SALF: SPATIALLY ATTENTIVE LAYER FUSION FOR PROBING VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting a pre-trained Vision Transformer (ViT) to a downstream task requires extracting a single compact representation from multi-layer patch tokens. Common probing baselines either attentively probe the patches of the final layer (discarding intermediate features) or fuse per-layer average-pooled summaries across depth (discarding fine-grained spatial detail). By probing all patches across all layers on multiple datasets, we demonstrate that the final layer is rarely the most informative and that attentive spatial aggregation significantly outperforms static average pooling. Notably, our analysis reveals that this spatial aggregation plays a more dominant role in downstream performance than depth-wise fusion, effectively reducing the reliance on complex multi-layer integration. To efficiently leverage both dimensions, we factorize the all-token readout into SALF-P (Spatially Attentive Layer Fusion with Per-layer aggregation), which first aggregates the spatial patches of each layer using a parameter-efficient shared cross-attention module, followed by a second cross-attention module to fuse these features across layers. However, further analysis reveals that depth-wise fusion has limited impact when paired with spatial attention. This motivates SALF-A, which averages over layer features before applying a single attentive pooling. SALF-A matches SALF-P across our benchmarks, while reducing the spatial readout to one pass and shrinking the token cache by a factor of the network depth. Evaluations across 19 downstream datasets and seven frozen backbones, alongside Image Quality Assessment (IQA) tasks under parameter-efficient fine-tuning, confirm that SALF-P and SALF-A substantially improve accuracy on fine-grained visual perception tasks by up to 16.20 percentage points over ALF while maintaining strong performance on globally semantic tasks.

1 INTRODUCTION

Adapting foundational vision models to downstream tasks is highly effective, particularly under limited data regimes (Evci et al., 2022; Xin et al., 2026; Zhang et al., 2020; Hu et al., 2022; Ciernik et al., 2026). By leveraging strong pre-trained representations, these models can achieve competitive performance with minimal training overhead. However, standard linear classification heads expect a single feature vector per image. This structural bottleneck presents a core challenge: efficiently extracting and compressing the rich, high-dimensional feature space of foundational models into a robust, unified representation for downstream tasks.

Two common approaches exist for extracting representations from Vision Transformers (ViTs). The first is attentive probing (Lee et al., 2019; Yu et al., 2022; Bardes et al., 2024; El-Nouby et al., 2024; Psomas et al., 2025; Chen et al., 2024), which applies cross-attention to summarize patch tokens strictly at the final layer. While this approach preserves fine-grained spatial information, it entirely discards intermediate representations. Conversely, Attentive Layer Fusion (ALF) (Ciernik et al., 2026) highlights the value of intermediate features by applying cross-attention across different network depths. However, ALF relies on average-pooled summary tokens from each layer, inherently destroying the spatial details preserved by attentive probing. Consequently, a natural question arises: can we dynamically aggregate features both spatially within layers and hierarchically across layers?

To investigate this, we apply probing on all layers and evaluate an unconstrained global probing baseline, FLAT, where a single learnable query token attends to all patch tokens and CLS tokens across all layers simultaneously. We compare this baseline against attentive probing and ALF on

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.

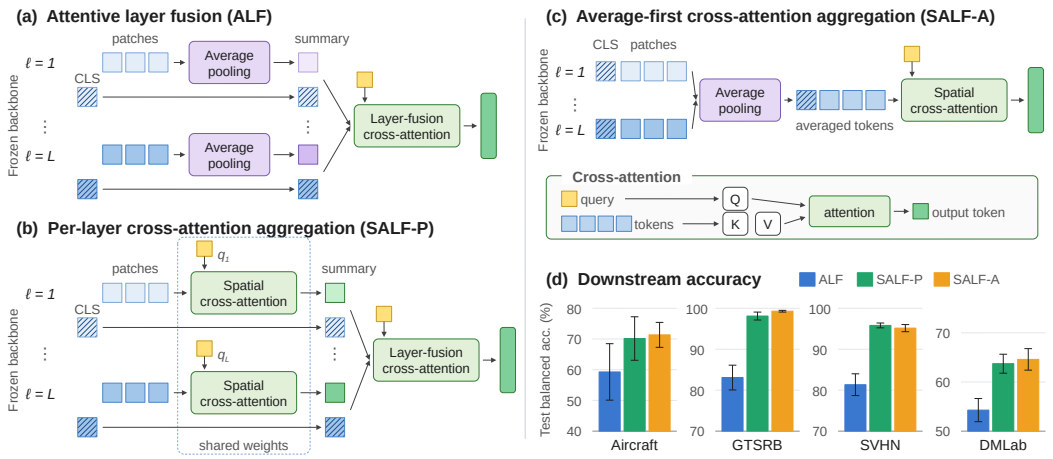

Figure 1: **Overview of SALF-P and SALF-A compared with ALF.** (a) ALF (Ciernik et al., 2026) average-pools the patch tokens of every layer and fuses the pooled summaries and the CLS tokens of all layers with a cross-attention module. (b) SALF-P replaces average pooling with shared spatial cross-attention, subsequently fusing the resulting summaries and CLS tokens across depth. (c) SALF-A first averages the tokens across layers and then applies a single spatial cross-attention to the averaged tokens. (d) Test balanced accuracy on four datasets that depend on spatial cues, averaged across seven frozen backbones (Table 1): both variants substantially outperform ALF.

downstream adaptation (see Fig. 2). We observe three key *findings*: (1) within any given layer, attentive patch aggregation significantly outperforms static average pooling, highlighting the need for attentive spatial aggregation; (2) earlier layers often match or even exceed the performance of the final layer, confirming the value of intermediate features; and (3) because attentive spatial attention inherently captures robust fine-grained features, it reduces the need for extensive cross-layer integration.

Although FLAT is effective, it is computationally expensive due to the massive key and value sequences in the cross-attention module. To achieve comparable representational power efficiently, we propose SALF-P, which decomposes the global aggregation problem into two sequential steps: intra-layer spatial aggregation and inter-layer depth fusion. Specifically, instead of relying on average pooling like ALF (Ciernik et al., 2026), we introduce a lightweight, shared cross-attention module that processes patch tokens within each intermediate layer. This allows the model to dynamically attend to salient regions (Fig. 4), producing a spatially aggregated summary token per layer. For cross-layer fusion, these summary tokens and the corresponding class tokens are further aggregated using a secondary cross-attention module. While this factorization is highly effective, we find that layerwise fusion contributes little once spatial attention is present, making per-depth spatial attention redundant (finding 3). Therefore, we propose a novel variant, SALF-A, that first averages every token, including the class token, across layers, then applies a single spatial cross-attention to the averaged tokens. By performing spatial aggregation only once, this variant significantly reduces the total computational overhead and the required cache sizes, yet does not sacrifice accuracy. Fig. 1 illustrates the high-level architectural design of both variants alongside a quantitative performance comparison against the state-of-the-art baseline (*i.e.*, ALF).

