Article

Not peer-reviewed version

Attention Heatmap Drift in a Contrastively Pretrained Vision–Language Model: A Controlled Matched-Learning-Rate Comparison of Full Fine-Tuning and Low-Rank Adaptation

Ruize Xia *

Posted Date: 6 April 2026

doi: 10.20944/preprints202604.0317.v1

Keywords: contrastive vision–language pretraining; full fine-tuning; low-rank adaptation; vision transformers; attention heatmap drift; transfer learning

 doi:10.20944/preprints202604.0317.v1

Article

Attention Heatmap Drift in a Contrastively Pretrained Vision–Language Model: A Controlled Matched-Learning-Rate Comparison of Full Fine-Tuning and Low-Rank Adaptation

Ruize Xia

Independent Researcher; xiaruize0911@gmail.com

Abstract

Downstream adaptation of a contrastively pretrained vision–language model can improve in-domain accuracy while degrading performance on unseen transfer tasks. This study examines how full fine-tuning and low-rank adaptation alter attention heatmaps under a controlled design that matches learning rate across adaptation methods. The completed matched-learning-rate matrix contains 80 runs using the OpenAI Contrastive Language–Image Pretraining model with a base 32-patch vision transformer image encoder, two datasets (EuroSAT and Oxford-IIIT Pets), four shared learning rates (1e-6, 5e-6, 1e-5, and 5e-5), and five random seeds. We measure classification-token-to-patch attention entropy, the fraction of patches required to capture 95% of attention mass, attention concentration, head diversity, in-domain validation accuracy, and adapter-aware zero-shot accuracy on CIFAR-100. Three findings emerge. First, learning rate is a primary determinant of structural drift: on EuroSAT, full fine-tuning moves from entropy broadening at 1e-6 (+1.83%) to marked contraction at 5e-5 (-3.99%), whereas low-rank adaptation remains entropy-positive across the full matched grid (+0.68% to +1.50%). Second, low-rank adaptation preserves out-of-domain transfer substantially better than full fine-tuning at matched learning rates: averaged across the EuroSAT grid, zero-shot accuracy on CIFAR-100 is 45.13% for low-rank adaptation versus 11.28% for full fine-tuning; on Oxford-IIIT Pets, the corresponding averages are 58.01% and 8.54%. Third, Oxford-IIIT Pets exhibits a clear interaction with optimization scale: low-learning-rate low-rank adaptation underfits the in-domain task, so method-only averages can obscure the regime in which it becomes competitive. Additional rollout, patch-to-patch, centered-kernel-alignment, and backbone analyses are directionally consistent with these controlled results. Across both controlled datasets, runs with broader retained attention support also retain more zero-shot performance. Taken together, these findings support attention heatmap drift as an informative descriptive lens on model adaptation while arguing against a universal interpretation of the observed behavior as a single collapse phenomenon.

Keywords: contrastive vision–language pretraining; full fine-tuning; low-rank adaptation; vision transformers; attention heatmap drift; transfer learning

1. Introduction

Contrastive Language–Image Pretraining (CLIP) has become a widely used foundation for visual recognition because zero-shot predictions often remain strong even before task-specific optimization [1]. Yet downstream adaptation introduces a recurring tension between in-domain accuracy and retention of the pretrained model's broader transfer behavior. In particular, full fine-tuning (Full FT) can improve target-task accuracy while degrading robustness under distribution shift, whereas parameter-efficient methods such as low-rank adaptation (LoRA) often preserve more of the pretrained representation [2–4]. For deployed vision–language systems, that trade-off matters as much as the final validation score.

 doi:10.20944/preprints202604.0317.v1

One informative, but deliberately limited, lens on this trade-off is the internal attention structure of the Vision Transformer (ViT) image encoder. Self-attention governs how information flows between the classification token and spatial patch tokens across depth [5]. Attention weights are not treated here as causal explanations of predictions [6,7]. Instead, they are used as descriptive structural measurements. We refer to the change in classification-token-to-patch attention support induced by downstream optimization as *attention heatmap drift*.

A central difficulty in the existing Full-FT-versus-LoRA literature is that the two methods are often compared under different learning-rate conventions. When optimization scales differ, apparent method effects become entangled with update-magnitude effects. That confound is especially problematic for structural analyses, because attention drift can change sharply with learning rate. A controlled matched-learning-rate design is therefore necessary if the goal is to distinguish method effects from optimization-regime effects.

This paper addresses three questions. First, under matched learning rates, how do Full FT and LoRA differ in attention heatmap drift? Second, how do those structural differences relate to in-domain validation accuracy and out-of-domain (OOD) zero-shot transfer retention? Third, are the observed patterns distributed uniformly across the model or concentrated in particular layers and training regimes?

