FROM UNIFORM CAPACITY TO FUNCTIONAL ROLES: EFFICIENT VISUAL LORA FOR SOURCE-FREE CROSS-DOMAIN FEW-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Source-Free Cross-Domain Few-Shot Learning (SF-CDFSL) aims to rapidly adapt pre-trained vision-language models to a target domain using only a few samples, without access to source domain data. However, existing vision LoRA methods typically employ a uniform Q/K/V adaptation structure across different Transformer layers, overlooking the distinct functional roles played by layers at different depths. This paper reveals an intriguing phenomenon regarding vision LoRA in SF-CDFSL: despite being assigned identical adaptation capacities, the layers contribute highly unevenly to the final decision. Specifically, shallow layers have a minimal impact, middle-to-deep layers rely primarily on the V-path, and final layers exhibit significant QK sensitivity. We investigate and explain this phenomenon; through counterfactual interventions on identical hidden states, we attribute this behavior to a stable functional division of labor emerging among Transformer depths and QK/V projection paths during target domain adaptation. Based on this insight, we propose Role-Aligned LoRA. This method aligns adaptation capacity with functional roles by freezing low-impact layers, retaining V-LoRA in V-sensitive regions, QKV-LoRA in transitional layers, and QK-LoRA in final layers. Experiments across four SF-CDFSL datasets validate our explanation and the proposed method. Results demonstrate that Role-Aligned LoRA maintains overall accuracy on CLIP while utilizing only 27.8% of the trainable parameters found in Uniform-QKV LoRA, and reduces support-set adaptation time by 34.3%.

1 INTRODUCTION

Source-free cross-domain few-shot learning (SF-CDFSL) adapts a pretrained model to a target domain from only a few labeled support examples, without revisiting source data. Vision–language models (VLMs) such as CLIP Radford et al. (2021) provide a transferable cross-modal decision space, and recent work increasingly asks *what* should be adapted to preserve it Zhang et al. (2026b;a); Zhao et al. (2026); Yi et al. (2026a;b). Visual LoRA Hu et al. (2022); Zanella & Ben Ayed (2024) is a strong parameter-efficient baseline, but its common Uniform-QKV form inserts the same Q/K/V adaptation into every visual Transformer block. This assumes that uniformly allocated capacity serves comparable functions across depth. **We find the opposite: uniform capacity gives rise to non-uniform decision function.**

Figure 1 summarizes the mismatch. Early blocks have small total effect, middle/deep blocks are predominantly V-sensitive, and the terminal block becomes strongly QK-sensitive; the transition is clearest on EuroSAT and CropDisease and weaker on ISIC and ChestX. The key question is therefore not simply which LoRA weights are large, but which adapted computation changes the final text-defined decision. Representation geometry and familiar local observables are insufficient for this purpose: even complete within-visual Gram geometry can be preserved while fixed text-defined logits change, and patch–text similarity, token norm, or CLS-to-patch attention remain weakly related to intervention effect. We address this with same-hidden-state counterfactuals. At a fixed input to one visual block, the adapted Q/K path and adapted V path are independently restored to their frozen counterparts while the rest of the network remains adapted. The resulting QK-sensitive and V-sensitive effects characterize changes in attention routing and propagated values while retaining

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.

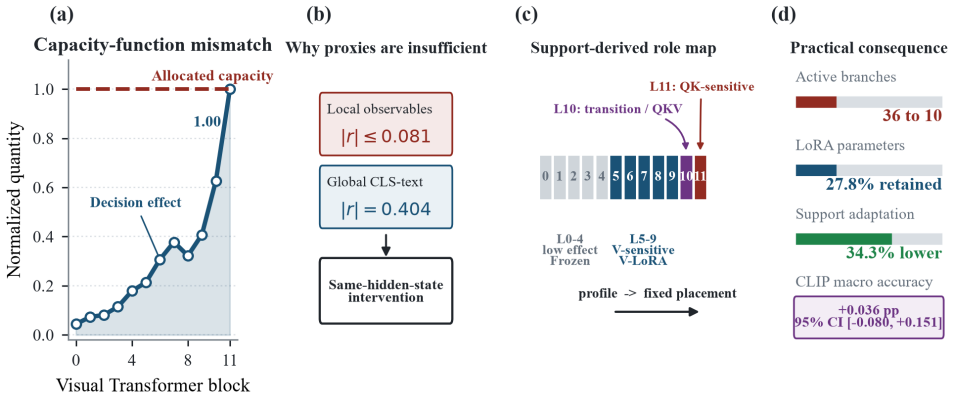

Figure 1: (a) Uniform LoRA capacity versus depth-varying decision effects. (b) Weak proxy–intervention correlation. (c) Support-derived functional roles. (d) Role-Aligned placement and accuracy–efficiency trade-off.

their interaction. Support-side profiles recover the same coarse organization observed on query data. Role-Aligned LoRA uses this organization to form a profile-guided sparse instantiation: freeze the low-effect prefix, retain V-LoRA in the V-sensitive region, keep QKV-LoRA at the mixed transition, and use QK-LoRA at the terminal block. On Base-only CLIP, this placement reduces active branches from 36 to 10 and trainable LoRA parameters from 884,736 to 245,760, while macro accuracy changes by $+0.036$ pp and support adaptation time falls by 34.3%. Matched low-budget controls compare against dense low-rank, deep-only, and random sparse alternatives; the CLIP-derived assignment transfers unchanged to PE-Core with a $+0.159$ pp macro change at the same parameter ratio. Our contributions are threefold:

- We expose a mismatch between uniform visual-LoRA capacity and heterogeneous decision-level function across Transformer depth.
- We introduce same-hidden-state QK/V interventions that reveal consistent support- and query-side structure in contribution magnitude and projection specialization.
- We instantiate this profile as a sparse role-aware layout and evaluate its accuracy–parameter–time trade-off across two 12-block visual encoders in the studied SF-CDFSL setting.

2 Delving into Functional Roles of Visual LoRA

2.1 Preliminaries

Source-free cross-domain few-shot episode. We consider an N-way K-shot target-domain episode with label set $\mathcal{Y} = \{1, \dots, N\}$, support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{NK}$, and query set $\mathcal{Q}$. The support set provides the labeled examples used for episode adaptation, whereas the query set is reserved for evaluation. A pretrained vision–language model contains a visual encoder $f_v(\cdot; \theta_v)$ and a text encoder $f_t(\cdot; \theta_t)$. Throughout target adaptation, the pretrained encoders remain frozen and trainable parameters are introduced only through visual LoRA updates.

Text-defined VLM decision. For class c with textual description p_c, the normalized text prototype and adapted visual representation are

$$\boldsymbol{t}_c = \frac{f_t(p_c; \theta_t)}{\|f_t(p_c; \theta_t)\|_2}, \qquad \boldsymbol{z}_v(x) = \frac{f_v(x; \theta_v, \Delta\theta)}{\|f_v(x; \theta_v, \Delta\theta)\|_2}. \tag{1}$$

The fixed text prototypes therefore define the classifier,

$$s_c(x) = \gamma\, \boldsymbol{z}_v(x)^\top \boldsymbol{t}_c, \qquad \hat{y} = \arg\max_{c \in \mathcal{Y}} s_c(x), \tag{2}$$

where γ is the logit scale. For any labeled example (x, y), we summarize separation in this text-defined decision space by the margin

$$m(x, y) = s_y(x) - \max_{c \neq y} s_c(x). \tag{3}$$

A positive change in this margin therefore means that an intervention increases separation of the correct text prototype from its strongest competitor.

Uniform visual LoRA. For a projection matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, LoRA keeps $\boldsymbol{W}$ frozen and learns a rank-r residual,

$$\boldsymbol{W}' = \boldsymbol{W} + \frac{\alpha}{\sqrt{r}}\boldsymbol{B}\boldsymbol{A}, \qquad \boldsymbol{A} \in \mathbb{R}^{r\times d},\ \boldsymbol{B} \in \mathbb{R}^{d\times r}. \tag{4}$$

Uniform-QKV baseline applies this same adaptation form to the Q, K, and V projections of every visual Transformer block, yielding three LoRA branches per block.

2.2 Phenomenon: Equal Capacity, Unequal Function

Uniform QKV is architecturally flat by construction: every block receives Q-, K-, and V-LoRA with the same rank and scaling. Its learned decision contribution is not. When a trained block is counterfactually reverted to its base Q/K/V computation while the remainder of the adapted network is held fixed, the mean absolute full-block effect is 8.53×10^{-4} over L0–4, 2.82×10^{-3} over L5–9, 5.43×10^{-3} at L10, and 8.67×10^{-3} at L11. Thus, the terminal block changes the text-defined margin by more than an order of magnitude more than the low-effect early prefix, despite receiving the same LoRA capacity. Equal adaptation capacity therefore produces unequal decision functions. A useful diagnostic must identify both *how much* a block matters and *which projections* support that effect.

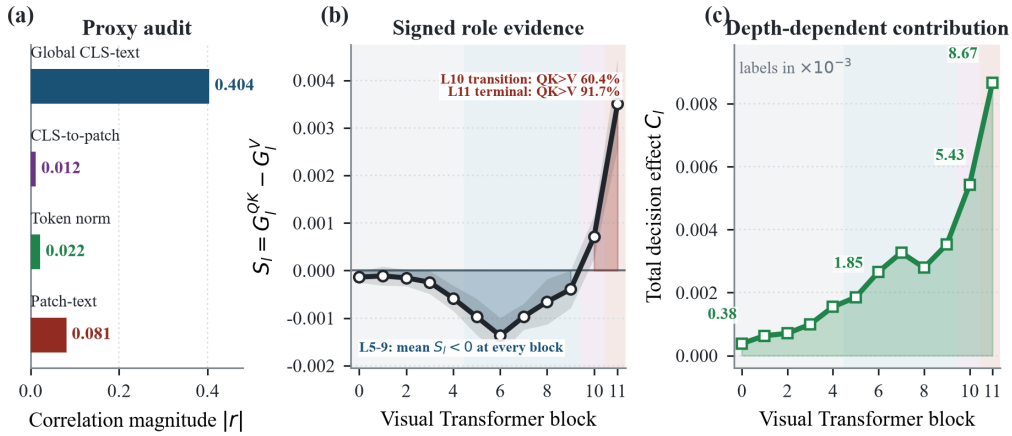

Figure 2: (a) Correlation between observational proxies and intervention contribution. (b) Layer-wise QK/V specialization S_l. (c) Complete-block contribution C_l across Transformer depth.

2.3 Interpretation: Functional Roles Are Decision-Level Properties

Representation geometry, attention, and token magnitude are natural candidates for adapter attribution, but they observe the visual representation rather than its consequence for the fixed text classifier. We first formalize this limitation, then intervene directly on QK and V adaptation.

2.3.1 Visual geometry does not determine the text-defined decision.

Let $\boldsymbol{F} \in \mathbb{R}^{M\times d}$ collect normalized visual features for M examples and $\boldsymbol{T} \in \mathbb{R}^{N\times d}$ collect normalized class text prototypes. The within-visual geometry is $\boldsymbol{F}\boldsymbol{F}^\top$, whereas the classifier uses $\boldsymbol{F}\boldsymbol{T}^\top$.

Proposition 1 (Visual Gram preservation does not imply decision preservation)**.** *Let $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ be orthogonal and $\boldsymbol{F}' = \boldsymbol{F}\boldsymbol{R}$. Then $\boldsymbol{F}'\boldsymbol{F}'^\top = \boldsymbol{F}\boldsymbol{F}^\top$, while the text-defined logits are preserved if and only if*

$$\boldsymbol{F}(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{T}^\top = \boldsymbol{0}. \tag{5}$$

Consequently, preserving the complete visual Gram geometry does not in general preserve the VLM decision.