Our primary contributions are summarized as follows:

- Through a comprehensive probing analysis, we show that attentive spatial aggregation and intermediate features improve downstream performance, while learned cross-layer fusion provides only small gains in mean accuracy over simple averaging once spatial attention is used.
- We propose SALF-P, which addresses both dimensions using shared cross-attention for intra-layer spatial aggregation, followed by cross-attention for depth fusion. Since depth fusion contributes little once spatial attention is present, we further introduce SALF-A, an efficient probe that averages features across layers before a single spatial aggregation step, substantially reducing computation and shrinking the token cache by a factor of the network depth.

- Extensive evaluations across 19 datasets and seven frozen backbones demonstrate that SALF improves balanced accuracy by up to 16.20 percentage points over ALF. We further evaluate SALF on Image Quality Assessment (IQA) under a parameter-efficient fine-tuning (PEFT) setup.

2 Related Work

Spatial aggregation in probing. Building on set-attention pooling (Lee et al., 2019) and cross-modal pretraining (Yu et al., 2022), attentive probing learns queries to dynamically aggregate spatial patch tokens. Although widely adopted as an evaluation benchmark for self-supervised vision models (Chen et al., 2024; Bardes et al., 2024; El-Nouby et al., 2024; Psomas et al., 2025), prior work largely confines probing to the representations of the final layer, inherently presuming that top-layer features are sufficient for downstream transfer.

Intermediate features in probing. Several studies indicate that earlier layers in both CNNs and ViTs encode highly localized spatial information before transitioning to abstract semantic representations (Zeiler & Fergus, 2014; Carter et al., 2019; Raghu et al., 2021; Caron et al., 2021). Leveraging these multi-depth features is well-established for transfer learning, utilizing mechanisms ranging from CNN feature pyramids (Ronneberger et al., 2015; Lin et al., 2017; Hariharan et al., 2015) to intermediate ViT readouts (Alain & Bengio, 2016; Raghu et al., 2021; Meza De la Jara et al., 2026; Evci et al., 2022; Singh et al., 2026). Among methods exploiting intermediate ViT features, Visual Query Tuning (VQT) (Tu et al., 2023) uses learnable per-layer queries with frozen backbone attention projections, whereas SALF learns its own spatial-attention projections. Closest to our work is Attentive Layer Fusion (ALF) (Ciernik et al., 2026), which aggregates depth representations using cross-attention. However, ALF first reduces spatial patch tokens via simple average pooling, inadvertently discarding fine-grained spatial information crucial for detailed recognition tasks. In contrast, SALF jointly optimizes feature aggregation both *within* and *across* layers.

Backbone adaptation. A parallel line of research explores parameter-efficient fine-tuning (PEFT) to adapt the backbone under restricted parameter budgets (Xin et al., 2026). Representative strategies include inserting bottleneck adapter modules (Houlsby et al., 2019; Chen et al., 2022), prepending learnable visual prompt tokens to the input sequence (Jia et al., 2022; Li & Liang, 2021), and injecting low-rank weight updates (LoRA) into attention layers (Hu et al., 2022; Zhang et al., 2020). Adaptation techniques update the underlying representations and require backpropagation through the backbone network, orthogonal to feature aggregation. Nevertheless, we demonstrate our method consistently enhances performance in combination with the PEFT setting.

Foundational models for no-reference image quality assessment (NR-IQA). Unlike semantic categorization, NR-IQA relies heavily on localized distortions, blur, noise, and compression artifacts, leading modern approaches to exploit multi-level and spatially resolved representations (Su et al., 2020; Ke et al., 2021; Wang et al., 2023). Closest to our experimental setting, Yadav et al. (2026) benchmark multiple pretrained backbones using a lightweight MLP readout. Applying our method to this evaluation framework yields further performance improvements across multiple datasets.

3 Problem and Motivation

In this section, we first discuss the design space for downstream adaptation. We then present a diagnostic analysis independent of the test set to empirically motivate our proposed architecture.

3.1 The design space of layer probes

The design space of probing mechanisms over a frozen backbone can be categorized along two primary axes: the spatial reduction strategy and the depth of the extracted features. Standard linear probing, last-layer attentive probing (ATT) (Chen et al., 2024), and Attentive Layer Fusion (ALF) (Ciernik et al., 2026) each occupies a quadrant in the 2×2 grid formed by these axes (average pooling vs. attention and final vs. all layers). The quadrant combining dynamic spatial aggregation and intermediate features, also explored by VQT (Tu et al., 2023), forms the central focus of our proposed methods (SALF-P and SALF-A). We formally define these architectural choices in §4.

3.2 LAYER-WISE PROBING AND UNCONSTRAINED READOUT

To investigate the impact of spatial aggregation strategies and feature depth on task adaptation, we design two diagnostic setups. First, we extract and probe the features of each network layer in isolation, comparing standard average pooling of patch tokens against attentive probing. This layer-wise evaluation identifies the most effective spatial aggregation method and reveals how feature utility varies across network depth. However, analyzing layers in isolation does not confirm whether a model benefits from integrating both spatial attention and intermediate features simultaneously.

Therefore, we introduce FLAT, an unconstrained readout mechanism that bypasses per-layer spatial reduction entirely. In this configuration, a single learnable query attends jointly over all patch tokens and class tokens across all layers. Because FLAT accesses all intermediate tokens without prior pooling, it serves as a global baseline for our diagnostic analysis.

To prevent any exposure to the standard test data, for this section, we repartition the standard training set into a 70/15/15 split for training, validation, and testing. All diagnostic statistics presented in this section are computed on this newly defined testing split. These comparisons examine the effects of spatial aggregation and feature depth.

Fig. 2 compares these single-layer probes against their all-layer counterparts on four datasets, drawn from a variety of image domains requiring features at different granularity. Across nearly every layer, attentive probing significantly outperforms average pooling. This substantial performance gap underscores the critical role of attentive spatial aggregation for robust downstream task adaptation, leading to our first key finding.

Finding 1: Attentive spatial aggregation improves over uniform pooling for tasks relying on spatial cues. Applying attentive spatial aggregation generally enhances downstream accuracy across network depths. These gains are especially pronounced on datasets that demand fine-grained visual perception, highlighting the value of preserving localized spatial details.

Furthermore, the layer-wise analysis reveals that intermediate features frequently yield stronger performance than those of the final layer. Indeed, integrating representations across all depths consistently boosts performance over single-layer baselines for both aggregation methods, with particularly large gains observed under average pooling. This demonstrates that discarding intermediate features severely limits downstream potential, leading to our second key finding.

Finding 2: Intermediate features improve downstream readout. Intermediate layers frequently outperform the final layer. Multi-layer readouts provide competitive performance without explicitly selecting a single layer.

Moreover, the figure highlights a critical interplay between spatial reduction and depth-wise information. The performance gap between single-layer attentive probing and FLAT has a consistent pattern: the best layers per dataset can match FLAT, while the other layers fall smoothly in performance away from them. This demonstrates the network's reduced reliance on depth-wise information, when attentive spatial attention already extracts informative representations at the individual layer level. Since a single-layer spatial attention probe with the right choice of layer can match the global FLAT probe, it should be possible to obtain strong performance even without highly sophisticated depth-wise integration, as long as the "core" spatial information from the most important layers is maintained. Thus, we argue that complex depth-wise aggregation mechanisms can be eschewed if the probe instead has access to the right spatial feature distribution, even if limited in layerwise extent.