To answer these questions, we analyze a completed 80-run matrix on the CLIP ViT-B/32 image encoder across EuroSAT and Oxford-IIIT Pets. Full FT and LoRA share the same four-learning-rate grid and five random seeds, and every run is evaluated with the same structural and transfer protocol. The matched matrix provides the primary evidence for the paper's claims. Earlier repository analyses—including rollout, patch-to-patch, centered kernel alignment (CKA), regularization, and backbone checks—are retained as supporting evidence because they broaden the empirical scope without displacing the controlled comparison.

The paper makes four contributions:

- it provides a completed 80-run matched-learning-rate comparison of Full FT and LoRA on CLIP ViT-B/32 across EuroSAT and Oxford-IIIT Pets;
- it shows that learning rate is a first-order determinant of attention heatmap drift, so method comparisons change materially when optimization scale is controlled;
- it shows that LoRA generally preserves substantially more zero-shot transfer than Full FT at matched learning rates; and
- it introduces a reproducible structural-analysis protocol and integrates the main controlled matrix with supporting analyses while maintaining a clear evidentiary hierarchy.

The next section reviews related work. A short preliminaries section then defines the adaptation setting and the structural quantities studied here, after which the materials and methods section describes the reproducible experimental protocol.

2. Related Work

2.1. CLIP Adaptation

CLIP established large-scale image–text pretraining as a strong basis for transfer learning across diverse downstream tasks [1]. A substantial subsequent literature then examined what happens when zero-shot models are adapted rather than used directly. In particular, Kumar et al. [2] showed that full fine-tuning can distort pretrained features and underperform under distribution shift, while Wortsman et al. [3] argued that robust adaptation of zero-shot models requires more care than simply optimizing for in-domain accuracy. These studies motivate the central question of the present paper: if downstream adaptation changes transfer behavior, what measurable internal structural changes accompany that shift?

Our work differs from this prior line in emphasis. The studies above focus primarily on predictive robustness and representation quality at the output level. We instead examine a structural correlate

 doi:10.20944/preprints202604.0317.v1

inside the visual encoder, namely the drift of attention heatmaps relative to the pretrained baseline. The paper therefore complements, rather than replaces, robustness-oriented analyses of CLIP adaptation.

2.2. Attention in Vision Transformers

The visual backbone of CLIP follows the Vision Transformer formulation of Dosovitskiy et al. [5], in which token interactions are mediated by multi-head self-attention. That architecture makes it natural to study how adaptation changes information routing across depth. Prior work on attention flow, especially attention rollout [8], provides one concrete mechanism for summarizing layerwise attention patterns beyond any single matrix. This broader literature supports the idea that attention can be analyzed as a structural property of the model even when it is not interpreted as a complete explanation of prediction behavior.

At the same time, the interpretability status of attention remains contested. Jain and Wallace [6] argued that attention weights should not automatically be equated with explanations, whereas Wiegreffe and Pinter [7] showed that attention can still support informative analysis when its role is framed more carefully. We adopt that more conservative position. In this paper, attention heatmaps are treated as descriptive measurements of how downstream optimization reshapes the visual encoder, not as stand-alone causal explanations of model decisions.

2.3. Parameter-Efficient Adaptation

LoRA introduced a low-rank parameterization of downstream updates that greatly reduces the number of trainable parameters while preserving much of the base model [4]. Although LoRA was originally developed and evaluated in language modeling settings, the broader idea of parameter-efficient adaptation has clear relevance for vision–language models: if the update space is constrained, one may expect the adapted model to remain closer to the pretrained representation geometry. This intuition is especially relevant for CLIP, where preserving pretrained transfer behavior is often as important as maximizing in-domain accuracy.

The present paper builds on that motivation but introduces a stronger experimental control than most method-level comparisons. Rather than assuming that any observed Full-FT-versus-LoRA gap is attributable to the adaptation method alone, we explicitly hold the learning-rate grid fixed. This control is important because optimization scale can itself induce substantial representational and structural drift.

2.4. Study Positioning

Taken together, the prior literature leaves an important gap. CLIP adaptation work shows that robustness can degrade after fine-tuning, transformer-analysis work provides tools for studying internal attention structure, and LoRA offers a mechanism for constraining adaptation. What is less well established is how these pieces fit together under a controlled comparison that disentangles adaptation method from learning rate.

The contribution of the present study is therefore not the novelty of attention analysis in CLIP per se. Rather, it is a controlled matched-learning-rate evaluation of attention heatmap drift in which method, learning rate, transfer retention, and layerwise structure are analyzed jointly. The broader analyses in the paper are included as supporting results that complement, but do not replace, the controlled matrix.

3. Preliminaries

3.1. Adaptation Setting

CLIP couples an image encoder and a text encoder trained with a contrastive objective [1]. The present study analyzes the image branch, which is a ViT with 12 transformer layers and 49 spatial patch tokens for a 224×224 input, together with a leading classification token (CLS) that aggregates global image information. Downstream adaptation adds a task-specific linear classifier for the target dataset. In the Full FT condition, all image-branch parameters used at inference are optimized on

 doi:10.20944/preprints202604.0317.v1

the downstream task. In the LoRA condition, low-rank update matrices are inserted into selected self-attention projections while most pretrained weights remain fixed [4].