We provide a proof in Appendix B. The proposition isolates a geometric limitation: preserving visual geometry alone does not certify preservation of a fixed text classifier, motivating decision-level attribution.

2.3.2 SAME-HIDDEN-STATE QK/V COUNTERFACTUALS.

For block l, let $\boldsymbol{H}^{(l)}$ denote the token sequence entering self-attention,

$$\boldsymbol{O}^{(l)} = \boldsymbol{A}^{(l)}\boldsymbol{V}^{(l)}, \qquad \boldsymbol{A}^{(l)} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}^{(l)}{\boldsymbol{K}^{(l)}}^{\top}}{\sqrt{d_h}}\right). \tag{6}$$

At fixed $\boldsymbol{H}^{(l)}$, Q/K change the attention mixing weights, while V changes the vectors propagated under those weights. We therefore use *QK-sensitive* and *V-sensitive* effects as operational measures of routing/selection and content/value influence while allowing interactions between the two paths. For a trained Uniform-QKV model, we distinguish between the LoRA-adapted projections, denoted by A, and the frozen pretrained base projections, denoted by B. Crucially, both versions are evaluated from the same hidden-state input $\boldsymbol{H}^{(l)}$, so that the intervention changes only the projection computation at the target block rather than its incoming representation.

We write $\mathcal{F}^{(l)}_{ab}$ for a counterfactual target-block computation, where the first subscript $a \in \{A, B\}$ specifies the source of the Q/K projections and the second subscript $b \in \{A, B\}$ specifies the source of the V projection. This gives four states:

$$\mathcal{F}^{(l)}_{AA} : \boldsymbol{A}(\boldsymbol{Q}_A, \boldsymbol{K}_A)\boldsymbol{V}_A, \qquad \text{adapted Q/K, adapted V,} \tag{7}$$

$$\mathcal{F}^{(l)}_{BA} : \boldsymbol{A}(\boldsymbol{Q}_B, \boldsymbol{K}_B)\boldsymbol{V}_A, \qquad \text{base Q/K, adapted V,} \tag{8}$$

$$\mathcal{F}^{(l)}_{AB} : \boldsymbol{A}(\boldsymbol{Q}_A, \boldsymbol{K}_A)\boldsymbol{V}_B, \qquad \text{adapted Q/K, base V,} \tag{9}$$

$$\mathcal{F}^{(l)}_{BB} : \boldsymbol{A}(\boldsymbol{Q}_B, \boldsymbol{K}_B)\boldsymbol{V}_B, \qquad \text{base Q/K, base V.} \tag{10}$$

All non-target blocks remain in their adapted state. Let $M^{(l)}_{ab}$ denote the mean decision margin obtained under $\mathcal{F}^{(l)}_{ab}$. We then define

$$G^{(l)}_{QK} = M^{(l)}_{AB} - M^{(l)}_{BB}, \qquad G^{(l)}_{V} = M^{(l)}_{BA} - M^{(l)}_{BB}, \tag{11}$$

$$G^{(l)}_{\text{full}} = M^{(l)}_{AA} - M^{(l)}_{BB}, \qquad S_l = G^{(l)}_{QK} - G^{(l)}_{V}. \tag{12}$$

Here, $G^{(l)}_{QK}$ measures the decision-margin change obtained by restoring only the adapted Q/K path while keeping V at its base value, whereas $G^{(l)}_{V}$ measures the corresponding change obtained by restoring only the adapted V path while keeping Q/K at their base values. $G^{(l)}_{\text{full}}$ measures the joint adapted effect of Q/K and V.

Accordingly, $S_l > 0$ indicates stronger QK sensitivity, whereas $S_l < 0$ indicates stronger V sensitivity under this intervention. Because the QK and V effects are not assumed to be additive, we interpret S_l as a comparative operational diagnostic rather than a complete causal factorization; the corresponding interaction residual is quantified in Appendix C.

2.4 EXPERIMENTAL EVIDENCE FOR FUNCTIONAL HETEROGENEITY

We test whether observational proxies recover intervention contribution and whether contribution magnitude and QK/V direction vary systematically with depth.

2.4.1 OBSERVATIONAL SURROGATES DO NOT RECOVER INTERVENTION CONTRIBUTION.

Let $\boldsymbol{h}^L_i$ and $\boldsymbol{h}^B_i$ be the adapted and base states of token i. Replacing only token i by its base counterpart gives the decision-level target

$$\Delta m_i = m_L(x) - m_{L \leftarrow B(i)}(x). \tag{13}$$

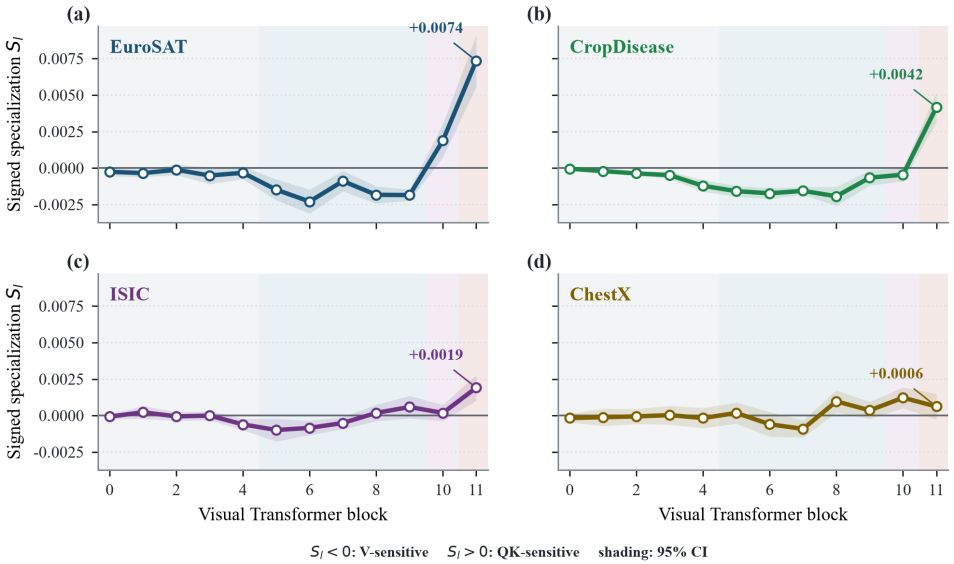

Figure 3: **Domain-wise QK/V specialization across Transformer depth.** Mean S_l with 95% confidence intervals; negative values indicate V sensitivity, and positive values indicate QK sensitivity.

As shown in Fig. 2(a), true-class patch–text similarity reaches only $|r| \approx 0.081$, token norm ≈ 0.022, and CLS-to-patch attention ≈ 0.012 with Δm_i; even the strongest tested global CLS–text observable reaches only ≈ 0.404. These measurements do not reliably identify the local update supporting the final decision.

2.4.2 Depth changes both contribution magnitude and role direction.

Figure 2(b,c) exposes two distinct structures. The total block effect grows strongly toward the end of the visual Transformer, while the aggregate signed role diagnostic is predominantly V-sensitive through the middle/deep region and turns QK-sensitive near the output, with the strongest reversal at L11. Across the 48 dataset–shot–text–seed profiles available at each layer, $P(G_{QK} > G_V)$ is 16.7%, 14.6%, 18.8%, 37.5%, and 41.7% at L5–9, rises to 60.4% at L10, and reaches 91.7% at L11. The same network therefore exhibits two forms of heterogeneity: *how much* a block changes the decision and *which projection family* contributes more strongly. Scalar importance alone cannot express this directional distinction.

2.4.3 The role pattern replicates across domains, with visible boundaries.

Figure 3 resolves the specialization profile by target domain. EuroSAT and CropDisease show the clearest transition: middle/deep blocks are predominantly V-sensitive, followed by strong terminal QK sensitivity. ISIC shows a weaker positive terminal shift, whereas ChestX remains more variable across depth and lacks a clear L11 QK-dominant pattern. Within these four target domains, the decomposition reveals a shared coarse depth organization with substantial variation in directional strength. This supports a stable region-level profile within the studied domain set, with domain-dependent directional strength, which motivates Sec. 3.

3 Method

Section 2 reveals that visual LoRA exhibits heterogeneous decision contribution and QK/V specialization across Transformer depth. Based on this observation, Role-Aligned LoRA summarizes each visual block by a support-derived functional profile and uses the resulting depth-wise organization to construct a sparse, projection-specific LoRA placement (Fig. 4).

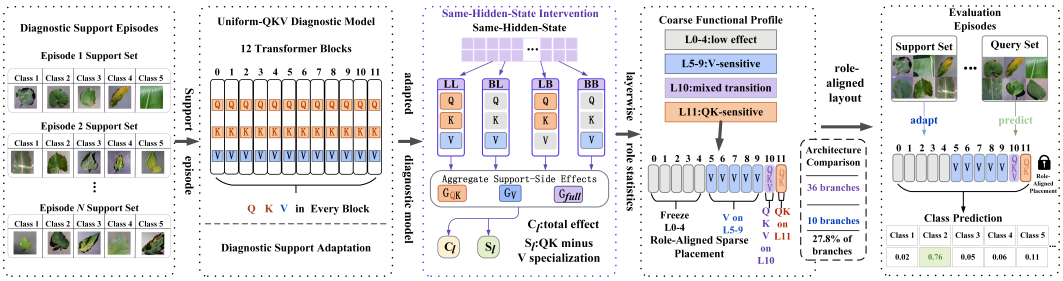

Figure 4: **Overview of Role-Aligned LoRA.** Same-hidden-state counterfactuals yield per-block contribution C_l and specialization S_l, which define the sparse placement: L0–4 frozen, L5–9 V-LoRA, L10 QKV-LoRA, and L11 QK-LoRA. The resulting 10-branch model retains 27.8% of Uniform-QKV trainable parameters.

3.1 Functional-Role Discovery

After Uniform-QKV adaptation on diagnostic episodes, we apply the same-hidden-state counterfactuals from Sec. 2.3.2 to the support examples and summarize each visual block l by

$$\mathcal{R}_l = \Big(C_l, S_l\Big), \qquad C_l = \mathbb{E}\Big[\Big|G^{(l)}_{\text{full}}\Big|\Big], \quad S_l = \mathbb{E}\Big[G^{(l)}_{QK} - G^{(l)}_{V}\Big]. \tag{14}$$

Here, C_l measures the decision contribution of the adapted block, while S_l characterizes its projection specialization: $S_l < 0$ indicates stronger V/content sensitivity, whereas $S_l > 0$ indicates stronger QK/routing sensitivity. Together, (C_l, S_l) separates two questions that a scalar importance score would conflate: *where does adaptation matter?* and *which projection family carries the dominant effect?*

3.2 Profile-Guided Sparse Instantiation

The aggregate support profile reveals four contiguous regimes in the studied 12-block visual Transformer: a low-effect prefix (L0–4), a V-sensitive middle region (L5–9), a mixed transition block (L10), and a QK-sensitive terminal block (L11). We instantiate each regime with the smallest projection set consistent with its observed functional role: low-effect blocks are frozen, V-sensitive blocks retain only V-LoRA, the mixed block retains QKV-LoRA, and the terminal QK-sensitive block retains Q- and K-LoRA. Let $\mathcal{P}_l \subseteq \{Q, K, V\}$ denote the active projections in visual block l. The resulting Role-Aligned placement is

$$\mathcal{P}_l^{\text{RA}} = \begin{cases} \varnothing, & l = 0, \ldots, 4, \\ \{V\}, & l = 5, \ldots, 9, \\ \{Q, K, V\}, & l = 10, \\ \{Q, K\}, & l = 11. \end{cases} \tag{15}$$

This placement reduces the number of active LoRA branches from 36 under Uniform-QKV to 10, while preserving the same rank and optimization setting for every retained branch. The resulting model therefore changes only *where* LoRA is applied and *which* projections are adapted.