Finding 3: Spatial information can mitigate the need for cross-layer attention. Because dynamic spatial attention extracts highly robust fine-grained representations within individual layers, it inherently reduces the reliance on depth-wise integration.

The FLAT readout mechanism, which simultaneously leverages both intermediate features and dynamic spatial attention, consistently yields the highest overall accuracy. However, this unconstrained approach requires computing a full attention matrix over all patch tokens across every layer. Consequently, its computational overhead and memory footprint make it challenging for practical deployment. This critical limitation motivates our primary research objective: developing an architectural framework that captures the representational power of both attentive spatial aggregation and multi-layer feature fusion, but does so in a highly efficient manner. This naturally leads to the design of our proposed methods, SALF-P and SALF-A.

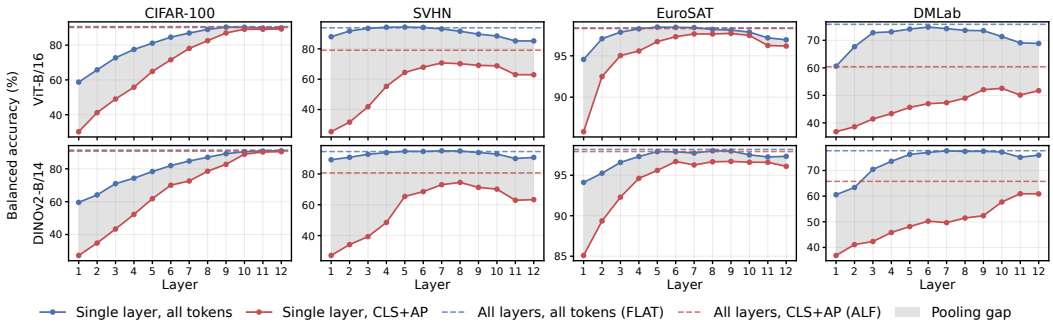

Figure 2: **Single-layer probes across depth.** Balanced accuracy (%) of a probe trained on one layer of ViT-B/16 (top) and DINOv2 ViT-B/14 (bottom), attending over all tokens of that layer (blue) or over its CLS and average-pooled (AP) tokens (red). Dashed lines show the corresponding probes over all layers jointly: FLAT (blue) and ALF (red). The shaded region is the accuracy lost by pooling the spatial tokens. The gap is the smallest on EuroSAT and CIFAR-100, which rely more on global semantics than spatial arrangements. Additional visualizations are provided in Supp.§C.1.

4 SALF: SPATIALLY ATTENTIVE LAYER FUSION

4.1 NOTATION AND FORMULATION

Let $\boldsymbol{H}^{(\ell)} \in \mathbb{R}^{N\times d}$ denote the patch tokens from layer $\ell \in \{1,\dots,L\}$ of a Vision Transformer (ViT) (Dosovitskiy et al., 2020), and let $\boldsymbol{h}_{\text{cls}}^{(\ell)}$ denote its class token. The common readout mechanisms can be formulated as a combination of three components: a spatial reduction function r applied per layer, a fusion function f applied over a selected subset of layers $\mathbb{L}$, and a final classification head g:

$$\hat{y} = g\Big(f\big(\{r(\boldsymbol{H}^{(\ell)})\}_{\ell\in\mathbb{L}}\big)\Big). \tag{1}$$

Standard linear probing utilizes $\mathbb{L} = \{L\}$, $r(\boldsymbol{H}^{(\ell)}) = \boldsymbol{h}_{\text{cls}}$, and $f = \text{Id}$. Last-layer attentive probing (Chen et al., 2024) (ATT) similarly restricts $\mathbb{L} = \{L\}$ but defines r as an attentive reduction. Attentive Layer Fusion (ALF) (Ciernik et al., 2026) operates over all layers ($\mathbb{L} = \{1,\dots,L\}$) but defines the spatial reduction as a simple concatenation of the class token and the average-pooled patches: $r(\boldsymbol{H}^{(\ell)}) = [\boldsymbol{h}_{\text{cls}}^{(\ell)}; \frac{1}{N}\sum_i \boldsymbol{H}_i^{(\ell)}]$, with f being a learned cross-attention over these pooled summaries. In contrast, our proposed method, SALF-P, utilizes an attentive reduction r while simultaneously spanning all intermediate layers $\mathbb{L}$. While Eq. (1) describes this standard sequential extraction ($f \circ r$), our SALF-A, deliberately reverses this composition ($r \circ f$) to drastically reduce computational overhead. We detail both architectures below.

4.2 FACTORIZING THE UNCONSTRAINED READOUT

The unconstrained FLAT baseline computes a single attention matrix over all $L(N+1)$ tokens. This forces the model to resolve spatial selection and depth selection simultaneously, resulting in a prohibitive computational overhead. To design a more efficient and scalable architecture, we explore two distinct factorizations, both anchored by dynamic attentive spatial aggregation. The first approach, termed SALF-P, decomposes the operation into a two-stage process: within-layer spatial aggregation followed by across-layer depth fusion. While highly expressive, this design still requires executing spatial attention at every intermediate layer. Motivated by Finding 3, we propose SALF-A, a lightweight variant that first averages the features across depth, and then applies a single spatial aggregation step to the resulting tokens, trading some depth-wise information for runtime efficiency, while preserving enough spatial information to maintain performance. Below, we detail the core spatial aggregation module alongside the complete SALF-P and SALF-A architectures.

Attentive spatial aggregation. The within-layer stage reduces each layer's spatial tokens via multi-head cross-attention. Given queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$, standard attention is defined as:

$$\text{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}_q)(\boldsymbol{K}\boldsymbol{W}_k)^\top}{\sqrt{d_h}}\right)\boldsymbol{V}\boldsymbol{W}_v, \tag{2}$$

omitting the multi-head formulation for brevity. To generate the summary for layer ℓ, we introduce K learnable queries $\boldsymbol{Q}^{(\ell)} \in \mathbb{R}^{K\times d}$ and use the patch tokens of that specific layer as both keys and values: $r(\boldsymbol{H}^{(\ell)}) = \mathrm{Attn}(\boldsymbol{Q}^{(\ell)}, \boldsymbol{H}^{(\ell)}, \boldsymbol{H}^{(\ell)}) \in \mathbb{R}^{K\times d}$. Prior to their respective linear projections, we normalize the query, key, and value inputs using three distinct layer normalization modules.

SALF-P. Given layer ℓ's independent query $\boldsymbol{Q}^{(\ell)}$, let $\boldsymbol{S}^{(\ell)} = \mathrm{Attn}(\boldsymbol{Q}^{(\ell)}, \boldsymbol{H}^{(\ell)}, \boldsymbol{H}^{(\ell)}) \in \mathbb{R}^{K\times d}$ denote the spatially aggregated summary of layer ℓ from Eq. (2). The L layer summaries are computed simultaneously and subsequently fused via an attentive pooling stage over the stacked depth axis:

$$\boldsymbol{S} = \big[\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(L)}, \boldsymbol{h}_{\mathrm{cls}}^{(1)}, \ldots, \boldsymbol{h}_{\mathrm{cls}}^{(L)}\big] \in \mathbb{R}^{L(K+1)\times d}, \qquad \boldsymbol{z} = \mathrm{Attn}\big(\boldsymbol{q}, \boldsymbol{S}, \boldsymbol{S}\big), \tag{3}$$

where $\boldsymbol{q}$ is a global learnable query and the concatenated summaries $\boldsymbol{S}$ serve as the keys and values. By flattening the summaries into $L(K+1)$ tokens, a single softmax operation jointly evaluates the aggregated spatial content across the network depth.