3.2. Attention Heatmap Drift and Structural Metrics

For layer ℓ and attention head h, let $\mathbf{a}^{(\ell,h)} \in \mathbb{R}^{49}$ denote the normalized attention weights from the CLS token to the 49 spatial patch tokens, with $\sum_{i=1}^{49} a_i^{(\ell,h)} = 1$. We summarize support breadth with entropy,

$$H^{(\ell,h)} = -\sum_{i=1}^{49} a_i^{(\ell,h)} \log a_i^{(\ell,h)}. \tag{1}$$

We also report the effective receptive field at 95% mass (ERF@0.95), defined as the smallest fraction of sorted patch weights needed to accumulate 95% of the total attention mass; the Gini coefficient, which measures concentration of the same distribution; and head diversity, computed as mean pairwise cosine dissimilarity across heads within a layer.

For any scalar structural metric m, drift is reported as percent change relative to the pretrained baseline,

$$\Delta m = 100 \times \frac{m_{\text{adapted}} - m_{\text{pretrained}}}{m_{\text{pretrained}}}. \tag{2}$$

Positive entropy or ERF drift indicates broader attention support, whereas positive Gini drift indicates increased concentration. Throughout the paper, the matched-learning-rate matrix is treated as the primary evidence, and auxiliary rollout, patch-to-patch, CKA, regularization, and backbone analyses are interpreted as contextual checks rather than replacements for the controlled comparison.

4. Materials and Methods

4.1. Model, Datasets, and Downstream Tasks

All controlled experiments use the OpenAI CLIP ViT-B/32 checkpoint (openai/clip-vit-base-patch32) with eager attention enabled so that per-head attention maps can be extracted from the visual encoder [1,5]. The matched matrix spans two downstream classification datasets, EuroSAT [9] and Oxford-IIIT Pets [10], and evaluates transfer retention on CIFAR-100 [11]. In the released code, the datasets are loaded from the Hugging Face entries `tanganke/eurosat`, `timm/oxford-iiit-pet`, and `uoft-cs/cifar100`. Images are resized to 224×224, center-cropped, normalized with the pretrained CLIP mean and standard deviation, and augmented during training with random horizontal flips.

EuroSAT runs are trained for 20 epochs and Oxford-IIIT Pets runs for 30 epochs. The downstream objective is cross-entropy loss on a task-specific linear classification head, and best validation accuracy is recorded across epochs rather than selected by early stopping. The CLIP text tower is not updated during downstream fine-tuning; it is used only for the zero-shot transfer evaluation described below.

4.2. Matched-Learning-Rate Design and Optimization

The controlled matrix is fully crossed over two adaptation methods, four shared learning rates (10^{-6}, 5×10^{-6}, 10^{-5}, 5×10^{-5}), two datasets, and five random seeds (7, 11, 19, 42, 123), yielding 80 completed runs. Optimization uses AdamW with weight decay 0.01, 5% linear warmup followed by cosine decay, batch size 64 for training, and gradient clipping at norm 1.0.

In Full FT, all image-branch parameters used during downstream inference are optimized together with the linear classifier. In LoRA, adapters with rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.05 are inserted into the `q_proj` and `v_proj` layers of the visual self-attention blocks, while the remaining pretrained image-branch weights stay fixed and the classifier head plus visual projection remain trainable [4]. The learning-rate grid is intentionally identical across Full FT and LoRA so that method effects can be compared without changing optimization scale.

 doi:10.20944/preprints202604.0317.v1

4.3. Structural Evaluation Protocol

For each checkpoint, structural metrics are computed against the pretrained baseline on five fixed, class-balanced validation subsets. Each subset contains 200 images and is generated once in the released pipeline so that all runs are evaluated on the same image pools. Structural evaluation uses batch size 32. Attention tensors are extracted with `output_attentions=True`. For every layer and head, the analysis keeps attention from the CLS token to the 49 spatial patch tokens, renormalizes the patch weights to sum to one, and then computes entropy, ERF@0.95, Gini concentration, and head diversity. Reported values are averaged over images, heads, and evaluation subsets, and then converted to percent change relative to the pretrained checkpoint.

Supplementary rollout metrics average heads, inject a residual identity contribution at each layer, and compose the resulting matrices across depth. Patch-to-patch entropy excludes the CLS token and averages row entropy over the patch-to-patch attention submatrix.

4.4. Transfer Evaluation, Statistical Analysis, and Released Artifacts

Zero-shot transfer is measured with CIFAR-100 using class prompts of the form "a photo of a {class name}". Text features are encoded once with the frozen CLIP text tower. Image features are extracted through the active adapted image encoder and visual projection, normalized, and compared with normalized text features by dot product. This adapter-aware path is crucial for LoRA checkpoints because evaluating images through the pretrained image branch would misstate retained transfer performance.