4 Experiments

4.1 Datasets

Following established Cross-Domain Few-Shot Learning (CDFSL) benchmarks, we evaluate our method on four challenging target domains: CropDiseases Mohanty et al. (2016), EuroSAT Helber et al. (2019), ISIC2018 Codella et al. (2019), and ChestX Wang et al. (2017). These datasets span agriculture, remote sensing, and medical imaging, and exhibit substantial domain shifts from generic source-domain data.

Table 1: **Validation of the support-derived functional profile.** Support/query statistics compare the aggregate 12-layer profiles. Aligned–Reversed tests whether the measured QK/V direction is informative for placement.

Validation statistic	Result
Specialization correlation	Pearson 0.962 / Spearman 0.734
Complete-effect correlation	Pearson 0.996
Specialization sign agreement	12/12 blocks
LODO: L5–9 mean $S_l < 0$	4/4 held-out domains
LODO: L11 $P(G_{QK} > G_V)$	88.9–100%
Aligned − Reversed, EuroSAT	$+2.58$ pp
Aligned − Reversed, CropDisease	$+1.69$ pp
Aligned − Reversed, ISIC	-0.11 pp
Aligned − Reversed, ChestX	$+0.13$ pp
EuroSAT/CropDisease aggregate	$+2.13$ pp; margin wins 120/120

Table 2: **Primary Base-only accuracy–efficiency results on CLIP.** Accuracy (%) is reported for all eight dataset–shot conditions. Params denotes trainable LoRA parameters and Time is the mean support-set adaptation time per episode.

	1-shot				5-shot				Overall		
Method	Euro	Crop	ISIC	ChestX	Euro	Crop	ISIC	ChestX	Macro	Params (K)	Time (s)
Uniform-QKV	82.38	85.16	36.73	21.78	93.69	96.61	53.07	24.07	61.688	884.7	17.20
Role-Aligned	83.37	85.97	35.85	21.75	93.43	96.59	52.31	24.54	61.724	245.8	11.30

4.2 Implementation Details

We use CLIP with a ViT-B/16 visual encoder as the primary backbone and adapt it directly to the target domain without access to source-domain data. Following standard episodic practice, support images are augmented during adaptation. We evaluate under 1-shot and 5-shot settings, reporting mean classification accuracy over 800 episodes for 1-shot and 400 episodes for 5-shot. The optimization protocol follows CLIP-LoRA Zanella & Ben Ayed (2024), with complete hyperparameter settings provided in Appendix F.

4.3 Main Results

Support-side profiling recovers stable functional roles. As shown in Table 1, the support-derived profile closely matches the query-side organization: specialization reaches a Pearson correlation of 0.962, complete-block effect reaches 0.996, and the specialization signs agree across all 12 blocks. The regional structure is also stable under leave-one-domain-out evaluation: L5–9 remain V-sensitive in all four held-out settings, while L11 is QK-dominant in 88.9–100% of the remaining profiles. Moreover, Role-Aligned outperforms Role-Reversed by 2.13 pp on EuroSAT/CropDisease and yields the larger decision margin in all 120 paired episodes, showing that the measured projection direction is predictive of useful placement. Full layerwise statistics and direction controls are provided in Appendices C–D.

Role-Aligned preserves accuracy while reducing adaptation cost. Table 2 reports the primary Base-only CLIP results. Role-Aligned reduces trainable LoRA parameters from 884.7K to 245.8K (27.8% retained) and mean support adaptation time from 17.20 s to 11.30 s (-34.3%), while macro accuracy changes only from 61.688% to 61.724% ($+0.036$ pp; 95% CI $[-0.080, +0.151]$). The condition-level results are domain-dependent: averaging 1-/5-shot differences gives $+0.36$ pp on EuroSAT, $+0.39$ on CropDisease, -0.82 on ISIC, and $+0.22$ on ChestX. This mirrors Fig. 3, where EuroSAT and CropDisease exhibit the clearest V-to-QK specialization and ISIC forms the main boundary of the shared sparse layout.

Table 3: **Matched low-budget placement controls.** All variants use the same independent Base-only evaluation stream. Δ = Role-Aligned minus the listed comparator.

Variant	Params	Param.%	Macro acc.	Time (s)
Uniform QKV $r16$	884,736	100.0	61.577	18.18
Role-Aligned $r16$	245,760	27.8	61.564	12.18
Uniform QKV $r4$	221,184	25.0	61.198	18.19
Uniform QKV $r5$	276,480	31.3	61.268	18.10
Deep-4 QKV $r16$	294,912	33.3	60.335	10.66
Random-10 #202	245,760	27.8	61.242	14.88

Table 4: **PE-Core transfer with the unchanged CLIP-derived layout.** Differences are Role-Aligned minus Uniform-QKV. Per-condition intervals are paired 95% confidence intervals.

Dataset	Shot	Uniform	Aligned	Δ pp	95% CI
EuroSAT	1	85.877	86.755	$+0.878$	±0.240
EuroSAT	5	94.930	95.223	$+0.293$	±0.196
CropDisease	1	92.457	92.743	$+0.287$	±0.151
CropDisease	5	97.860	97.950	$+0.090$	±0.114
ISIC	1	44.688	44.750	$+0.062$	±0.312
ISIC	5	60.640	60.457	-0.183	±0.396
ChestX	1	22.262	22.283	$+0.022$	±0.337
ChestX	5	24.397	24.223	-0.173	±0.567
Macro, 8 conditions		65.389	65.548	$+0.159$	$[+0.047, +0.273]$

The gain is not explained by parameter reduction alone. Table 3 compares Role-Aligned with matched low-budget alternatives. Reducing Uniform-QKV to rank 4 or 5 lowers the parameter count but leaves all 36 branches active, resulting in adaptation times near 18 s and lower accuracy. Deep-4 is faster (10.66 s) but loses 1.229 pp. Most importantly, Random-10 uses exactly the same parameter budget as Role-Aligned, yet is 0.321 pp less accurate and slower (14.88 vs. 12.18 s). Thus, the trade-off depends not only on sparsity, but also on *where* branches are retained and *which* projections are adapted. Additional boundary and budget controls are reported in Appendix D.

The CLIP-derived role map transfers to PE-Core. We transfer the same 12-block assignment to PE-Core without re-estimating its layer boundaries. Macro accuracy increases from 65.389% to 65.548% ($+0.159$ pp; 95% CI $[+0.047, +0.273]$), while support adaptation time decreases from 22.05 s to 15.21 s (-31.0%) at the same 27.8% parameter ratio. As summarized in Table 4, gains concentrate on EuroSAT and CropDisease, while ISIC and ChestX remain near zero or slightly negative in 5-shot, reproducing the same domain boundary observed with CLIP. PE-Core also reproduces terminal QK sensitivity at L11; the complete intervention and Role-Reversed controls are reported in Appendix E.

The role map is robust to neighboring layer boundaries. We further perturb the frozen/V-LoRA boundary by one block while keeping L10 as QKV-LoRA and L11 as QK-LoRA. As shown in Table 5, neighboring layouts form a smooth accuracy–time–capacity trade-off rather than exhibiting

Table 5: **One-block frozen/V boundary perturbation.** Only the boundary between the frozen prefix and V-LoRA region is shifted; L10 and L11 remain QKV- and QK-LoRA, respectively. Results use a separate two-seed robustness suite.

Frozen	V-LoRA	Branches	Param.%	Accuracy	Time (s)
L0–3	L4–9	11	30.6	61.150	12.03
L0–4	L5–9	10	27.8	60.758	11.38
L0–5	L6–9	9	25.0	60.500	10.67

a sharp collapse. This supports a regional interpretation of the functional organization, rather than dependence on a single precisely selected layer boundary.

Profiling overhead. The functional profile is obtained once from 240 diagnostic runs, requiring 4,432 s (1.23 GPU-hours) on a single RTX 3090. With the measured 5.91 s reduction in Base adaptation time per episode, this one-time cost is amortized after approximately 751 adaptations. A detailed cost breakdown is provided in Appendix F.

5 Related Work

Source-free cross-domain few-shot adaptation. CDFSL studies adaptation from a handful of target examples under substantial domain shift Guo et al. (2020); Tseng et al. (2020); the source-free setting removes access to source data and places greater weight on pretrained structure. VLMs provide a transferable image–text space Radford et al. (2021); Jia et al. (2021); Zhai et al. (2022; 2023); Yu et al. (2022), which prompt learning, cache/adaptor methods, linear-probe refinements, and visual LoRA exploit under scarce supervision Zhou et al. (2022b;a); Khattak et al. (2023a;b); Yao et al. (2023); Zhang et al. (2022); Tang et al. (2024); Huang et al. (2024); Li et al. (2025); Zanella & Ben Ayed (2024); Xu et al. (2026). Recent SF-CDFSL work increasingly studies *what* should be adapted, including discriminability shortcuts, under-used text layers, cross-modal alignment, and attention allocation Zhang et al. (2026b;a); Zhao et al. (2026); Yi et al. (2026a;b). We instead ask how low-rank Q/K/V updates inside the visual Transformer divide responsibility for the final cross-modal decision.

Parameter-efficient adaptation and selective capacity allocation. PEFT restricts adaptation to small trainable subspaces, including adapters, prompt/prefix tuning, sparse bias updates, activation scaling, and LoRA Houlsby et al. (2019); Li & Liang (2021); Lester et al. (2021); Ben Zaken et al. (2022); Liu et al. (2022); Hu et al. (2022). LoRA variants redistribute capacity through dynamic rank, shared factors, magnitude–direction decomposition, or asymmetric optimization Zhang et al. (2023); Valipour et al. (2023); Kopiczko et al. (2024); Liu et al. (2024); Hayou et al. (2024); other work removes lower-layer adapters or prunes adapter/LoRA capacity Rücklé et al. (2021); He et al. (2022); Guo et al. (2021); Zhou et al. (2025). These methods mainly use rank, magnitude, optimization, or output importance. Our placement is instead driven by intervention-measured functional role.

Transformer functional analysis. Vision Transformers exhibit depth-wise organization Dosovitskiy et al. (2021); Touvron et al. (2021); Raghu et al. (2021); Park & Kim (2022), while attention weights alone need not provide faithful causal attribution Jain & Wallace (2019). Rollout/relevance methods and intervention-based analyses probe model decisions more directly Abnar & Zuidema (2020); Chefer et al. (2021); Vig et al. (2020). We apply this principle to task-induced visual LoRA residuals, holding a block's hidden-state input fixed while separating QK- and V-sensitive effects on decisions. Appendix A provides extended context and implementation details.

6 Conclusion

Uniform visual LoRA is uniform in capacity but not in decision function. Same-hidden-state interventions reveal depth-dependent contribution magnitude and QK/V specialization, while support/query agreement and direction reversal substantiate the resulting functional profile. Role-Aligned LoRA instantiates this organization as a 10-branch layout retaining only 27.8% of the trainable parameters of Uniform-QKV. On Base-only CLIP, macro accuracy remains essentially unchanged while per-episode support adaptation time decreases by 34.3%. Matched low-budget controls, neighboring-boundary perturbations, and cross-backbone transfer further support functional role as a practical signal for sparse visual-LoRA allocation. Whether the same regional organization extends to visual Transformers with different depths or to unseen domain families remains an open question.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research-methodology and experiment-design feedback, code and analysis refinement, interpretation of experimental results, and LaTeX formatting. We reviewed all AI-assisted work. We executed and checked all AI-assisted code and analyses ourselves, and we reviewed all numerical results and scientific claims for correctness and consistency with the reported experiments. We take responsibility for the final content of this work, including text, scientific claims, code, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We support reproducibility through the main text, appendix, and an anonymized code release. The main text specifies the functional-role measurements, the Role-Aligned sparse instantiation, the evaluation protocol, and the parameter and runtime accounting. The appendix provides the complete theoretical proof, full experimental analyses and controls, implementation and dataset details, and pseudocode. An anonymized implementation, experiment scripts, and step-by-step reproduction instructions are available at `https://anonymous.4open.science/r/RoleAligned_CLIP_LoRA`.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pp. 1–9, 2022. doi: 10.18653/v1/2022.acl-short.1.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60884–60937, 2025. doi: 10.52202/085713-2036.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

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

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Demi Guo, Alexander M. Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pp. 4884–4896, 2021.