This design yields two critical advantages. First, the projection matrices $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ and the normalization parameters are shared across all L layers. Only the queries $\boldsymbol{Q}^{(\ell)}$ remain layer-specific. Consequently, the main parameter overhead of this spatial reduction stage equals exactly one cross-attention block, regardless of the network depth. Second, the approach is highly parallelizable: the L within-layer attention operations are independent, allowing for highly efficient concurrent execution prior to the final depth fusion.

SALF-A. To avoid the computational cost of layer-wise spatial attention, and directly motivated by Finding 3, we introduce SALF-A. It collapses the multi-layer feature hierarchy into a single representation via mean pooling, prior to spatial aggregation:

$$\bar{\boldsymbol{H}} = \frac{1}{L}\sum_{\ell=1}^{L}[\boldsymbol{H}^{(\ell)}, \boldsymbol{h}_{\mathrm{cls}}^{(\ell)}] \in \mathbb{R}^{(N+1)\times d}, \qquad \boldsymbol{z} = \mathrm{Attn}\big(\boldsymbol{q}, \bar{\boldsymbol{H}}, \bar{\boldsymbol{H}}\big). \tag{4}$$

This approach assumes that depth-wise averaging does not destroy critical information from individual layers. This may hold in ViTs, due to residual connections inducing more uniform representations across layers (Raghu et al., 2021). Consequently, depthwise summarization, followed by a single spatial attention operation, is sufficient to extract the necessary fine-grained details. Empirically, SALF-A achieves average classification accuracy comparable to SALF-P with substantially lower readout cost, although performance differences remain on individual tasks. Crucially, SALF-A requires caching only the depth-pooled features, significantly improving storage requirements, feature loading, and memory pressure during hyper-parameter search and training. Both variants conclude with a classification head: $\hat{y} = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b})$.

5 Evaluation

5.1 Experimental Setup

Datasets. We evaluate on 19 image classification datasets following prior work (Ciernik et al., 2026), and on three additional IQA datasets. The full list and the dataset statistics are given in Supp.§A.1.
Backbones. We probe seven frozen Vision Transformers that cover three pre-training paradigms and several model sizes: supervised ViT-S/16, ViT-B/16 and ViT-L/16 (Dosovitskiy et al., 2020), self-supervised DINOv2-S/14 and DINOv2-B/14 (Oquab et al., 2023), and language-supervised CLIP-B/32 and CLIP-B/16 (Radford et al., 2021; Cherti et al., 2023). See Supp.§A.2 for more details.
Baselines. Linear probing, last-layer attentive probing (ATT) (Chen et al., 2024) and ALF (Ciernik et al., 2026) each occupies a quadrant in the feature aggregation design space, see Table 6. We also include VQT (Tu et al., 2023) for completeness. Details in Supp.§A.3.
Evaluation metric. We report top-1 balanced accuracy (BA) on the test set, which is the mean of the per-class recalls and is robust to the class imbalance present in several datasets. For IQA, we use Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC).
Implementation details. All methods are trained on extracted features from frozen backbones for all datasets except PCAM. VQT is trained end to end on every dataset. We search 21 combinations of learning rate and weight decay. The best candidate is selected via BA on the validation split. See Supp.§A.4 for details.

Table 1: **Balanced accuracy gains (pp) over the last-layer CLS linear probe.** Mean ± standard deviation over seven backbones. Bold and underline mark the largest and second-largest improvement. †Trained end-to-end. Horizontal lines separate the four dataset categories, in order: natural multi-domain, natural single-domain, specialized, and structured.

Dataset	Linear probe	Δ ALF	Δ ATT	Δ VQT†	Δ SALF-P	Δ SALF-A
STL-10	$99.20_{\pm 0.52}$	**+0.08**$_{\pm 0.11}$	+0.00$_{\pm 0.15}$	+0.00$_{\pm 0.30}$	<u>+0.06</u>$_{\pm 0.09}$	-0.17$_{\pm 0.23}$
CIFAR-10	$96.41_{\pm 1.89}$	+0.89$_{\pm 0.92}$	+0.80$_{\pm 0.79}$	<u>+1.30</u>$_{\pm 0.98}$	+1.27$_{\pm 1.20}$	**+1.35**$_{\pm 1.23}$
Caltech-101	$95.10_{\pm 1.29}$	+0.83$_{\pm 0.35}$	+0.09$_{\pm 0.78}$	**+1.58**$_{\pm 0.75}$	+1.22$_{\pm 0.60}$	<u>+1.52</u>$_{\pm 0.89}$
VOC2007	$87.42_{\pm 1.87}$	<u>+0.90</u>$_{\pm 0.54}$	-0.22$_{\pm 0.88}$	**+1.91**$_{\pm 0.76}$	+0.61$_{\pm 0.73}$	+0.60$_{\pm 1.01}$
CIFAR-100	$84.30_{\pm 5.73}$	+3.48$_{\pm 2.91}$	+2.61$_{\pm 2.22}$	+4.51$_{\pm 2.74}$	<u>+4.67</u>$_{\pm 3.46}$	**+4.77**$_{\pm 3.52}$
Country-211	$19.93_{\pm 4.64}$	**+3.52**$_{\pm 0.66}$	-0.71$_{\pm 1.80}$	+1.54$_{\pm 0.83}$	<u>+1.84</u>$_{\pm 1.50}$	-2.21$_{\pm 2.70}$
Pets	$92.79_{\pm 2.64}$	<u>+0.96</u>$_{\pm 1.59}$	+0.68$_{\pm 1.84}$	+0.55$_{\pm 1.85}$	**+1.35**$_{\pm 2.38}$	+0.56$_{\pm 2.54}$
Flowers	$97.23_{\pm 4.04}$	+0.87$_{\pm 1.91}$	+1.16$_{\pm 2.19}$	+1.17$_{\pm 2.60}$	**+1.63**$_{\pm 3.22}$	<u>+1.59</u>$_{\pm 2.96}$
Aircraft	$52.30_{\pm 11.05}$	+6.98$_{\pm 3.37}$	+10.28$_{\pm 4.45}$	+9.45$_{\pm 8.39}$	<u>+17.85</u>$_{\pm 5.85}$	**+19.04**$_{\pm 7.69}$
GTSRB	$69.59_{\pm 5.32}$	+13.49$_{\pm 3.51}$	+20.15$_{\pm 5.10}$	+27.36$_{\pm 5.52}$	<u>+28.51</u>$_{\pm 5.97}$	**+29.69**$_{\pm 5.31}$
SVHN	$54.54_{\pm 3.30}$	+26.80$_{\pm 2.33}$	+31.12$_{\pm 3.30}$	+38.06$_{\pm 2.77}$	**+41.27**$_{\pm 3.15}$	<u>+40.61</u>$_{\pm 3.02}$
PCAM†	$82.03_{\pm 2.30}$	+5.18$_{\pm 1.74}$	+5.27$_{\pm 1.15}$	<u>+5.73</u>$_{\pm 2.68}$	+4.02$_{\pm 3.07}$	**+5.89**$_{\pm 2.57}$
EuroSAT	$93.97_{\pm 2.55}$	+4.05$_{\pm 2.48}$	+3.27$_{\pm 2.25}$	+4.35$_{\pm 2.41}$	**+4.59**$_{\pm 2.51}$	<u>+4.55</u>$_{\pm 2.45}$
RESISC45	$89.64_{\pm 1.24}$	+5.31$_{\pm 1.04}$	+4.60$_{\pm 1.00}$	+5.54$_{\pm 0.89}$	<u>+6.57</u>$_{\pm 0.93}$	**+6.83**$_{\pm 1.20}$
Retinopathy	$45.65_{\pm 2.68}$	+6.24$_{\pm 2.15}$	+4.07$_{\pm 1.97}$	<u>+6.76</u>$_{\pm 2.60}$	+5.64$_{\pm 2.30}$	**+7.04**$_{\pm 2.41}$
DTD	$76.54_{\pm 2.34}$	<u>+2.78</u>$_{\pm 2.20}$	+0.07$_{\pm 1.14}$	**+2.90**$_{\pm 1.09}$	+2.52$_{\pm 1.44}$	+1.86$_{\pm 1.50}$
FER2013	$58.14_{\pm 3.52}$	+8.81$_{\pm 1.48}$	+8.50$_{\pm 1.76}$	+8.63$_{\pm 1.23}$	**+11.14**$_{\pm 2.51}$	<u>+10.00</u>$_{\pm 2.27}$
DMLab	$44.24_{\pm 2.77}$	+10.05$_{\pm 1.94}$	+14.53$_{\pm 3.08}$	+19.08$_{\pm 2.69}$	<u>+19.48</u>$_{\pm 2.68}$	**+20.37**$_{\pm 3.04}$
Mean	$74.39_{\pm 1.81}$	+5.62$_{\pm 0.75}$	+5.90$_{\pm 0.56}$	+7.80$_{\pm 0.97}$	**+8.57**$_{\pm 0.94}$	<u>+8.55</u>$_{\pm 0.98}$