Learning-rate trends over the five-point grid are analyzed with exact permutation correlations. Comparisons among regularization settings are adjusted with the Holm–Bonferroni procedure to control family-wise error, and multi-seed method contrasts are reported with Welch tests and effect sizes [12]. The public repository releases the training code, analysis scripts, manuscript assets, and run-level JSON histories used to regenerate the paper tables and figures, thereby providing the information needed to reproduce the reported matched-learning-rate study.

5. Results

5.1. Learning-Rate Effects

Figure 1 summarizes the controlled matrix. Several consistent patterns emerge.

On EuroSAT, Full FT is highly sensitive to optimization scale. At the smallest learning rate, it mildly broadens attention (+1.83% entropy, +2.42% ERF on average across seeds), but as the learning rate increases it becomes progressively more contractive, reaching −3.99% entropy and −8.25% ERF at 5×10^{-5}. LoRA shows a distinct pattern: it remains entropy-positive across the full matched grid, with mean entropy shifts between +0.68% and +1.50%. The in-domain accuracy gap is largest at the smallest learning rate and narrows substantially at larger values. By 5×10^{-5}, LoRA reaches 98.83% mean validation accuracy, close to the best Full FT regime, while preserving substantially higher CIFAR-100 performance.

Oxford-IIIT Pets presents a more nuanced pattern. Full FT is contractive throughout the grid, with mean entropy shifts ranging from −1.57% to −3.63%. LoRA is structurally milder, ranging from essentially neutral at 10^{-6} (+0.01%) to moderately contractive at 5×10^{-5} (−0.75%). However, low-learning-rate LoRA substantially underfits the task: at 10^{-6} it reaches only 19.13% mean validation accuracy despite retaining near-baseline zero-shot transfer. This underfitting regime explains why method-only averages can be misleading on Pets. Once the learning rate is large enough to optimize the task effectively, LoRA becomes competitive in-domain while retaining substantially more of the pretrained transfer behavior than Full FT. The best Pets LoRA regime in the matched grid (5×10^{-5}) attains 91.91% mean validation accuracy, slightly above the best Pets Full FT regime (91.20% at 10^{-5}), while preserving substantially higher CIFAR-100 accuracy (51.91% versus 5.01%).