Yunhui Guo, Noel C. Codella, Leonid Karlinsky, James V. Codella, John R. Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, 2020.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning*, pp. 17783–17806, 2024.

Shwai He, Liang Ding, Daize Dong, Miao Zhang, and Dacheng Tao. SparseAdapter: An easy approach for improving the parameter-efficiency of adapters. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 2184–2190, 2022. doi: 10.18653/v1/2022.findings-emnlp.160.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with CLIP for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25411–25421, 2025.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pp. 4582–4597, 2021.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, 2022.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, pp. 32100–32121, 2024.

Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016. doi: 10.3389/fpls.2016.01419.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, 2021.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. AdapterDrop: On the efficiency of adapters in transformers. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7930–7946, 2021.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems*, 2020.

Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. AMU-Tuning: Effective logit bias for CLIP-based few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23323–23333, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357, 2021.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2020.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Sakenis, and Jason Huang. Causal mediation analysis for interpreting neural nlp: The case of gender bias. In *Advances in Neural Information Processing Systems*, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017. doi: 10.1109/CVPR.2017.369.

Huali Xu, Li Liu, Tianpeng Liu, Shuaifeng Zhi, Shuzhou Sun, and Ming-Ming Cheng. Step-wise distribution-aligned style prompt tuning for source-free cross-domain few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1098–1114, 2026. doi: 10.1109/TPAMI.2025.3610039.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6757–6767, 2023.

Shuai Yi, Yixiong Zou, Yuhua Li, and Ruixuan Li. Improving CLIP adaptation by breaking tail alignment for source-free cross-domain few-shot learning. In *International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a.

Shuai Yi, Yixiong Zou, Yuhua Li, and Ruixuan Li. Addressing exacerbated attention sink for source-free cross-domain few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29494–29503, 2026b.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1593–1603, 2024.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510, 2022.

Zhenyu Zhang, Guangyao Chen, Yixiong Zou, Yuhua Li, and Ruixuan Li. Reclaiming lost text layers for source-free cross-domain few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15080–15090, 2026a.

Zhenyu Zhang, Yixiong Zou, Yuhua Li, Ruixuan Li, and Guangyao Chen. Mind the discriminability trap in source-free cross-domain few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36978–36988, 2026b.

Yaze Zhao, Yixiong Zou, Yuhua Li, and Ruixuan Li. Interpretable cross-domain few-shot learning with rectified target-domain local alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41605–41615, 2026.

Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao, and Muyun Yang. LoRA-drop: Efficient LoRA parameter pruning based on output evaluation. In *International Conference on Computational Linguistics*, pp. 5530–5543, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

APPENDIX

A DETAILED RELATED WORK

This section provides additional context for three lines of work closely related to Role-Aligned LoRA: source-free cross-domain few-shot adaptation, parameter-efficient capacity allocation, and functional analysis of Transformer computation.

A.1 SOURCE-FREE CROSS-DOMAIN FEW-SHOT ADAPTATION

CDFSL studies few-shot target adaptation under domain shift, with early work combining episodic/meta-learning and transfer-learning strategies and later benchmarks exposing how strongly performance depends on the type of visual shift Tseng et al. (2020); Guo et al. (2020). Source-free variants remove access to source-domain data, making target adaptation rely on the prior structure of the pretrained model rather than on renewed source training. This setting has become especially natural for VLMs: CLIP Radford et al. (2021), ALIGN Jia et al. (2021), LiT Zhai et al. (2022), SigLIP Zhai et al. (2023), CoCa Yu et al. (2022), and PE-Core Bolya et al. (2025) provide transferable image–text representations whose downstream predictions depend on cross-modal similarity.

Within VLM-based few-shot adaptation, different methods intervene at different parts of the model. CoOp Zhou et al. (2022b), CoCoOp Zhou et al. (2022a), MaPLe Khattak et al. (2023a), PromptSRC Khattak et al. (2023b), and KgCoOp Yao et al. (2023) learn textual or multimodal contexts; VPT Jia et al. (2022) learns visual prompts; CLIP-Adapter Gao et al. (2024), Tip-Adapter Zhang et al. (2022), and TaskRes Yu et al. (2023) modify features, cache evidence, or classifiers; CLIP-LoRA Zanella & Ben Ayed (2024) adapts visual Transformer weights through low-rank residuals. Strong source-free baselines further improve logit calibration, linear probing, deconfusion, or distribution alignment Tang et al. (2024); Huang et al. (2024); Li et al. (2025); Xu et al. (2026).

Recent work is particularly relevant because it increasingly treats SF-CDFSL as a mechanism problem rather than only a benchmark problem. The discriminability-trap analysis shows that stronger visual discriminability can conflict with VLM performance Zhang et al. (2026b); other studies recover under-used text-layer information, rectify local cross-modal alignment, break harmful tail-token alignment, or mitigate attention-allocation failures Zhang et al. (2026a); Zhao et al. (2026); Yi et al. (2026a;b). These findings collectively argue that successful adaptation depends on *how* pretrained cross-modal structure is modified. Our work narrows this question to the visual LoRA itself: instead of changing the loss, prompt, or alignment target, we ask which adapted Q/K/V computations actually support the final cross-modal decision.

A.2 PARAMETER-EFFICIENT VLM ADAPTATION AND CAPACITY ALLOCATION

PEFT restricts the trainable state while preserving most pretrained parameters. Bottleneck adapters Houlsby et al. (2019), prefix tuning Li & Liang (2021), prompt tuning Lester et al. (2021), BitFit Ben Zaken et al. (2022), and IA^3 Liu et al. (2022) represent distinct choices of trainable subspace. LoRA Hu et al. (2022) instead parameterizes each adapted weight update as a low-rank residual, which is attractive for visual Transformers because Q, K, and V projections expose a clear set of candidate adaptation locations.

Subsequent work improves LoRA by changing how low-rank capacity is represented or optimized. AdaLoRA Zhang et al. (2023) reallocates rank dynamically; DyLoRA Valipour et al. (2023) supports multiple effective ranks in one training process; VeRA Kopiczko et al. (2024) shares random low-rank matrices and learns lightweight scaling vectors; DoRA Liu et al. (2024) separates magnitude and direction; LoRA+ Hayou et al. (2024) uses asymmetric learning rates for the two low-rank factors. These methods answer how much low-rank capacity should be assigned or how that capacity should be optimized, but not whether different projection types play different decision-level roles after adaptation.

A related line removes trainable capacity altogether. AdapterDrop Rücklé et al. (2021) drops adapters from lower layers, SparseAdapter He et al. (2022) prunes adapter weights, Diff Pruning Guo et al. (2021) learns sparse task-specific differences, and LoRA-drop Zhou et al. (2025) removes LoRA modules using output-based importance. Structured reduction methods such as Lay-

erDrop Fan et al. (2020) and Movement Pruning Sanh et al. (2020) further show that parameter removal can itself act as compression or regularization. These methods provide natural references for separating sparse capacity allocation from functional placement. Our experiments compare Role-Aligned LoRA with Role-Reversed, dense low-rank, deep-only, random sparse, and matched-budget alternatives, allowing projection role and branch topology to be examined independently of parameter count.

A.3 Transformer functional analysis and causal attribution

ViT Dosovitskiy et al. (2021) and DeiT Touvron et al. (2021) establish the visual Transformer architecture, while representation studies show that Transformer features evolve non-uniformly across depth Raghu et al. (2021); Park & Kim (2022). Such depth dependence makes uniform adaptation a questionable default, but observational structure alone does not identify causal contribution. Attention weights, for example, need not constitute faithful explanations Jain & Wallace (2019). Attention rollout/flow Abnar & Zuidema (2020) and relevance propagation Chefer et al. (2021) strengthen attribution by aggregating or propagating evidence through the network, whereas causal mediation directly intervenes on internal states Vig et al. (2020).

Our diagnostic follows the intervention principle but targets *adaptation-induced* computation. At a selected visual Transformer block, the hidden-state input is held fixed while adapted or base QK/V computations are substituted, and the resulting change is measured at the final cross-modal decision. This differs from post-hoc attention visualization and from pruning scores computed from parameter magnitude or module outputs: the quantity of interest is the decision effect caused by restoring a specific adapted computation. Because QK and V interactions are not strictly additive, we interpret the resulting scores as operational QK-sensitive and V-sensitive effects rather than a complete causal decomposition; Supplementary Sec. C reports the residual interaction explicitly.

B Proof of the Cross-Modal Geometry Proposition

Proposition 1 isolates a simple but consequential distinction between preserving visual geometry and preserving the text-defined classifier. We give the complete argument below.

Proof. Let $\boldsymbol{F} \in \mathbb{R}^{M\times d}$ denote visual features and let $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ be orthogonal, i.e., $\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}$. Define $\boldsymbol{F}' = \boldsymbol{F}\boldsymbol{R}$. Then

$$\boldsymbol{F}'\boldsymbol{F}'^\top = \boldsymbol{F}\boldsymbol{R}\boldsymbol{R}^\top\boldsymbol{F}^\top = \boldsymbol{F}\boldsymbol{F}^\top. \tag{16}$$

Thus every pairwise inner product among the rows of $\boldsymbol{F}$ is preserved. When the rows are normalized, pairwise Euclidean distances are preserved as well because

$$\|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2 = 2 - 2\boldsymbol{f}_i^\top \boldsymbol{f}_j. \tag{17}$$

Let $\boldsymbol{T} \in \mathbb{R}^{N\times d}$ contain the fixed normalized text prototypes. The original and transformed logit matrices are

$$\boldsymbol{L} = \boldsymbol{F}\boldsymbol{T}^\top, \qquad \boldsymbol{L}' = \boldsymbol{F}\boldsymbol{R}\boldsymbol{T}^\top. \tag{18}$$

Therefore

$$\boldsymbol{L}' = \boldsymbol{L} \quad \Longleftrightarrow \quad \boldsymbol{F}(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{T}^\top = \boldsymbol{0}. \tag{19}$$

This is the exact invariance condition for fixed $\boldsymbol{F}$. A sufficient condition that is independent of the query features is

$$\boldsymbol{R}\boldsymbol{T}^\top = \boldsymbol{T}^\top, \tag{20}$$

i.e., $\boldsymbol{R}$ fixes the text-prototype subspace pointwise. Conversely, if $\boldsymbol{F}$ has full column rank, $\boldsymbol{L}' = \boldsymbol{L}$ implies $\boldsymbol{R}\boldsymbol{T}^\top = \boldsymbol{T}^\top$. Hence an orthogonal transformation can leave all visual pairwise geometry unchanged while altering the logits of a fixed text classifier. □

Implication for the analysis. The proposition does not depend on a particular LoRA parameterization; it establishes that visual-geometry preservation alone cannot certify preservation of cross-modal decisions. The proposition therefore motivates decision-level measurements of adapted computation, complementing geometry-based observables with their consequences for the fixed cross-modal classifier.

Table 6: Absolute correlation with intervention-measured token margin contribution.

Signal	$\|r\|$
Global CLS–text compatibility	0.404
Patch true-text similarity	0.081
Patch max-text similarity	0.042
ATHA-style score	0.042
TIR-sum score	0.040
Token norm	0.022
AVWO	0.016
CLS-to-patch attention	0.012
Patch specificity	0.011

Table 7: **Query-side CLIP intervention validation.** Each layer contains 48 dataset–shot–text–seed profiles. $P(QK > V)$ reports the fraction with $G_{QK} > G_V$; $\mathbb{E}|G_{\text{full}}|$ measures complete block effect magnitude; $\mathbb{E}[S_l]$ reports signed QK-minus-V specialization.