5.2 Experimental Results

Performance on various tasks. Table 1 reports the balanced-accuracy gain of every baseline over the linear probe on 18 datasets, averaged over seven backbones. Additional results on ImageNet-1k are in Supp.§B.3. Both SALF-P and SALF-A slightly outperform FLAT, which attends over every token of every layer, as detailed in Supp.§B.5. SALF-P and SALF-A both improve the linear probe by 8.6 pp on average, against 5.6 pp for ALF and 5.9 pp for ATT. VQT, which attends over the patches of every layer through the backbone's frozen attention, gains 7.8 pp. The largest gaps to SALF-P are on Aircraft and SVHN, where spatial selection matters. SALF-A matches SALF-P within 0.1 pp, suggesting that fusing the layers before a single spatial attention step preserves most of the information recovered by attending within every layer, at 6× fewer probe GFLOPs per image (§5.2). However, SALF-P continues to outperform SALF-A on specific datasets, such as Country-211, Pets, and FER2013, highlighting its value as a robust alternative for challenging recognition scenarios.

The gain varies widely across dataset groups. The natural multi-domain group, which covers canonical object recognition closest to the backbones' pre-training setting, is largely saturated. The linear probe already exceeds 95% on STL-10, CIFAR-10, and Caltech-101, and all probes lie within 2 pp of each other. The natural single-domain group contains fine-grained tasks within one domain, such as aircraft variants, traffic signs, and house numbers, where discriminative cues often lie in specific, sometimes very small, regions of the image. This is where SALF gains the most, with SALF-P achieving +41.3 pp on SVHN, +28.5 pp on GTSRB, and +17.9 pp on FGVC Aircraft with respect to the linear probe baseline, and SALF-A is within 1.2 pp of it on each. ATT, which attends over tokens but reads only the last layer, already outperforms ALF on these three datasets, suggesting that spatial attention is more important than layer fusion for these tasks. The specialized group covers imagery from non-natural sensors, and the structured group covers texture, facial expression, and depth estimation in game frames; both are far from the pre-training distribution. SALF is the best probe on six of these seven datasets, demonstrating that our formulation generalizes well.

Performance on various backbone sizes. Fig. 3 breaks the aggregate numbers of Table 1 down by backbone. SALF-P and SALF-A have the highest mean gain on every backbone. The margin of SALF-P over both attentive baselines narrows as the backbone size grows, suggesting that larger backbones encode more of the task-relevant information in the last layer and pooled tokens. On DINOv2,

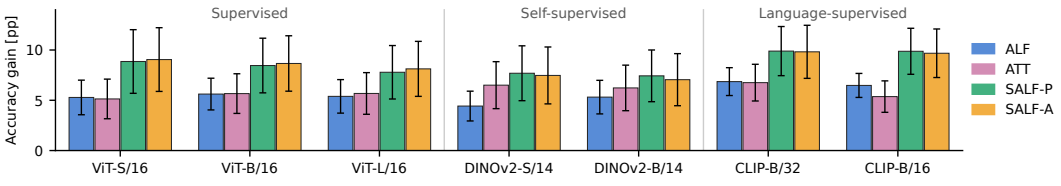

Figure 3: **Gains over the linear probe per backbone.** Balanced-accuracy gain in percentage points over the last-layer CLS linear probe for every probe and backbone, grouped by pre-training paradigm. Bars are the mean over 18 datasets, and error bars are the standard error of the mean. SALF-P and SALF-A have the two highest means on every backbone.

Table 2: **Performance on NR-IQA.** We report the mean and standard deviation for SRCC and PLCC over 3 random seeds in comparison with B_Gated (Yadav et al., 2026) and ALF (Ciernik et al., 2026).

	KADID10K		**KonIQ10K**		**KADID10K → KonIQ10K**	
Method	SRCC (↑)	PLCC (↑)	SRCC (↑)	PLCC (↑)	SRCC (↑)	PLCC (↑)
B_Gated	$0.841_{\pm 0.039}$	$0.840_{\pm 0.035}$	$0.928_{\pm 0.011}$	$0.943_{\pm 0.008}$	$0.630_{\pm 0.098}$	$0.662_{\pm 0.099}$
ALF	$0.910_{\pm 0.013}$	$0.912_{\pm 0.011}$	$0.944_{\pm 0.004}$	$0.954_{\pm 0.001}$	$0.651_{\pm 0.032}$	$0.670_{\pm 0.023}$
SALF-P	$\mathbf{0.924_{\pm 0.011}}$	$\mathbf{0.924_{\pm 0.011}}$	$\mathbf{0.948_{\pm 0.001}}$	$0.956_{\pm 0.001}$	$0.663_{\pm 0.007}$	$0.689_{\pm 0.006}$
SALF-A	$0.917_{\pm 0.022}$	$0.921_{\pm 0.022}$	$0.948_{\pm 0.002}$	$\mathbf{0.957_{\pm 0.002}}$	$\mathbf{0.705_{\pm 0.020}}$	$\mathbf{0.731_{\pm 0.026}}$

whose patch-level pre-training objective keeps the last-layer patch tokens spatially informative, ATT outperforms ALF and comes within 1.2 pp of SALF-P. On the supervised and CLIP backbones, which are trained through the CLS token alone, ATT performs on par with or below ALF. SALF reads both axes and is the best probe under all three supervision regimes.