 doi:10.20944/preprints202604.0317.v1

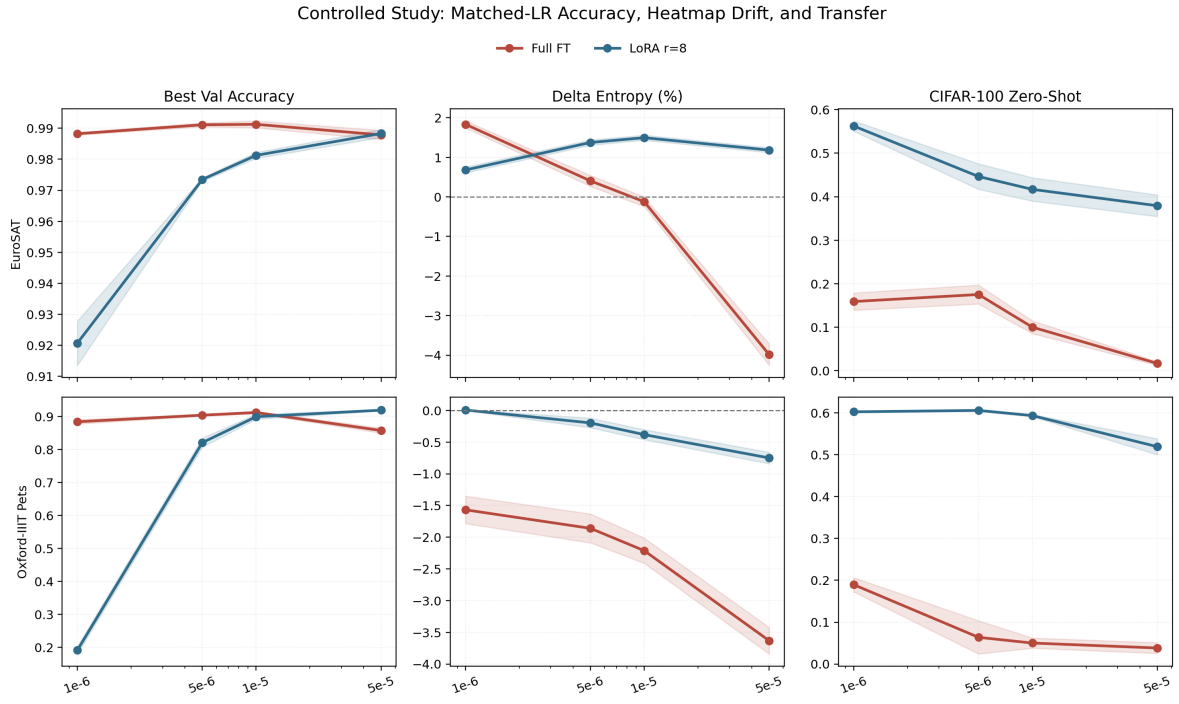

Figure 1. Controlled matched-learning-rate comparison across EuroSAT and Oxford-IIIT Pets. Each panel reports the mean and standard deviation across five seeds for best validation accuracy, CLS-to-patch entropy drift, and adapter-aware CIFAR-100 zero-shot accuracy. The figure shows that learning rate strongly modulates structural drift, and that LoRA generally preserves transfer better than Full FT at matched optimization scales.

Table 1. Aggregate means across the completed matched-learning-rate grid. Each row summarizes 20 runs (four learning rates, five seeds).

Dataset	Method	Mean ΔEntropy (%)	Mean ΔERF (%)	Mean Best Val Acc (%)	Mean CIFAR-100 ZS (%)
EuroSAT	Full FT	-0.47	-1.97	98.96	11.28
EuroSAT	LoRA r=8	+1.18	+1.55	96.59	45.13
Oxford-IIIT Pets	Full FT	-2.32	-5.18	88.94	8.54
Oxford-IIIT Pets	LoRA r=8	-0.33	-1.44	70.76	58.01

Table 1 provides a compact summary over the full grid. On EuroSAT, LoRA is both structurally broader and substantially more transfer-preserving than Full FT, albeit with a modest average in-domain accuracy penalty driven largely by the smallest learning rate. On Pets, the method averages should be interpreted with caution because the 10^{-6} LoRA regime is dominated by underfitting. Even so, the aggregate comparison preserves the same structural contrast: Full FT is markedly more contractive than LoRA.

5.2. Transfer Retention

The completed controlled study also sharpens the relationship between structure and transfer. The pretrained CLIP baseline reaches 60.22% zero-shot accuracy on CIFAR-100. Against that reference point, the mean transfer loss under Full FT is severe across both datasets: 11.28% average CIFAR-100 accuracy over the EuroSAT grid and 8.54% over the Pets grid. LoRA retains much more of the pretrained capability: 45.13% on EuroSAT and 58.01% on Pets.

Importantly, the run-level association between structural preservation and transfer preservation is stronger than the association between structural preservation and in-domain validation accuracy. The entropy-vs-CIFAR correlation is 0.61 on EuroSAT and 0.90 on Pets, whereas the entropy-vs-accuracy correlation is weak to moderate (-0.14 and -0.43, respectively). This pattern should be interpreted cautiously. It does not imply that broader attention support causes better zero-shot transfer, but it

 doi:10.20944/preprints202604.0317.v1

suggests that attention heatmap drift can serve as a descriptive proxy for how much of the pretrained representation is retained during adaptation.

5.3. Layerwise Analysis

Figure 2 visualizes mean per-layer entropy drift over the matched grid. The layerwise view adds an important qualification to the run-level averages. Structural contraction is not uniformly distributed across depth. Under high-learning-rate Full FT, the largest negative shifts accumulate in the later layers, especially on Pets. At 5×10^{-5}, the mean layer-12 entropy shift is -20.29% for Full FT and -11.53% for LoRA. EuroSAT shows the same directional tendency under high-learning-rate Full FT, albeit less strongly.

LoRA does not eliminate drift, but it changes its profile. On EuroSAT, the LoRA heatmaps remain positive across all learning rates, with especially large broadening in the middle and later layers. On Pets, LoRA becomes mildly contractive as the learning rate increases, yet the contraction remains substantially smaller than under Full FT. This distinction supports a more precise interpretation of the adaptation effect: the dominant pattern is not uniform collapse across optimization settings, but pronounced late-layer contraction under aggressive Full FT regimes.