Layer	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
$P(QK > V)$ (%)	29.2	35.4	37.5	41.7	25.0	16.7	14.6	18.8	37.5	41.7	60.4	**91.7**
$\mathbb{E}\|G_{\text{full}}\|$	.00038	.00063	.00071	.00099	.00155	.00185	.00266	.00327	.00279	.00353	.00543	**.00867**
$\mathbb{E}[S_l]$ ($\times 10^{-3}$)	-0.138	-0.113	-0.153	-0.246	-0.581	-0.973	-1.369	-0.971	-0.655	-0.394	+0.708	**+3.511**

C Full Counterfactual and Layerwise Analysis

This section provides the complete counterfactual analysis underlying the layerwise functional profile, including the surrogate correlation audit, intervention implementation, support- and query-side profiles, leave-one-domain-out statistics, condition-level trajectories, and training-time evolution.

C.1 Surrogate Definitions and Full Correlation Summary

Table 6 reports the absolute correlations used in Figure 2(a) of the main paper. Global CLS–text compatibility is reported as a decision-oriented reference alongside the local visual observables.

C.2 Intervention Implementation and Non-Additivity

For every target block l, the adapted network is run normally up to $H^{(l)}$. Base and adapted Q/K/V projections are then evaluated from this same hidden state, after which the four $LL/BL/LB/BB$ attention states defined in the main paper are propagated through the unchanged remainder of the network. Residual paths, normalization, MLPs, and all non-target blocks remain in their trained adapted state. PE-Core additionally preserves its native rotary transformation after choosing the projection source.

The QK/V construction is not assumed to be additive. Across all 576 query-side CLIP layerwise measurements, the interaction residual

$$\epsilon_l = G_{\text{full}}^{(l)} - G_{QK}^{(l)} - G_V^{(l)} \tag{21}$$

has mean absolute magnitude 2.49×10^{-4}, compared with 2.71×10^{-3} for $|G_{\text{full}}|$. The residual is therefore approximately 9.2% of the full-effect magnitude on average. This supports S_l as a comparative operational diagnostic while ruling out its interpretation as a complete additive causal factorization.

C.3 Query-Side All-Layer Validation

Averaging the query-side full-block effect over the four coarse regions gives 8.53×10^{-4} for L0–4, 2.82×10^{-3} for L5–9, 5.43×10^{-3} for L10, and 8.67×10^{-3} for L11. The signed specialization is negative throughout L5–9, becomes mildly positive at L10, and increases sharply at L11. The query-

Table 8: Aggregate agreement between support- and query-side intervention profiles.

Statistic	Value
Specialization Pearson	0.962
Specialization Spearman	0.734
Complete-effect Pearson	0.996
Specialization sign agreement	12/12

Table 9: Support-only leave-one-domain-out stability. "Neg." means every L5–9 mean S_l is negative.

Held out	L5–9 neg.	$P_{10}(QK > V)$	$P_{11}(QK > V)$	$C_{11}/C_{0:4}$
EuroSAT	yes	33.3%	88.9%	5.38×
CropDisease	yes	55.6%	88.9%	6.67×
ISIC	yes	66.7%	100.0%	6.75×
ChestX	yes	44.4%	88.9%	10.92×

side profile therefore exhibits the same depth-dependent organization observed in the support-side measurements.

C.4 Support- and Query-Side Profile Agreement

We apply the same blockwise counterfactual analysis to clean views of the labeled support examples after Uniform-QKV support adaptation. The Base-only study covers four domains, 1-/5-shot settings, three seeds, and ten diagnostic episodes per condition. Table 8 compares the resulting 12-layer aggregate profile with the query-side measurements. Although the absolute intervention scale is larger on support examples the relative layer organization is highly consistent: specialization Pearson correlation is 0.962, complete-effect Pearson correlation is 0.996, and the specialization signs agree at all 12 layers.

The support profile itself has negative mean S_l at every block from L5 through L9, $P(QK > V) = 50.0\%$ at L10, and $P(QK > V) = 91.7\%$ at L11. Mean complete-block effect rises from 2.76×10^{-3} over L0–4 to 7.34×10^{-3} over L5–9 and 1.96×10^{-2} at L11. The larger support-side scale is expected because these samples were used for adaptation; the placement depends on the coarse ordering and role direction rather than matching absolute magnitudes.

Leave-one-domain-out aggregation further characterizes the stability of the regional profile. Removing each target domain in turn preserves negative mean specialization throughout L5–9, while L11 remains QK-dominant in at least 88.9% of the remaining profiles. The L10 proportion ranges from 33.3% to 66.7%, consistent with its mixed QK/V behavior.

C.5 Condition-Level Specialization

The condition-level trajectories reveal that the aggregate role profile is not driven by a single shot or semantic setting. EuroSAT and CropDisease exhibit consistently negative specialization through the middle/deep blocks and a pronounced positive terminal shift across most conditions. ISIC shows a weaker terminal tendency, whereas ChestX remains substantially more variable across conditions. These trajectories therefore expose the condition-level variation underlying the domain aggregates reported in the main paper.

C.6 Temporal Diagnostic

The checkpoint trajectories provide a temporal view of the measured specialization. Figure 6 shows L5, L7, L10, and L11 over normalized adaptation progress. V- and QK-sensitive effects evolve jointly across conditions rather than following a shared V-first/QK-later sequence. The final depth organization is therefore better characterized as a property of the adapted solution than as a common temporal ordering.

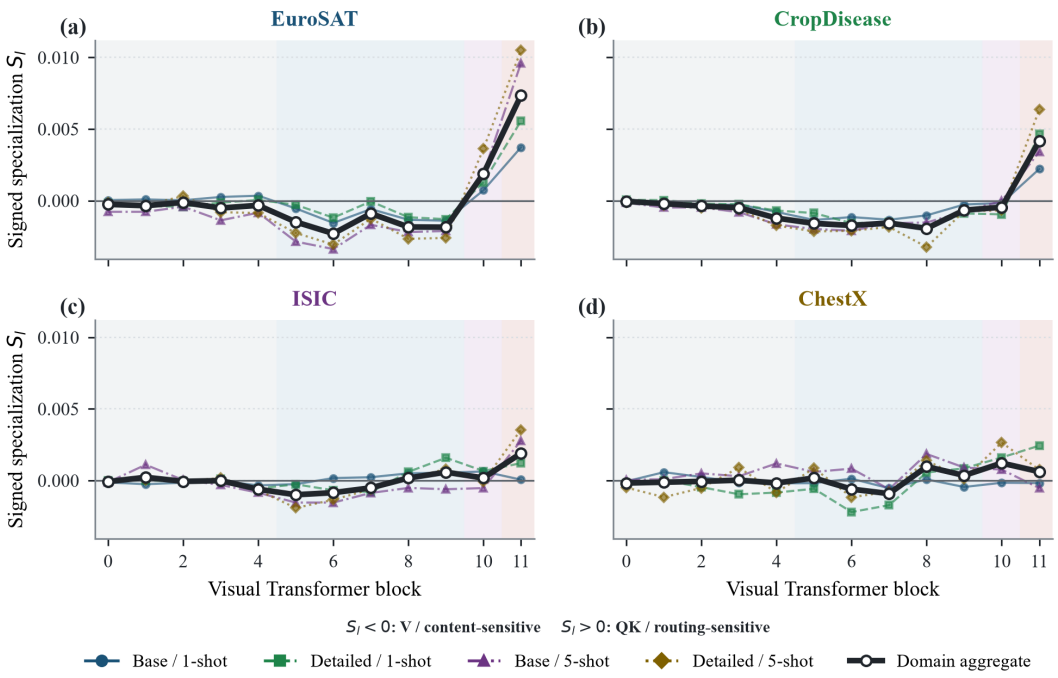

Figure 5: **Query-side condition trajectories expose the consistency behind each domain aggregate.** Thin colored curves show three-seed mean signed specialization separately for Base/Detailed text and 1-/5-shot adaptation, and the thick black curve is the domain aggregate. Negative values indicate V/content sensitivity and positive values indicate QK/routing sensitivity. EuroSAT and CropDisease show a consistent V-to-QK transition; ISIC retains a weaker positive terminal tendency, whereas ChestX remains mixed and does not show consistent terminal QK dominance.

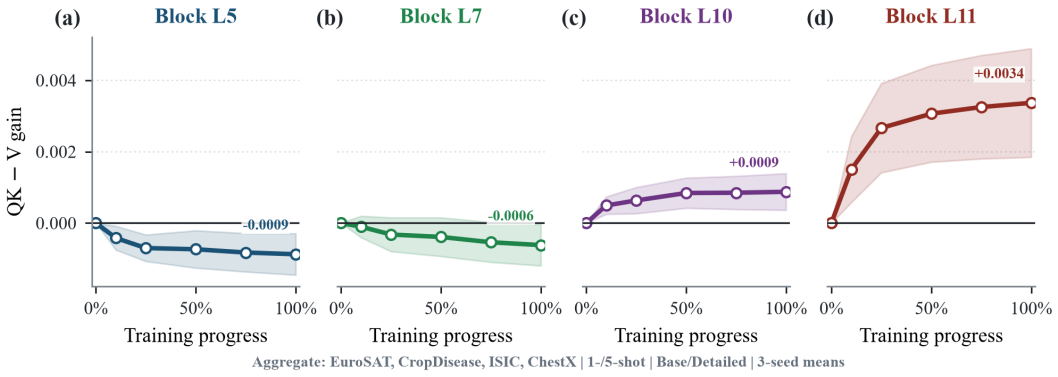

Figure 6: **Training trajectories do not support a universal V-first, QK-later optimization chronology.** Signed QK-minus-V decision effects are shown over normalized training progress for blocks L5, L7, L10, and L11, aggregated across EuroSAT, CropDisease, ISIC, and ChestX, 1-/5-shot adaptation, and Base/Detailed text conditions; each source row is a three-seed mean. Shaded regions are 95% confidence intervals across the 16 domain–shot–text profiles, and annotations give final-checkpoint means. The absence of a shared temporal sign transition indicates that the depth-dependent functional profile emerges without a universal V-first/QK-later training trajectory.

D Functional-Placement and Budget Controls

This supplementary section expands four complementary controls used in Sec. 6 and the method discussion: projection-direction reversal, formal low-budget alternatives, one-block boundary robust-

Table 10: **Functional-direction diagnostic.** Aggregate results over EuroSAT/CropDisease, 1-/5-shot, and base/detailed semantics. Role-Aligned follows the discovered depth-wise direction; Role-Reversed swaps the middle and terminal assignments.

Variant	Params	Param ratio	Accuracy	Margin	Adapt time (s)
Uniform QKV	884,736	1.000	91.43	0.05978	20.29
V-only	294,912	0.333	91.44	0.05679	16.70
QK-only	589,824	0.667	89.93	0.04744	18.47
Role-Reversed	344,064	0.389	89.68	0.05024	13.88
Role-Aligned	245,760	0.278	**91.81**	0.05943	**13.26**

Table 11: Role-Aligned minus Role-Reversed accuracy (pp) over the complete four-domain diagnostic screen.

Dataset	Base-1	Base-5	Detail-1	Detail-5	Mean
EuroSAT	+4.62	+1.16	+4.36	+0.18	+2.58
CropDisease	+2.49	+0.71	+2.76	+0.80	+1.69
ISIC	+1.33	-0.80	-0.44	-0.53	-0.11
ChestX	-0.71	-0.27	+0.09	+1.42	+0.13

ness, and an exact-high-budget placement test. Together, these experiments characterize projection direction, low-budget topology, regional boundary sensitivity, and parameter-count effects.