Ablation studies. Table 15 in Supp.§B.6 varies how the L layers are combined. For SALF-P, which fuses after the spatial attention step, cross-attention over the per-layer summaries gives the best mean accuracy, while mean pooling loses only 0.3 pp. For SALF-A, which fuses before the spatial attention step, mean pooling, depth attention and two learned per-image layer weightings are within 0.3 pp of each other, and all of them are about 3 pp above ALF and ATT. This suggests that the gain comes from the spatial attention, not from the fusion method, consistent with §3. Capacity ablations of SALF-P and layer-subset ablations are in Supp.§B.6.

Performance on NR-IQA. We use parameter-efficient fine-tuning (PEFT) to fine-tune SigLIP2 (Tschannen et al., 2025) for image quality scoring. As shown in Table 2, our proposed variants consistently outperform the B_Gated (Yadav et al., 2026) and ALF baselines across both in-domain and cross-domain NR-IQA evaluations. In strictly in-domain settings, SALF-P achieves the highest correlation on KADID10K, while both variants perform comparably on KonIQ10K. Most notably, in the challenging cross-domain setting (KADID10K to KonIQ10K), the highly efficient SALF-A model surpasses all other methods, achieving an SRCC of 0.705 and a PLCC of 0.731. This confirms that parameter-free depth-wise aggregation before spatial attention maintains in-domain accuracy while improving generalization. Moreover, SALF outperforms ALF under PEFT, proving our methods remain advantageous in the finetuning setup (see Supp.§E for more results).

Computational overhead and full finetuning. Supp.§B.4 reports the cost of each probing method and of full finetuning on ViT-B/16. During training, since the backbone is frozen, we can cache the tokens and perform hyperparameter search and retraining on the cache only. Both FLAT and SALF-P store 2,364 tokens per image, equivalent to every token of every layer. SALF-A, on the other hand, stores 12× fewer tokens, since we can directly cache the average pooled spatial tokens across layers. At inference time, the backbone costs 17.56 GFLOPs per image. FLAT is the most expensive probing baseline, adding 5.59 (32%) GFLOPs on top and taking 1.6× the training memory of SALF-P (13.96 against 8.75 GiB at batch size 256), because it projects all 2,364 tokens at the full head width. SALF-A peaks at 1.18 GiB, since the layer mean reduces the 2,364 tokens to 197 before any projection. VQT caches the input tokens of every layer plus the final CLS token (2,365 per image) and recomputes the frozen keys and values at each step, which costs 2.97 training GFLOPs per image and 2.17 GiB, 2.1× and 1.8× those of SALF-A. Note that VQT accuracies in Table 1

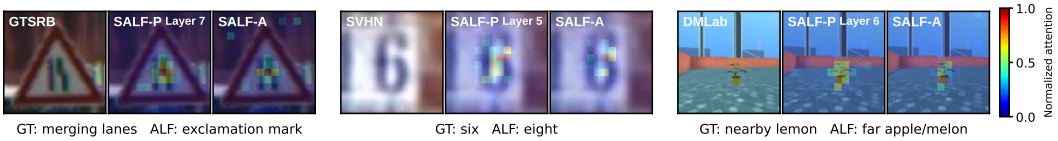

Figure 4: **Where the spatial query looks.** Test images that ALF misclassifies and both SALF variants classify correctly, with the SALF-P summarizer map at its most-attended layer and the SALF-A map over the averaged tokens. Ground truth (GT) label and ALF's prediction are given below each image.

come from the end-to-end training setup described in Supp.§A.4, while Table 13 separately profiles a cached implementation of the same readout.

All attentive probes have approximately 4.7M parameters with details outlined in Supp.§A.3. Both of the SALF variants perform on par with or exceed full finetuning, while training 18× fewer parameters (4.7M against 85.8M) on cached features, and using 1.9× (SALF-P) to 14× (SALF-A) less training memory per step. See Supp.§B.5 for more details.

5.3 Interpreting SALF

Attention map visualizations. Fig. 4 shows the spatial maps on images that ALF misclassifies and both SALF variants get correct. SALF-P shows the summarizer's map at its most-attended layer, while SALF-A shows a single map over the averaged tokens. SALF-P and SALF-A select similar patches, even though SALF-P chooses its positions separately in each layer and SALF-A chooses them once over the depth-averaged tokens. The class evidence stays at the same positions across depth, suggesting why averaging over depth before spatial attention costs SALF-A little accuracy.

When Is Spatial Averaging Costly? SALF's gains over ALF vary substantially across tasks, suggesting spatial mean pooling may remove task-relevant fine-grained cues. We perform two analyses of this phenomenon. First, we define a training-free, Fisher-style feature statistic (Fisher, 1936), *averaging efficiency* η_a, which compares uniform vs. optimal linear aggregation, measuring how much class signal averaging over axis a preserves. Spatial efficiency η_S strongly predicts SALF-P's advantage over ALF ($r = -0.84$) *without* fitting a probe. Combined with its layerwise analogue η_L, it isolates high-gain datasets: on tasks with $\eta_S < 0.5$ and $\eta_L > 0.5$, SALF-A gains 13.5 ± 4.5 BA points vs. 0.4 ± 3.3 otherwise (mean $\pm$ std. dev.). Second, attention non-uniformity alone does not predict SALF's gains: divergence from uniform attention, $\bar{\Delta}$, is not significantly correlated with performance gains ($r = 0.13$, $p > 0.05$). Instead, we quantify whether the spatial attention pattern is consistent across images by decomposing $\bar{\Delta}$ into dataset-level (static) and image-dependent (dynamic) components. Their balance is captured by the *spatial stationarity*, ξ, which strongly predicts SALF-P's advantage ($r = -0.86$, $p < 10^{-4}$); further, static-dominant tasks ($\xi < 0$) gain 12.6 ± 3.7 BA points, versus 0.9 ± 2.5 when $\xi \geq 0$. See Supp.§F and §G for complete results.

6 Conclusion

In this work, we investigate the challenge of extracting robust, compact representations from pre-trained Vision Transformers. Through a comprehensive layer-wise diagnostic analysis against an unconstrained readout mechanism, we demonstrate that dynamic spatial aggregation plays a more dominant role in downstream performance than depth-wise fusion, particularly for tasks demanding fine-grained visual perception. Motivated by these findings, we introduce SALF-P and, based on the limited contributions of depthwise attention, the more efficient SALF-A. These architectures optimize feature aggregation across both the spatial and depth dimensions, effectively preserving critical localized details without sacrificing global semantic context. Extensive evaluations across downstream datasets using frozen backbones, as well as parameter-efficient fine-tuning on image quality assessment tasks, confirm that our proposed methods consistently achieve enhanced downstream performance while maintaining highly competitive computational efficiency. Future work could explore strategies to further reduce feature caching requirements per input image. In particular, non-parametric token reduction techniques could decrease the number of spatial tokens stored for downstream adaptation.