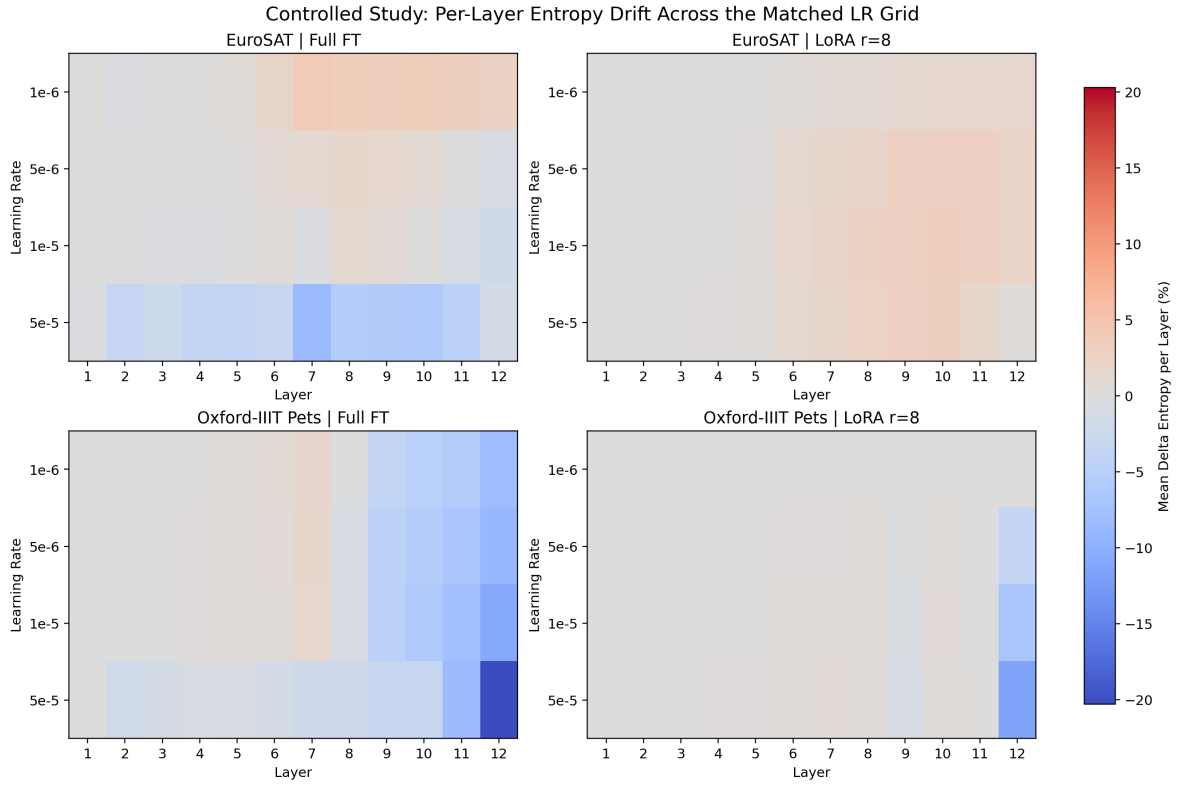

Figure 2. Mean per-layer entropy drift across the matched learning-rate grid. Rows correspond to learning rates and columns correspond to transformer layers. High-learning-rate Full FT produces the strongest late-layer contraction, especially on Oxford-IIIT Pets, whereas LoRA remains structurally milder across the grid.

6. Additional Analyses

The paper also includes earlier analyses that were not part of the strictly matched-learning-rate matrix. These results are not treated as replacements for the controlled matrix, because they mix learning-rate conventions or evaluate additional settings outside the main grid. They are retained because they reinforce the central interpretation and extend the empirical scope of the study.

6.1. Structural Validation

The first set of retained results triangulates the controlled entropy findings with additional structural measurements. Table 2 shows that on EuroSAT, LoRA r=8 produces slightly higher rollout

 doi:10.20944/preprints202604.0317.v1

entropy, larger rollout ERF@0.95, lower rollout Gini, and higher patch-to-patch entropy than Full FT. These auxiliary metrics are consistent with the controlled matched-LR conclusion that LoRA is generally more structurally conservative than Full FT.

The same table also reports a low coefficient of variation across five balanced evaluation subsets, suggesting that the reported entropy measurements are not highly sensitive to the specific subset of images used for the structural analysis. This stability reduces the concern that the observed drift patterns are artifacts of evaluation-image selection.

Representational similarity provides an additional secondary perspective. Table 3 shows that, in the earlier follow-up analysis, higher mean CKA to the pretrained encoder is associated with broader retained attention support ($r = 0.585$ across 39 checkpoints). Runs that remain closer to the pretrained representation geometry also tend to preserve broader attention support. This auxiliary result does not establish mechanism, but it is directionally consistent with the controlled entropy-versus-transfer correlations reported above [13].

Table 2. 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 [13] 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 stratified, class-balanced evaluation subsets drawn from the in-domain validation split.

Model	Rollout Entropy (bits)	Rollout ERF@0.95	Rollout Gini	Patch-to-Patch Entropy (bits)	Mean Layerwise 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 3. Correlation between representational fidelity (CKA to the pretrained model [13]) and structural change (Δentropy).

Analysis	n	Pearson r	Pearson p	Spearman ρ	Spearman p
Run-level mean CKA vs. mean ΔEntropy	39	0.585	0.0003	0.406	0.0106
Strongest per-layer association (layer 4)	39	0.892	0.0000	0.638	0.0000

6.2. Backbone and Regularization

The paper also includes two contextual checks. First, Table 4 shows that backbone choice changes the magnitude and even the sign of structural drift on EuroSAT. On ViT-B/16, Full FT increases entropy where ViT-B/32 is nearly neutral or mildly contractive. This result reinforces the caution implied by the controlled study: structural drift is not uniform across pretrained backbones.