D.1 Functional-Direction Control

Table 10 reports the five-variant diagnostic comparison over EuroSAT/CropDisease, 1-/5-shot, and base/detailed semantics. Role-Aligned reaches 91.81% accuracy with 27.8% of Uniform's trainable LoRA parameters, whereas reversing the middle/terminal role assignment yields 89.68%. Because the reversed assignment uses QK on five middle/deep blocks and V only at L11, it contains 344,064 parameters versus 245,760 for Role-Aligned—40% more trainable capacity. Role-Aligned nevertheless yields the larger paired decision margin in all $120/120$ Aligned–Reversed episodes.

D.2 Cross-Domain Direction Boundary

The strongest aggregate effects occur on EuroSAT and CropDisease. Table 11 extends this diagnostic to all four target domains. The Aligned–Reversed separation remains large on EuroSAT and CropDisease but contracts toward zero on ISIC and ChestX, mirroring the specialization analysis in Sec. 4.

D.3 Formal Low-Budget Alternatives

We run an independent Base-only formal control suite with the same four domains and 800/400 episode counts, for 4,800 paired episodes per variant. The estimand is the equal-weight mean over the eight dataset–shot conditions; paired bootstrap resampling is performed within each condition before macro averaging. This avoids assigning twice the weight to 1-shot conditions merely because they contain 800 rather than 400 episodes.

Uniform-r4/r5 retain Q/K/V LoRA in all 12 blocks while reducing rank, whereas Deep-4 retains full-rank QKV only in the four deepest blocks. Three 10-branch random layouts are evaluated in the screening study; Random-10 #202 attains the highest screening accuracy and is included in the full low-budget comparison.

The control suite separates parameter count from branch topology. Uniform-r4/r5 have comparable or smaller parameter counts but remain dense across all 36 projection branches, so they retain essentially the full Uniform adaptation time and are lower in macro accuracy. Deep-4 is the fastest control but loses 1.229 pp despite using more parameters than Role-Aligned. The exact-budget Random-10 challenger uses the same 245,760 parameters yet is lower by 0.321 pp and requires 18.1% more

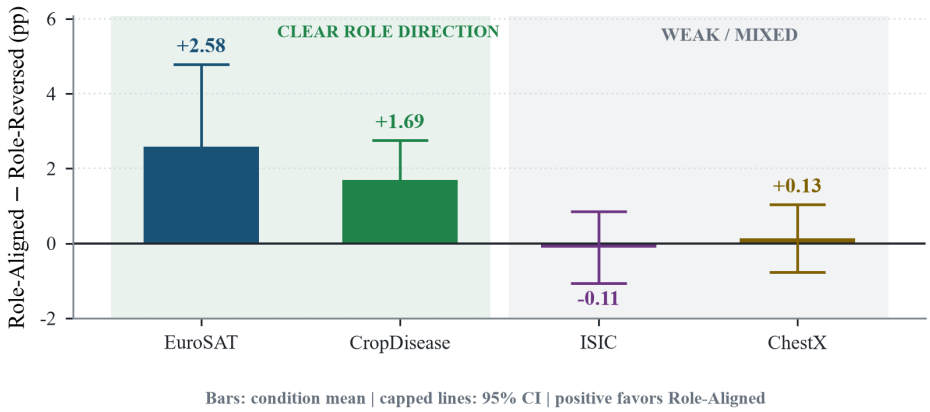

Figure 7: **Reversing the discovered role direction identifies where the mechanism is informative.** Bars show mean paired Role-Aligned-minus-Role-Reversed accuracy in percentage points, and capped lines show 95% confidence intervals across the four shot/prompt conditions. The separation is pronounced on EuroSAT and CropDisease but contracts on ISIC and ChestX, matching the clear-versus-mixed role boundary in Fig. 3 in the main paper.

Table 12: **Formal Base-only low-budget control suite.** Δ denotes Role-Aligned minus the listed comparator using equal-condition stratified paired bootstrap. Each variant is evaluated over the same eight dataset–shot conditions.

Variant	Params	Param ratio	Macro acc.	Δ pp [95% CI]	Time (s)
Uniform QKV $r16$	884,736	1.000	61.577	-0.014 $[-0.129, +0.102]$	18.18
Role-Aligned $r16$	245,760	0.278	61.564	ref.	12.18
Uniform QKV $r4$	221,184	0.250	61.198	$+0.366$ $[+0.252, +0.483]$	18.19
Uniform QKV $r5$	276,480	0.313	61.268	$+0.296$ $[+0.182, +0.410]$	18.10
Deep-4 QKV $r16$	294,912	0.333	60.335	$+1.229$ $[+1.100, +1.357]$	**10.66**
Random-10 #202	245,760	0.278	61.242	$+0.321$ $[+0.201, +0.442]$	14.88

adaptation time. Decision margins tell the same low-budget story: Role-Aligned exceeds Uniform-r4 by $+0.00197$, Uniform-r5 by $+0.00168$, Deep-4 by $+0.00188$, and Random-10 by $+0.00180$ under the equal-condition estimand.

D.4 One-Block Boundary Robustness

We vary the boundary between the frozen prefix and the V-LoRA region by one Transformer block while retaining QKV-LoRA at L10 and QK-LoRA at L11. The three resulting layouts contain 11, 10, and 9 active LoRA branches, corresponding to 30.6%, 27.8%, and 25.0% of the Uniform-QKV parameter count. The Base-only study uses two seeds and ten episodes per dataset–shot condition.

Moving V-LoRA one block earlier increases macro accuracy from 60.758% to 61.150%, together with an increase from 10 to 11 active branches and from 11.38 s to 12.03 s adaptation time. Moving the boundary one block later reduces the active set to 9 branches and the adaptation time to 10.67 s, with macro accuracy of 60.500%.

Relative to the L0–4/L5–9 configuration, the paired macro differences are -0.392 pp (95% CI $[-0.783, +0.008]$) for the earlier V-LoRA boundary and $+0.258$ pp (95% CI $[-0.208, +0.725]$) for the later boundary. The three configurations form a smooth capacity–time–accuracy continuum, indicating that the measured organization is regional rather than concentrated at a single sharp layer transition.

Table 13: **One-block frozen/V boundary perturbation.** The frozen-prefix/V-LoRA boundary is shifted by one block while the L10 QKV and L11 QK assignments are retained. Results use the Base-only two-seed robustness setting.

Frozen	V-LoRA	Branches	Param %	Accuracy	Time (s)
L0–3	L4–9	11	30.6	**61.150**	12.03
L0–4	L5–9	10	27.8	60.758	11.38
L0–5	L6–9	9	25.0	60.500	**10.67**

Table 14: **Exact-high-budget placement control.** All variants in this table are evaluated on the same 16-condition diagnostic setting; BM-Flat and BM-RoleWeighted match the 884,736-parameter Uniform-QKV budget.

Variant	Params	Rank sum	Accuracy	Margin	Time (s)
Uniform QKV $r16$	884,736	576	62.84	0.02267	17.51
Role-Aligned $r16$	245,760	160	62.83	0.02209	11.49
BM-Flat	884,736	576	**63.08**	0.02485	11.47
BM-RoleWeighted	884,736	576	63.01	**0.02486**	**11.47**

D.5 Exact-High-Budget Placement Control

We further evaluate the Role-Aligned branch locations at the full Uniform-QKV parameter budget. BM-Flat and BM-RoleWeighted keep the same projection locations while increasing rank to a total of 884,736 trainable parameters. BM-Flat distributes the matched rank approximately uniformly across active branches, whereas BM-RoleWeighted allocates more rank to roles with larger intervention-measured magnitude.

BM-Flat increases mean decision margin relative to Uniform QKV, whereas BM-RoleWeighted remains nearly tied with BM-Flat. The result indicates that the selected projection locations retain their effectiveness when additional low-rank capacity is assigned to them, while proportional rank weighting provides little additional benefit in this setting.

E PE-Core Replication and Transfer

This supplementary section expands the second-backbone study along four complementary axes: intervention replication inside the adapted visual Transformer, full formal transfer over all target domains, role-direction statistics, and a frozen-pooler endpoint test that marks the boundary of the observed depth pattern.

E.1 Intervention Replication in the Adapted Visual Transformer

At L11, the PE-Core specialization score is 0.00718 ± 0.00143 on EuroSAT and 0.00451 ± 0.00128 on CropDisease (95% CI), with QK exceeding V in $20/20$ and $18/20$ episodes. The second backbone therefore reproduces the terminal QK-sensitive behavior in the adapted Transformer under the tested diagnostic conditions.

E.2 Full Eight-Condition Formal Transfer

The PE-Core formal benchmark covers all four target domains under both 1-shot and 5-shot adaptation. Table 15 reports the complete eight-condition Role-Aligned–Uniform comparison over $4{,}800$ paired episodes.

Across all eight conditions, the macro difference is $+0.159$ pp with stratified paired-bootstrap 95% CI $[+0.047, +0.273]$ pp. The two small negative 5-shot conditions on ISIC and ChestX remain visible in the full table. At the same 27.8% trainable-parameter ratio, support-set adaptation time

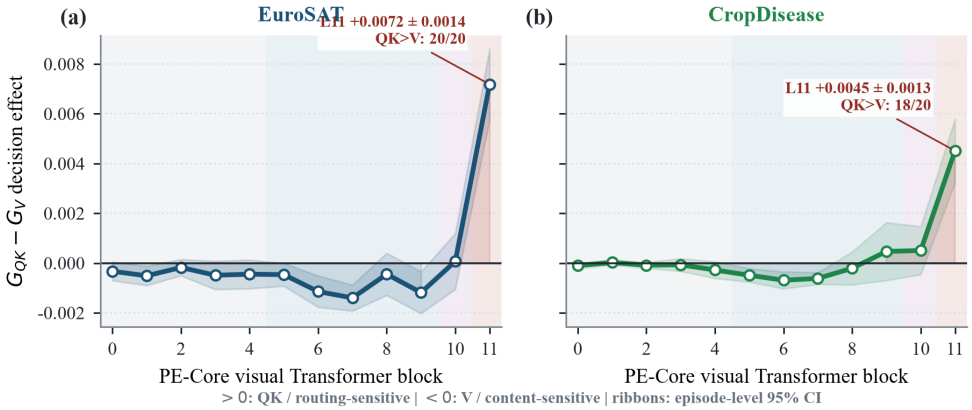

Figure 8: **PE-Core independently reproduces terminal QK/routing sensitivity under the CLIP intervention definition.** Curves show episode-level $G_{QK} - G_V$ decision effects for EuroSAT and CropDisease; positive values indicate QK/routing sensitivity, negative values indicate V/content sensitivity, and translucent ribbons show paired 95% confidence intervals over the diagnostic episodes. The annotated L11 effects are positive on both domains. The PE-Core experiment applies the same Role-Aligned layer/projection assignment used with CLIP.

Table 15: **PE-Core full formal transfer.** Differences are Role-Aligned minus Uniform QKV. Per-condition intervals use paired episode differences; the macro interval uses a stratified paired bootstrap over the eight condition means.

Dataset	Shot	Uniform	Role-Aligned	Δ Acc. (pp)	95% CI (pp)
EuroSAT	1	85.877	86.755	+0.878	±0.240
EuroSAT	5	94.930	95.223	+0.293	±0.196
CropDisease	1	92.457	92.743	+0.287	±0.151
CropDisease	5	97.860	97.950	+0.090	±0.114
ISIC	1	44.688	44.750	+0.062	±0.312
ISIC	5	60.640	60.457	-0.183	±0.396
ChestX	1	22.262	22.283	+0.022	±0.337
ChestX	5	24.397	24.223	-0.173	±0.567
Macro	8 cond.	**65.389**	**65.548**	**+0.159**	**[+0.047,+0.273]**

decreases from 22.047 s to 15.210 s (31.0%). Together, these results extend the Role-Aligned assignment to a second visual encoder.