AI USE STATEMENT

In this work, we used generative AI tools to locate a few relevant related papers, polish the writing, and generate specific code snippets for our experiments. We have rigorously reviewed all AI-assisted work. Specifically, we manually verified the accuracy and relevance of all AI-suggested literature, thoroughly reviewed and edited all AI-polished text, and rigorously tested the AI-generated code to guarantee its correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our results, we have comprehensively documented all dataset characteristics, evaluation protocols, and experimental configurations in §5. Additional implementation details, including specific hyper-parameters and training setups, are provided in Supp.§A. These sections are to provide all the necessary information to faithfully reproduce our results. Furthermore, to facilitate future research and verify our empirical claims, we will publicly release the complete source code upon approval.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab, 2016. URL `https://arxiv.org/abs/1612.03801`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision (IJCV)*, 132(1):208–223, 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Laure Ciernik, Marco Morik, Lukas Thede, Luca Eyring, Shinichi Nakajima, Zeynep Akata, and Lukas Muttenthaler. Beyond the final layer: Attentive multi-layer fusion for vision transformers. In *International Conference on Machine Learning (ICML)*, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. `https://www.kaggle.com/competitions/diabetic-retinopathy-detection`, 2015.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024.

Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning (ICML)*, pp. 6009–6033, 2022.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88 (2):303–338, 2010.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683. IEEE, 2020.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 178–178. IEEE, 2004.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp. 117–124. Springer, 2013.

Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 447–456, 2015.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 2020.

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 (ICML)*, pp. 2790–2799. PMLR, 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. *International Conference on Learning Representations (ICLR)*, 2022.

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 (ECCV)*, pp. 709–727. Springer, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *International Conference on Computer Vision (ICCV)*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*, pp. 3744–3753. PMLR, 2019.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, pp. 4582–4597, 2021.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *Eleventh international conference on quality of multimedia experience (qomex)*, pp. 1–3. IEEE, 2019.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

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

Ignacio Meza De la Jara, Cristian Rodriguez-Opazo, Damien Teney, Damith Ranasinghe, and Ehsan Abbasnejad. Mysteries of the deep: Role of intermediate representations in out of distribution detection. In *Neural Information Processing Systems (NeurIPS)*, 2026.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505. IEEE, 2012.

Bill Psomas, Dionysis Christopoulos, Eirini Baltzi, Ioannis Kakogeorgiou, Tilemachos Aravanis, Nikos Komodakis, Konstantinos Karantzalos, Yannis Avrithis, and Giorgos Tolias. Attention, please! revisiting attentive probing for masked image modeling. *arXiv preprint arXiv:2506.10178*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7725–7735. IEEE, 2023.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2023.

Yi Xin, Jianjiang Yang, Siqi Luo, Yuntao Du, Qi Qin, Haoxing Chen, Kangrui Cen, Yangfan He, Bin Fu, Yuewen Cao, et al. Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark. *International Journal of Computer Vision*, 134(6):304, 2026.

Ankit Yadav, Ta Duc Huy, and Lingqiao Liu. Revisiting vision–language foundations for no-reference image quality assessment. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 5416–5425. IEEE, 2026.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833. Springer, 2014.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European Conference on Computer Vision (ECCV)*, pp. 698–714. Springer, 2020.

SALF: Spatially Attentive Layer Fusion for Probing Vision Transformers

Supplementary Material

The supplementary material provides further details, extended results, and in-depth analyses. Specifically, §A details our experimental setup, including dataset statistics, backbone configurations, probe architectures, and hyperparameter search protocols. §B presents additional classification results, such as per-backbone accuracy breakdowns, evaluations on the MAE backbone, comparisons with full fine-tuning, and extended ablation studies. §C expands on our motivating layer-sweep experiments across all datasets and backbones. §D provides supplementary visualizations for both the spatial and cross-layer attention maps. §E reports further details and results for the Image Quality Assessment (IQA) tasks. Finally, §F and §G offer comprehensive mathematical definitions, empirical estimations, and extended evaluations for our two analytical metrics: the averaging-efficiency statistic (η) and the spatial stationarity measure (ξ).

A Additional implementation details

A.1 Datasets

Table 3 lists the 19 classification datasets and Table 4 shows the three IQA datasets.

Following ALF (Ciernik et al., 2026), we group the classification datasets into four categories. **Natural multi-domain** datasets contain natural photographs of many object categories, close to the pre-training distribution of the backbones. **Natural single-domain** datasets contain natural images from one domain with fine-grained classes, such as pet breeds, aircraft variants and traffic signs. **Specialized** datasets contain images from non-natural sensors: histopathology slides, satellite and aerial images, and retinal fundus photographs. **Structured** datasets test texture, facial expression and the distance to objects in rendered 3D scenes.

Table 3: **Classification datasets.** Train and test are the official splits. Imbalance is the ratio between the largest and the smallest class in the training set.

Dataset	Task	Classes	Train	Test	Imbalance
Natural Multi-Domain					
STL-10 Coates et al. (2011)	Objects	10	5,000	8,000	1.0
CIFAR-10 Krizhevsky (2009)	Objects	10	45,000	10,000	1.0
Caltech-101 Fei-Fei et al. (2004)	Objects	102	2,753	6,085	1.3
VOC2007 Everingham et al. (2010)	Objects	20	7,844	14,976	20.6
CIFAR-100 Krizhevsky (2009)	Objects	100	45,000	10,000	1.1
ImageNet-1k Russakovsky et al. (2015)	Objects	1000	1,281,167	50,000	1.8
Country-211 Radford et al. (2021)	Geolocation	211	31,650	21,100	1.0
Natural Single-Domain					
Pets Parkhi et al. (2012)	Pet breeds	37	2,944	3,669	1.2
Flowers Nilsback & Zisserman (2008)	Flower species	102	1,020	6,149	1.0
Aircraft Maji et al. (2013)	Aircraft variants	100	3,334	3,333	1.0
GTSRB Stallkamp et al. (2012)	Traffic signs	43	26,640	12,630	10.0
SVHN Netzer et al. (2011)	House numbers	10	65,931	26,032	3.0
Specialized					
PCAM Veeling et al. (2018)	Histopathology	2	262,144	32,768	1.0
EuroSAT Helber et al. (2019)	Satellite	10	16,200	5,400	1.6
RESISC45 Cheng et al. (2017)	Aerial scenes	45	18,900	6,300	1.2
Retinopathy Dugas et al. (2015)	Fundus grading	5	35,126	42,670	36.5
Structured					
DTD Cimpoi et al. (2014)	Textures	47	1,880	1,880	1.0
FER2013 Goodfellow et al. (2013)	Facial expression	7	28,709	7,178	16.5
DMLab Beattie et al. (2016); Zhai et al. (2019)	3D distance	6	65,550	22,735	2.0

Table 4: **IQA datasets.** Summary of the evaluation benchmarks including dataset size, experimental splits, and mean opinion score (MOS) ranges.

