroSAT)	39.35	0.98

CLS-to-patch metric. LoRA also yields a higher mean layerwise centered kernel alignment to the pretrained model (0.815 versus 0.694), suggesting less representational drift. Finally, the coefficient of variation of the entropy metric across five balanced evaluation subsets is below 0.001 for both methods, indicating that the reported structural metrics are not highly sensitive to the choice of evaluation images.

4.6. Zero-Shot Transfer: Evidence and Caveats

The zero-shot table is included for completeness, but its interpretation requires care. Reliable evaluation of LoRA models requires an inference procedure that activates the learned adapters in the image encoder. Accordingly, the EuroSAT LoRA models were reevaluated using this procedure, while the pretrained baseline was retained from the standard CLIP evaluation pipeline.

The revised results materially change the interpretation. LoRA does not preserve the full pretrained CIFAR-100 accuracy of 60.22%, but it degrades far less than full fine-tuning: LoRA $r = 8$ reaches 47.28%, compared with 9.43% for full fine-tuning. LoRA, therefore, remains structurally and functionally more conservative than full fine-tuning in this setting. Still, the earlier, stronger claim of near-perfect transfer preservation is not supported once adapters are evaluated correctly.

The Flowers102 values are also uniformly low (all below 1.37%), suggesting that the current prompt and label specification do not provide a sufficiently reliable primary benchmark for this study. For that reason, Flowers102 is treated as supplemental rather than central evidence.

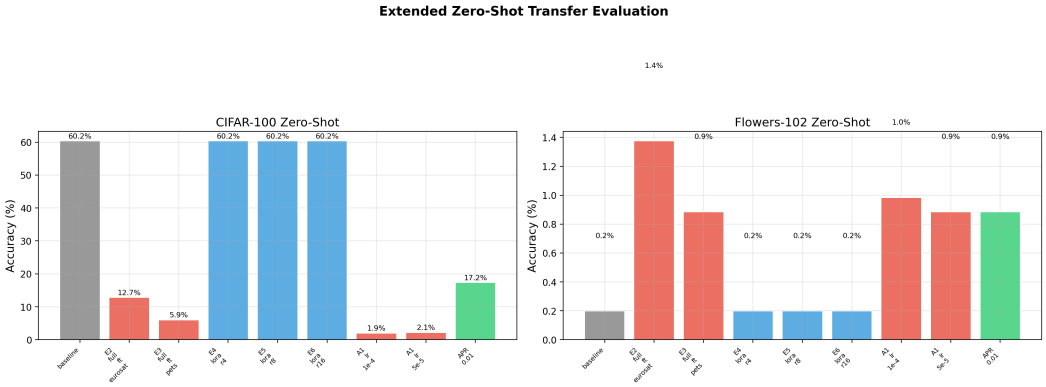

Figure 8. Extended zero-shot evaluation under the study's evaluation protocol. CIFAR-100 is the more interpretable benchmark, whereas the Flowers102 values are uniformly too low to support strong conclusions.

5. Discussion

The results suggest that attention structural change is neither negligible nor uniform across adaptation strategies. Across the experiments analyzed here, full fine-tuning tends to narrow attention support, particularly on Oxford-IIIT Pets. In contrast, EuroSAT LoRA maintains or slightly broadens attention support relative to the pretrained baseline. The auxiliary rollout and patch-to-patch analyses reinforce this interpretation, and the centered kernel alignment results indicate that LoRA remains closer to the pretrained representation geometry than full fine-tuning.

The learning-rate sweep further indicates that the magnitude of optimization steps is a key determinant of structural change. This observation is noteworthy because it implies that attention contraction is not simply an intrinsic property of full fine-tuning; rather, it is strongly modulated by training scale. From a practical perspective, this suggests that structurally conservative adaptation may be achievable not only through parameter-efficient methods such as LoRA but also through careful optimization control.

The regularization results provide only cautious support for this interpretation. Although APR and entropy-floor regularization shift the model toward broader attention distributions, the available per-layer evidence does not survive a family-wise multiple-comparison correction. These results, therefore, motivate but do not yet establish regularization as a reliable solution.

6. Conclusions

This study examined structural changes in attention during fine-tuning of CLIP through a comparative analysis of full fine-tuning, LoRA, and regularized adaptation. The analysis combines exact tests, where appropriate, with supplementary rollout, patch-to-patch, centered-kernel alignment, and subset-sensitivity measurements to triangulate the main findings. Under this deliberately conservative inferential framing, the central qualitative conclusion remains intact: adaptation regime and learning rate materially affect attention structure, and LoRA is substantially less disruptive than full fine-tuning on the clearest EuroSAT comparison. At the same time, the study delineates what remains unresolved, particularly regarding run-level inference and the generality of the findings across seeds, backbones, and transfer benchmarks.

Author Contributions: R.X. conceived the study, designed and conducted all experiments, performed the statistical analyses, and wrote the manuscript. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code, manuscript source, figures, tables, and analysis artifacts supporting this study are publicly available at https://github.com/xiaruize0911/Attention_Collapse_in_CLIP_Fine-tuning_repo. Additional derived outputs used in this study are available from the corresponding author on reasonable request.

Acknowledgments: The author thanks the open-source software and dataset communities whose tools and benchmarks enabled this study. Computational resources for the experiments were provided through rented NVIDIA A40 instances on RunPod.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

APR	Attention preservation regularization
CKA	Centered kernel alignment
CLIP	Contrastive Language–Image Pretraining
ERF	Effective receptive field
FT	Fine-tuning
LoRA	Low-rank adaptation

References

1. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 8748–8763.
2. Shao, W.; Zhang, M.; Hong, Z.; Zhou, Z.; Qiu, S.; Liang, F.; Yu, X.; Li, K.; Zhang, P.; Wu, Y.; et al. VLM-Eval: A General Evaluation Benchmark for Vision-Language Models. *arXiv preprint arXiv:2311.08419* **2023**.
3. Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; Liang, P. Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution. In Proceedings of the International Conference on Learning Representations, 2022.
4. Wortsman, M.; Ilharco, G.; Kim, J.W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R.G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. Robust Fine-Tuning of Zero-Shot Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7959–7971.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, 2017.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, 2021.
7. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks? In Proceedings of the Advances in Neural Information Processing Systems, 2021.
8. Caron, M.; Touvron, H.; Misra, I.; Hervé, J.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
9. Wiegreffe, S.; Pinter, Y. Attention is not not Explanation. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 11–20.
10. Chefer, H.; Gur, S.; Wolf, L. Transformer Interpretability Beyond Attention Visualization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782–791.
11. Jain, S.; Wallace, B.C. Attention is not Explanation. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 3543–3556.
12. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, 2022.
13. Abnar, S.; Zuidema, W. Quantifying Attention Flow in Transformers. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4190–4197.
14. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. 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* **2019**, *12*, 2217–2226.

15. Parkhi, O.M.; Vedaldi, A.; Zisserman, A.; Jawahar, C.V. Cats and Dogs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3498–3505.
16. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, 2019.
17. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, 2015.