E.3 Role-Direction Statistics Across Domains

PE-Core also provides Role-Reversed controls under the formal episode stream. Averaged over 1-/5-shot conditions, the Aligned–Reversed difference is $+1.304$ pp on EuroSAT and $+0.607$ pp on CropDisease, but only $+0.128$ pp on ISIC and $+0.042$ pp on ChestX. Table 16 gives the per-shot values, reproducing the same clear-versus-mixed domain boundary observed with CLIP.

E.4 Frozen-Pooler Endpoint Boundary

PE-Core ends with a frozen learned attention pooler that is outside the adapted 12-block Transformer. A corresponding K/V token-stream intervention at this endpoint produces the opposite ordering from the terminal adapted block: on EuroSAT the K-path gain is 0.00963 and the V-path gain is 0.02311 ($K - V = -0.01348$); on CropDisease the corresponding values are 0.00555 and 0.02825 ($K - V = -0.02270$). K exceeds V in only $3/20$ and $0/20$ episodes. This endpoint test shows that depth alone does not determine the measured role direction.

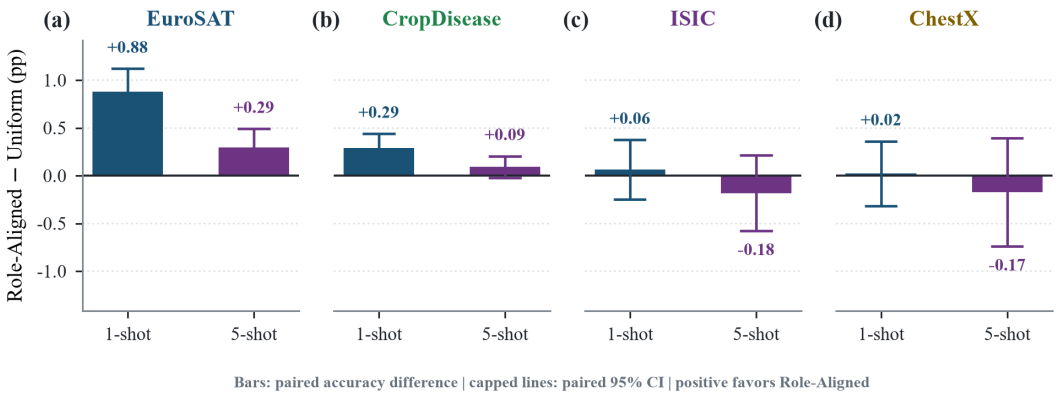

Figure 9: **The Role-Aligned placement on PE-Core exhibits domain-dependent condition effects.** Bars show paired Role-Aligned-minus-Uniform accuracy differences for 1-shot and 5-shot adaptation, and capped lines show paired 95% confidence intervals; positive values favor Role-Aligned. Gains concentrate on EuroSAT and CropDisease, while ISIC and ChestX retain the mixed-domain boundary.

Table 16: PE-Core Role-Aligned minus Role-Reversed accuracy (pp) with paired 95% confidence intervals.

Dataset	Shot	Δ Acc.	95% CI
EuroSAT	1	+1.325	±0.311
EuroSAT	5	+1.283	±0.239
CropDisease	1	+0.620	±0.206
CropDisease	5	+0.593	±0.180
ISIC	1	+0.252	±0.375
ISIC	5	+0.003	±0.506
ChestX	1	+0.367	±0.365
ChestX	5	-0.283	±0.607

F Dataset, Protocol, and Implementation Details

F.1 Detailed Dataset Description

We evaluate on the four target domains of the Broader Study of Cross-Domain Few-Shot Learning (BSCD-FSL) benchmark Guo et al. (2020): CropDisease, EuroSAT, ISIC2018, and ChestX. These domains deliberately span progressively different acquisition processes and visual statistics, from natural RGB leaf photographs to satellite imagery, dermoscopy, and grayscale radiography. Table 17 summarizes the target pools used by the standard benchmark. The image counts refer to the BSCD-FSL target subsets rather than the complete upstream repositories, which can contain additional images or labels. As shown in Fig 10, these four target domains—spanning agriculture, remote sensing, dermatology, and medical radiology—collectively constitute a series of challenging cross-domain scenarios characterized by an increasing degree of domain shift relative to the source domain. shifts from the source domain

CropDisease. CropDisease is derived from the PlantVillage plant-disease collection Mohanty et al. (2016). The BSCD-FSL target pool contains 38 crop/disease categories and 43,456 images. Images are close-up photographs of healthy or diseased plant leaves acquired under relatively controlled conditions. Compared with the other target domains, CropDisease remains closest to conventional natural-image statistics, but recognition depends heavily on fine texture, color, and lesion-pattern differences that are highly specialized to plant pathology.

EuroSAT. EuroSAT Helber et al. (2019) contains 27,000 Sentinel-2 satellite images in 10 land-use/land-cover categories. The benchmark uses RGB imagery with a top-down remote-sensing

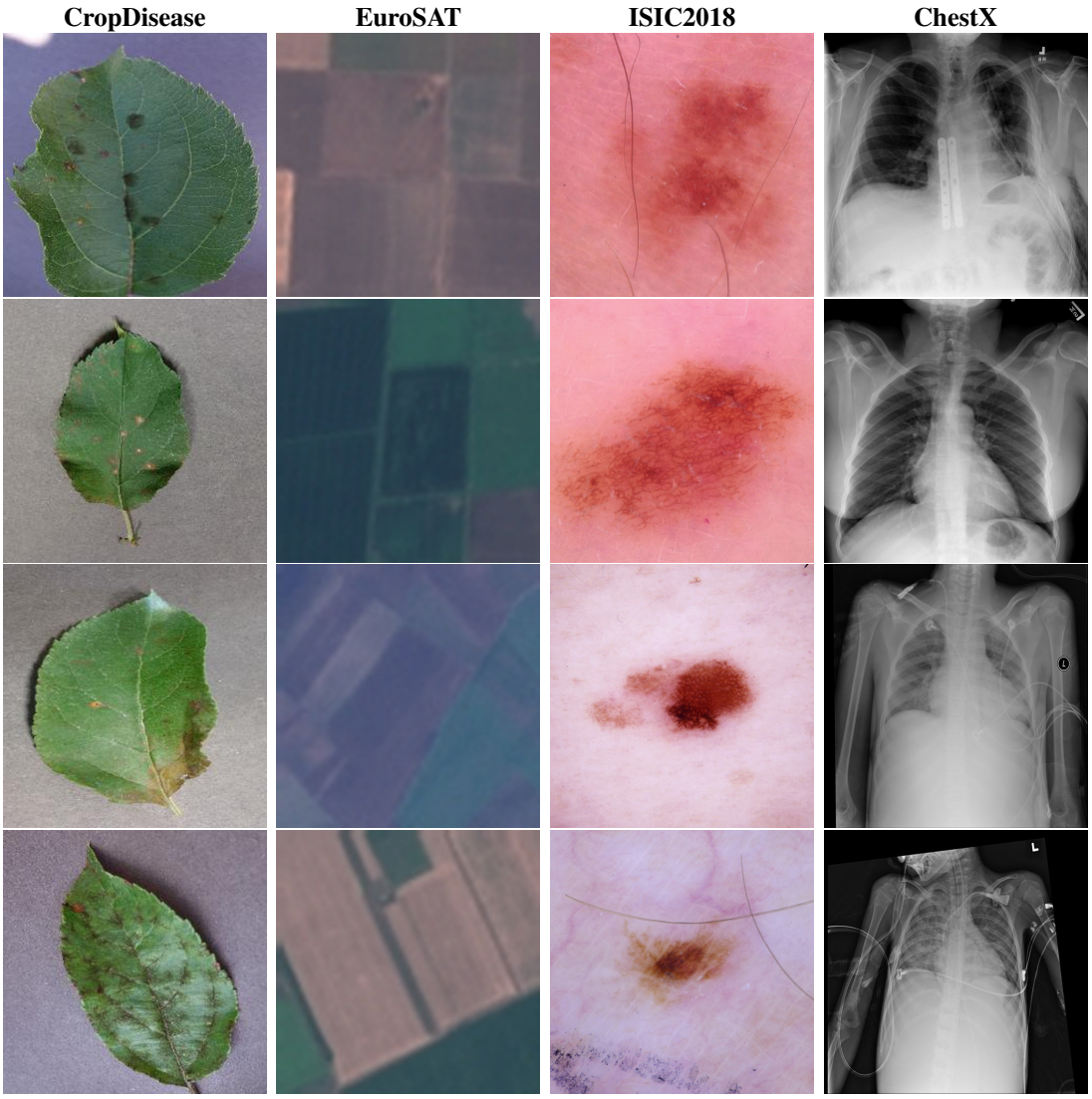

Figure 10: **Sample images from the four target-domain datasets.** CropDisease (agriculture), EuroSAT (remote sensing), ISIC2018 (dermoscopy), and ChestX (chest radiography).

Table 17: **Target-domain datasets used in the BSCD-FSL benchmark.** "Images" denotes the benchmark target pool from which few-shot episodes are sampled.

Dataset	Domain	Classes	Images	Modality
CropDisease	Agriculture	38	43,456	RGB leaf images
EuroSAT	Remote sensing	10	27,000	RGB satellite images
ISIC2018	Dermoscopy	7	10,015	RGB skin lesions
ChestX	Radiography	7	25,847	Grayscale chest X-rays

viewpoint. The lack of perspective distortion and the prevalence of repeated spatial patterns make its image statistics substantially different from ordinary object-centric photographs, while retaining full color information.

ISIC2018. ISIC2018 Codella et al. (2019) is a dermoscopic skin-lesion classification dataset with 10,015 images in seven diagnostic categories. The images are medical close-ups rather than natural

scenes; discriminative evidence is concentrated in lesion morphology, pigmentation, border structure, and local texture. This domain therefore introduces both a strong semantic shift and a change in acquisition geometry relative to natural-image pretraining.

ChestX. ChestX is the BSCD-FSL subset derived from the NIH ChestX-ray8 collection Wang et al. (2017). It contains 25,847 frontal chest radiographs across seven target categories. Unlike the other three datasets, ChestX is grayscale, and its class evidence is expressed through radiographic structures rather than surface appearance. It is consequently the most visually distant target domain in the original BSCD-FSL ordering Guo et al. (2020).

Source-free scope. The experiments treat CLIP and PE-Core as pretrained foundation models and construct all episodic adaptation tasks from the four BSCD-FSL target-domain pools. Each episode updates visual LoRA parameters from its labeled target support set. The four target datasets therefore provide the few-shot adaptation and evaluation domains, while the pretrained backbone supplies the source knowledge.

F.2 Episodic Evaluation Protocol

Every task is 5-way. For a K-shot episode, $K \in \{1, 5\}$ labeled support images are sampled per class together with 15 disjoint query images per class, giving 75 query images per episode. The support set drives LoRA adaptation, and the resulting adapted model produces predictions for the query set.

The support-side role profile is estimated from a separate collection of diagnostic episodes using the same blockwise intervention defined in Sec. C. Query-side intervention measurements provide the corresponding second view of the layerwise organization.

For CLIP, each 1-shot condition contains 800 episodes, and each 5-shot condition contains 400 episodes. The Base setting therefore contains eight dataset–shot conditions and 4,800 paired episodes, corresponding to 360,000 query predictions per method. The Detailed-semantics setting contributes another eight conditions and 4,800 paired episodes. Uniform QKV and Role-Aligned share the same class stream, support/query split, augmentation seed, and training random state within each paired comparison.