Second, the retained regularization results show that explicit structural penalties can move Full FT toward broader attention support, but the evidence is not yet sufficient to support a central claim of reliable mitigation. Table 5 indicates that APR and entropy-floor regularization can improve entropy relative to the unregularized Full FT baseline, and Table 7 shows that none of the tested regularizers survives family-wise correction in the current analysis [12]. These results are therefore presented as follow-up experiments rather than as central evidence.

For Table 7, the learning-rate correlations are evaluated with exact permutation tests because the matched grid contains only five learning-rate settings, and the regularizer comparisons are adjusted with Holm–Bonferroni across the regularizer family to control family-wise error [12].

At the run level, the same matched-grid pattern is supported by the multi-seed follow-up table: on EuroSAT, Full FT and LoRA r=8 differ significantly in mean entropy ($p = 0.0127$), ERF ($p = 0.0026$), Gini ($p = 0.0368$), head diversity ($p = 0.0430$), and best validation accuracy ($p = 0.0699$ by Welch test); on Pets, the entropy and ERF differences remain significant ($p = 0.0064$ and $p = 0.0228$), as do Gini ($p = 0.0019$) and best validation accuracy ($p = 0.0115$). Accordingly, the principal EuroSAT/Pets method contrast is interpreted as statistically supported rather than exclusively descriptive [12].

 doi:10.20944/preprints202604.0317.v1

Table 4. Backbone comparison between CLIP ViT-B/32 and ViT-B/16 on EuroSAT. Metric cells show B/32 / B/16 values.

Method	ΔEntropy (%)	ΔERF (%)	ΔGini (%)	Task Acc. (%)	Entropy Gap
Full FT on EuroSAT	-0.14 / +3.57	-1.73 / +6.66	-3.54 / -8.83	99.30 / 99.15	+3.72
LoRA r=8 on EuroSAT	+0.58 / -0.59	+0.37 / +4.88	-6.48 / -4.09	98.96 / 98.89	-1.18

Table 5. 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

Table 6. Run-level multi-seed comparisons on mean Δentropy. Values are reported as percentage change relative to the corresponding pretrained baseline, aggregated across completed seed runs.

Comparison	Seeds	Mean ΔEntropy (%)	Welch t	Welch p	Cohen's d
Full FT vs. LoRA r=8 (EuroSAT)	3 vs 3	-0.22 vs +0.87	-7.157	0.0127	-5.844
Full FT vs. LoRA r=8 (Pets)	3 vs 3	-2.10 vs -0.84	-7.969	0.0064	-6.507
Full FT vs. APR $\lambda = 0.1$ (EuroSAT)	3 vs 3	-0.22 vs +0.55	-18.441	0.0020	-15.057
Full FT vs. Entropy floor $\lambda = 0.1$ (EuroSAT)	3 vs 3	-0.22 vs +0.70	-14.439	0.0002	-11.789

Table 7. Corrected inferential statistics.

Comparison	Test	Statistic	p-value
LR vs. ΔEntropy	Exact permutation Pearson r	-0.893	0.0333
LR vs. ΔEntropy	Exact permutation Spearman ρ	-1.000	0.0167
APR $\lambda = 0.01$ vs. baseline	Paired per-layer t-test	2.132	0.2194
Entropy floor $\lambda = 0.1$ vs. baseline	Paired per-layer t-test	2.638	0.1154

6.3. Single-Configuration Results

Some earlier single-configuration results remain informative when interpreted in light of the controlled matrix. In the original EuroSAT rank ablation, LoRA r=4, r=8, and r=16 all produced positive entropy shifts (+0.98%, +0.58%, and +0.44%, respectively) and materially higher adapter-aware CIFAR-100 zero-shot accuracy than the corresponding original Full FT baseline (40.51%, 47.28%, and 39.35% versus 9.43%). These values are not given greater weight than the controlled matched-LR comparison, but they are consistent with the same overall interpretation: lower-rank adaptation tends to retain more of the pretrained structure and transfer behavior than unconstrained Full FT.

7. Discussion

The controlled results, together with the retained auxiliary analyses, support a narrower framing than a broad "attention collapse" characterization. Three conclusions are supported consistently.

 doi:10.20944/preprints202604.0317.v1

First, learning rate is a first-order determinant of attention heatmap drift. The same adaptation method can appear expansive, neutral, or strongly contractive depending on optimization scale. Any claim about method-level structural behavior should therefore be stated jointly with learning rate rather than in isolation.

Second, LoRA is generally more conservative than Full FT under matched optimization scales, especially when transfer retention is part of the evaluation target. This conclusion is strongest on EuroSAT, where LoRA remains entropy-positive throughout the grid and preserves substantially more CIFAR-100 performance. On Pets, the comparison is more conditional because LoRA underfits at very small learning rates, but once that regime is excluded, LoRA remains both competitive in-domain and markedly more transfer-preserving.