The PE-Core benchmark follows the same 5-way episodic construction and the same 800/400 episode counts across the four target domains. Its eight dataset–shot conditions contain 4,800 paired episodes and 360,000 query predictions per method. The CLIP-derived 12-block Role-Aligned layer/projection assignment is transferred unchanged to the PE-Core visual Transformer.

Additional diagnostics use smaller collections for profile agreement, temporal behavior, role reversal, high-budget placement, and boundary robustness. The low-budget comparison uses all eight Base dataset–shot conditions and 4,800 paired episodes per variant.

F.3 Input Processing and Paired Comparison

Each backbone uses its native image preprocessing pipeline, and the same input transforms are applied to Uniform QKV, Role-Aligned, and the corresponding controls. No method-specific crop, resize, or dataset-dependent preprocessing is introduced to favor a placement. Within each paired experiment, the image identities, support/query partition, augmentation random state, and optimizer random state are shared across compared variants. Within each paired comparison, image identities, support/query partitions, augmentation randomness, and optimizer randomness are shared across variants, reducing episode-sampling variance in the reported method differences.

F.4 LoRA and Optimization

For every active branch,

$$W' = W + \frac{\alpha}{\sqrt{r}} BA, \tag{22}$$

with rank $r = 16$, $\alpha = 8$, and dropout 0.25. Uniform and Role-Aligned use AdamW with learning rate 2×10^{-4}, weight decay 0.01, and 250 optimizer updates; ChestX follows the inherited 125-update protocol. The pretrained visual and text parameters remain frozen, so Role-Aligned changes

Table 18: Role-Aligned minus Uniform accuracy (pp). Positive values favor Role-Aligned.

Dataset	Base-1	Base-5	Detail-1	Detail-5	Mean
EuroSAT	+0.99	-0.26	+0.83	-0.41	+0.29
CropDisease	+0.80	-0.03	+0.42	-0.15	+0.26
ISIC	-0.88	-0.77	-0.40	-0.99	-0.76
ChestX	-0.03	+0.46	-0.07	+0.29	+0.16

only the active Q/K/V LoRA locations. Similarity and decision calculations use dtype-compatible tensors, and the final intervention statistics are accumulated in FP32.

Support-set adaptation time is measured on a single NVIDIA RTX 3090. Under the Base setting, Uniform-QKV requires 17.20 s per episode, and Role-Aligned requires 11.30 s, corresponding to a 34.3% reduction. The support-side role-profiling cost is reported separately in the following subsection.

F.5 Support-Side Role Profiling

The support-side role profile is obtained from Uniform-QKV adapted diagnostic episodes. Same-hidden-state counterfactuals are evaluated on clean views of their labeled support examples and aggregated over four domains, two-shot settings, three seeds, and ten episodes per condition.

The resulting profile contains a low-effect early region, a V-sensitive middle region, a mixed L10 transition, and a strongly QK-sensitive L11 endpoint. This produces the L0–4/L5–9/L10/L11 Role-Aligned assignment used throughout the CLIP and PE-Core experiments. Supplementary Sec. C reports the support/query profile agreement and leave-one-domain-out statistics.

The support-side profiling experiment contains 240 diagnostic runs. On one RTX 3090, mean support adaptation requires 17.52 s per run and mean intervention evaluation requires 0.94 s. The complete profiling workload is 4,432 s (73.9 min, 1.23 GPU-hours), including 4,206 s of support adaptation and 226 s of intervention evaluation. Relative to the measured 5.91 s reduction in Base episode-adaptation time, this workload corresponds to approximately 751 episodes of accumulated adaptation-time savings.

F.6 Auxiliary Detailed-Semantics Condition

Detailed semantics provide an auxiliary text condition constructed from 1–3 real class-reference images. The extracted visual evidence is converted into self-contained CLIP-compatible descriptions and held constant across episodic comparisons. Because this condition introduces additional class-reference images, the Base template is reported separately as the standard no-extra-reference setting.

G Extended Benchmark Results and Domain Boundaries

G.1 Primary Base-Only Result and Full Domain Boundary

The standard no-extra-reference Base protocol is the primary CLIP benchmark. Averaged equally over the eight dataset–shot conditions, Uniform QKV reaches 61.688% and Role-Aligned 61.724%, a difference of $+0.036$ pp with stratified paired-bootstrap 95% CI $[-0.080, +0.151]$. Table 18 then separates Base and auxiliary Detailed conditions by domain.

ISIC favors Uniform in all four conditions, illustrating a domain boundary of the sparse placement. The intervention measures the decision contribution of components within an adapted Uniform model, whereas sparse retraining follows a different optimization trajectory. The resulting accuracy effects therefore remain domain dependent, with the largest negative shift occurring on ISIC.

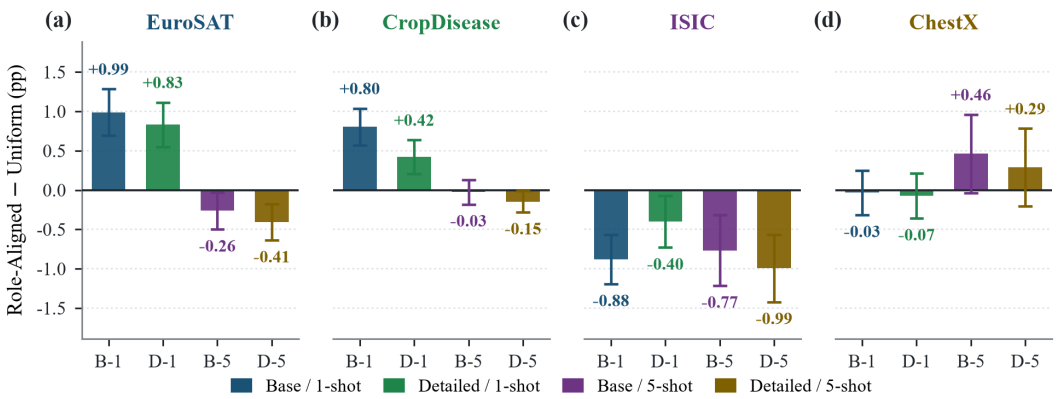

Figure 11: **Condition-wise formal CLIP differences.** Bars show paired Role-Aligned-minus-Uniform accuracy differences for Base (B) and Detailed (D) text under 1-shot and 5-shot adaptation; capped lines are episode-level paired 95% confidence intervals. Across the 16 Base+Detailed conditions, the macro difference is -0.012 pp with stratified paired-bootstrap 95% CI $[-0.094, +0.070]$.

Table 19: **Published SF-CDFSL context.** Accuracy (%) on the four target datasets. Published rows reproduce the reported benchmark values, and the final two rows show the Base Uniform-QKV and Role-Aligned results from our experiments.

	5-way 1-shot					5-way 5-shot				
Method	ISIC	EuroSAT	CropDisease	ChestX	Avg.	ISIC	EuroSAT	CropDisease	ChestX	Avg.
StepSPT	32.97	70.01	84.84	22.84	52.68	52.12	89.40	96.01	26.36	65.97
Tip-Adapter	32.68	75.44	77.15	22.24	51.87	46.96	87.24	94.19	24.07	63.12
AMU-Tuning	32.29	72.24	80.20	21.56	51.57	44.60	88.47	94.26	23.34	62.66
LP++	33.63	73.05	81.84	21.72	52.56	48.49	87.48	94.47	23.89	63.58
LDC	33.72	74.39	84.07	22.32	53.62	49.70	90.82	96.71	25.89	65.78
CLIP-LoRA	36.40	81.72	84.62	21.86	56.07	52.22	93.31	95.88	24.61	66.50
DTrap	38.12	85.02	87.20	22.68	58.26	56.14	94.14	96.64	26.61	68.38
VtT	38.20	85.01	87.00	22.70	58.23	56.20	94.58	97.21	26.42	68.57
ATHA	38.86	82.56	87.99	24.00	58.35	56.42	93.41	97.62	26.67	68.53
CC-CDFSL	38.13	86.07	88.91	22.21	58.83	54.72	94.35	97.08	25.47	67.90
Uniform QKV (ours, Base)	36.73	82.38	85.16	21.78	56.51	53.07	93.69	96.61	24.07	66.86
Role-Aligned (ours, Base)	35.85	83.37	85.97	21.75	**56.73**	52.31	93.43	96.59	24.54	66.72

G.2 Published SF-CDFSL Context

Published SF-CDFSL methods use different episode counts, optimization schedules, objectives, and prompt conditions. Table 19 summarizes their reported performance together with our Base results to provide benchmark context across the four target domains.

G.3 Independent Effect of Detailed Semantics

Across the eight dataset–shot settings, Detailed semantics increase average accuracy by approximately $+1.40$ pp for Uniform-QKV and $+1.30$ pp for Role-Aligned. The similar shift across both placements indicates that the semantic condition acts primarily through the text representation rather than through the LoRA branch layout.

H Algorithmic

Algorithm 1 summarizes the construction of the layerwise functional profile and the tested 12-block Role-Aligned LoRA instantiation.

Algorithm 1: Role-Aligned LoRA: Support-Side Functional Profiling and Episodic Adaptation

Require: Diagnostic support episodes $\mathcal{E}_{\text{diag}}$; pretrained visual encoder f in the studied $L = 12$ block configuration; LoRA configuration Θ_{LoRA}; episode $(\mathcal{S}, \mathcal{Q})$.

Ensure : Role-Aligned placement $\mathcal{P}^{\text{RA}}$; query predictions $\hat{\mathbf{y}}_{\mathcal{Q}}$.

Functional-role profiling

Initialize $\mathcal{B}_l^C, \mathcal{B}_l^S \leftarrow \emptyset$ for $l = 0, \ldots, L-1$

foreach $e \in \mathcal{E}_{\text{diag}}$ **do**

  Adapt Uniform-QKV LoRA on the labeled support set of e

  for $l \leftarrow 0$ **to** $L-1$ **do**

    Cache $H^{(l)}$ and evaluate $\mathcal{F}_{LL}^{(l)}, \mathcal{F}_{BL}^{(l)}, \mathcal{F}_{LB}^{(l)}, \mathcal{F}_{BB}^{(l)}$

    Compute the corresponding mean margins $M_{LL}^{(l)}, M_{BL}^{(l)}, M_{LB}^{(l)}, M_{BB}^{(l)}$

    $G_{QK}^{(l)} \leftarrow M_{LB}^{(l)} - M_{BB}^{(l)}; \quad G_V^{(l)} \leftarrow M_{BL}^{(l)} - M_{BB}^{(l)}$

    $G_{\text{full}}^{(l)} \leftarrow M_{LL}^{(l)} - M_{BB}^{(l)}$

    Append $|G_{\text{full}}^{(l)}|$ to $\mathcal{B}_l^C$ and $G_{QK}^{(l)} - G_V^{(l)}$ to $\mathcal{B}_l^S$

$C_l \leftarrow \text{mean}(\mathcal{B}_l^C)$ and $S_l \leftarrow \text{mean}(\mathcal{B}_l^S)$ for all l

Tested 12-block sparse instantiation

$\mathcal{P}_l^{\text{RA}} \leftarrow \emptyset$ for $l = 0{:}4; \quad \mathcal{P}_l^{\text{RA}} \leftarrow \{V\}$ for $l = 5{:}9$

$\mathcal{P}_{10}^{\text{RA}} \leftarrow \{Q, K, V\}; \quad \mathcal{P}_{11}^{\text{RA}} \leftarrow \{Q, K\}$

Episodic adaptation

Initialize LoRA branches according to $\mathcal{P}^{\text{RA}}$ and optimize them on $\mathcal{S}$

$\hat{\mathbf{y}}_{\mathcal{Q}} \leftarrow f_{\mathcal{P}^{\text{RA}}}(\mathcal{Q})$

return $\mathcal{P}^{\text{RA}}, \hat{\mathbf{y}}_{\mathcal{Q}}$