Third, the completed results argue for treating attention heatmap drift as a descriptive structural measurement rather than a causal explanation. The positive association between broader retained attention support and better zero-shot retention is informative, and the retained CKA analysis is directionally consistent with that relationship, but neither result by itself establishes mechanism. Transfer degradation likely reflects a broader set of representational changes than any single attention metric can capture.

7.1. Limitations

This study has four main limitations. First, the controlled matrix is restricted to CLIP ViT-B/32; although a ViT-B/16 auxiliary check is retained, the main controlled claims should not be generalized automatically across backbones. Second, the transfer analysis centers on CIFAR-100 as the only primary out-of-domain benchmark. Third, the paper compares Full FT only to LoRA r=8 in the controlled matrix; it does not provide exhaustive coverage of the parameter-efficient fine-tuning design space. Fourth, the auxiliary repository analyses are heterogeneous in design and are therefore used only as supporting evidence, not as substitutes for the controlled matrix.

7.2. Open Research Questions and Future Directions

The present results also expose several questions that remain unresolved. First, the relationship between attention heatmap drift and transfer degradation is correlational rather than causal. Future work should test interventions that directly manipulate layerwise attention support, such as entropy regularization, adapter constraints, or controlled re-initialization, and then determine whether transfer changes accordingly. Second, the generality of the observed patterns beyond the current CLIP ViT-B/32 setting remains open. Matched-learning-rate studies across larger backbones, hybrid architectures, joint image–text adaptation, and other parameter-efficient tuning methods are needed to determine which findings are model-specific and which are robust.

Third, the temporal onset of harmful drift is not yet well understood. Dense checkpoint analysis could reveal whether late-layer contraction appears gradually, emerges abruptly, or provides an early warning signal that can guide stopping or learning-rate adjustment. Fourth, the current transfer evaluation centers on CIFAR-100. Broader out-of-domain benchmarks, retrieval tasks, and multimodal evaluation protocols would clarify when attention heatmap drift predicts practical generalization loss and when it does not. Finally, a more complete theory of adaptation should explain why some forms of structural drift are compatible with improved transfer while others accompany representation damage. Answering that question will require tighter links between internal measurements, optimization dynamics, and downstream generalization.

8. Conclusions

This paper presented a completed 80-run controlled comparison of Full FT and LoRA for CLIP ViT-B/32 under matched learning rates and incorporated supporting analyses from the same experiment set into the overall interpretation. The results show that structural drift in attention heatmaps is strongly modulated by optimization scale, that LoRA is usually more structurally conservative than Full FT under the same learning-rate budget, and that broader retained attention support is associated

 doi:10.20944/preprints202604.0317.v1

with better zero-shot transfer retention. The supporting analyses extend this interpretation without displacing it: rollout, patch-to-patch, and CKA summaries are consistent with the controlled findings, regularization results suggest partial but not yet definitive mitigation, and backbone validation shows that structural drift depends on backbone choice. The overall conclusion is therefore deliberately restrained: attention heatmap drift provides an informative empirical lens on CLIP adaptation when method, learning rate, downstream task, and backbone are analyzed jointly, while the remaining research agenda is to determine how general these drift patterns are and whether they can support controlled intervention rather than descriptive monitoring alone.

Author Contributions: Conceptualization, R.X.; methodology, R.X.; software, R.X.; validation, R.X.; formal analysis, R.X.; investigation, R.X.; data curation, R.X.; writing—original draft preparation, R.X.; writing—review and editing, R.X.; visualization, R.X. 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 assets, and analysis pipeline are available at https://github.com/xiaruize0911/Attention_Collapse_in_CLIP_Fine-tuning_repo.

Acknowledgments: The author acknowledges the open-source software and datasets used in this study.

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

Abbreviations

CKA	centered kernel alignment
CLIP	Contrastive Language-Image Pretraining
CLS	classification token
ERF	effective receptive field
FT	fine-tuning
LoRA	low-rank adaptation
OOD	out-of-domain
ViT	Vision Transformer

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. *arXiv preprint arXiv:2103.00020* **2021**.
2. 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.
3. 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. *arXiv preprint arXiv:2109.01903* **2021**.
4. 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.
5. 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.
6. 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, Volume 1 (Long and Short Papers), 2019, pp. 3543–3556. https://doi.org/10.18653/v1/N19-1357.

 doi:10.20944/preprints202604.0317.v1

7. 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. https://doi.org/10.18653/v1/D19-1002.
8. 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. https://doi.org/10.18653/v1/2020.acl-main.385.
9. 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.
10. 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.
11. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.
12. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **1979**, *6*, 65–70.
13. Kornblith, S.; Norouzi, M.; Lee, H.; Hinton, G. Similarity of Neural Network Representations Revisited. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning, 2019, Vol. 97, *Proceedings of Machine Learning Research*, pp. 3519–3529.