Dataset	Distortions	Images	Train	Val	Test	MOS range
KADID10k (Lin et al., 2019)	Synthetic	10,125	6,125	2,000	2,000	$[1, 5]$
KonIQ10k (Hosu et al., 2020)	Authentic	10,073	6,043	2,014	2,016	$[1, 5]$
SPAQ (Fang et al., 2020)	Authentic	11,125	6,675	2,225	2,225	$[0, 100]$

The IQA datasets pair each image with a mean opinion score (MOS) from human raters, and the task is to regress that score. KADID10k is synthetic: 81 reference images, each degraded by 25 distortion types at 5 levels. KonIQ10k and SPAQ are authentic. KonIQ10k contains in-the-wild photographs with crowdsourced ratings, and SPAQ contains photographs taken with 66 smartphones.

A.2 Backbones

We evaluate seven Vision Transformers, grouped by pre-training paradigm into supervised, self-supervised and language-supervised models. Each group covers two to three model sizes. All backbones use their public pre-trained weights and stay frozen during probe training. Every multi-layer probe reads one intermediate representation per transformer block, which gives 12 layers for the small and base models and 24 for ViT-L/16. We list the pre-training data and weight source of each backbone below.

Supervised. ViT-S/16, ViT-B/16 and ViT-L/16 (Dosovitskiy et al., 2020), pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k.

Self-supervised. DINOv2-S/14 and DINOv2-B/14 (Oquab et al., 2023), pre-trained on LVD-142M without register tokens.

Language-supervised. CLIP-B/32 and CLIP-B/16 with the original OpenAI CLIP weights (Radford et al., 2021), loaded through OpenCLIP (Cherti et al., 2023), contrastively pre-trained on 400M image-text pairs.

A.3 Probe and baseline architectures

This section describes the cross-attention block and the classification head shared by all attentive probes, followed by how the baselines and SALF use them. Table 5 lists every probe with its parameter count on ViT-B/16. Table 6 shows the design space for layer probes.

Cross-attention block. The block takes n_q query vectors and n_{kv} key and value tokens of width d. Queries, keys and values pass through three separate LayerNorms, then through bias-free linear projections to H heads of width d_h. The attention output is mapped back to d by a linear layer with bias. For every probe except ATT we keep the head width fixed across backbones and scale the number of heads with d: 12 heads for $d = 384$ (ViT-S/16, DINOv2-S/14), 24 for $d = 768$ (ViT-B/16, DINOv2-B/14, CLIP-B/16 and CLIP-B/32) and 32 for $d = 1024$ (ViT-L/16). ATT uses 8 heads on every backbone, so its head width grows with d.

Classification head. The pooled d-dimensional vector passes through BatchNorm with a learned per-feature scale and shift, then a linear layer over the C classes.

Linear probe. The classification head above on the class token of the last layer, with no attention block.

Attentive baselines. Each baseline reads its tokens with one learned query through one cross-attention block with the setting listed in Table 5. ALF reads the class token and the mean patch token of every layer, $2L$ tokens. ATT reads the $N{+}1$ tokens of the last layer, with 8 heads as in the ATT baseline of ALF. Flat in §3 reads all $L(N{+}1)$ tokens. The single-layer probes of the layer sweep read either the $N{+}1$ tokens of one layer (attentive) or its class token and mean patch token (pooled).

VQT. VQT learns one query token per transformer layer, applies 0.1 dropout to the queries during training, and classifies from the concatenation of the final CLS token and every layer's query output. The pre-trained backbone remains frozen.

Table 5: Probe configurations on ViT-B/16 ($d = 768$, $L = 12$ layers, $N = 196$ patch tokens plus one class token per layer). Every attentive probe uses the cross-attention block and the BatchNorm-plus-linear head of Supp.§A.3. The difference lies in which tokens the queries read, and the number and width of the attention heads. Heads × width is the number of heads and the width of each head; the inner width of every stage is their product. Parameters are for a 10-class head.

Probe	Keys and values	Queries	Stages	Heads × width	Params
Linear probe	class token, last layer	–	0	–	7.7K
ATT	N+1 tokens, last layer	1	1	8×192	4.73M
ALF	CLS and mean patch, every layer	1	1	24×64	4.73M
FLAT	all $L(N+1)$ tokens	1	1	24×64	4.73M
SALF-P	N patches per layer; L summaries + L CLS	L+1	2	24×32, 24×32	4.75M
SALF-A	layer mean of the N+1 tokens	1	1	24×64	4.73M

SALF-P. The within-layer stage has one learned query per layer ($K = 1$). All L layers share a single summarizer block, so only the queries are layer-specific. Each query reads the N patch tokens of its layer, with the class token excluded, giving the summary $\boldsymbol{S}^{(\ell)} = \mathrm{Attn}(\boldsymbol{Q}^{(\ell)}, \boldsymbol{H}^{(\ell)}, \boldsymbol{H}^{(\ell)})$. A second block with a single query fuses $2L$ tokens: the class token and the summary of each layer. Both blocks have half the inner width of the ALF block, so the probe has about as many parameters as ALF.

SALF-A. The N+1 tokens of a layer, class token included, are averaged position-wise over the L layers. A single learned query reads the averaged tokens through one cross-attention block with 24 heads of width 64, as in ALF.

A.4 Training and hyperparameter search

Every number in the main tables comes from the protocol below, search and retraining included, which follows the linear-probing protocol of ALF (Ciernik et al., 2026). Ablation rows reuse the learning rate and weight decay selected for their control and run only the retraining stage.

Optimization. Tokens are L2-normalized along the feature dimension. We train with AdamW with a cosine schedule and no warmup, gradient clipping at norm 5 and a cross-entropy loss whose class weights are inversely proportional to the class frequencies of the training split.

Search. We hold out a stratified 20% of the training set as validation. The grid is the learning rate in $\{10^{-1}, 10^{-2}, 10^{-3}\}$ and the weight decay in $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, total of 21 candidates. Each candidate is trained for 20 epochs and scored by balanced accuracy on the validation split. Attention dropout is not searched; we reuse the rate that ALF selected for each dataset and backbone from $\{0, 0.1, 0.3\}$.

Retraining. The selected configuration is retrained on the full training set for 40 epochs, or for as many epochs as it takes to reach 1000 gradient steps on small datasets. No early stopping is used.

Cached training. The tokens of the frozen backbone are stored in half precision, and probes train on them with a batch size of 1024 unless noted otherwise. On datasets with fewer than five such batches, the batch size is reduced so that an epoch has at least five steps. Following ALF, we add Gaussian noise with $\sigma = 0.05$ to the cached tokens of a training sample with probability 0.5 to prevent overfitting.

End-to-end training. On PCAM, and for VQT on every dataset, the backbone runs inside the training loop. Images are resized to 256 and center-cropped to 224, with no augmentation and no feature noise. Training uses 8 GPUs with a batch size of 256 per GPU in mixed precision. VQT is trained on 1 GPU with batch size of 256 for all datasets except for PCAM, where 2 GPUs are used with the same global batch size.

Training on IQA datasets. For the backbone we fine-tune with LoRA (rank 4, α=8, dropout 0.05, applied to the query/key projections of the vision tower only) for 15 epochs, using a per-GPU batch size of 2 with 6 gradient-accumulation steps (effective batch size 12), Adam with learning rate 1e-4 for both the LoRA adapters and the fusion/regression head as in (Yadav et al., 2026). Input images are resized (no cropping) to 512×512, matching SigLIP2's native pretraining resolution. Each run

Table 6: **The design space of layer probes.** Rows indicate the source layers, while columns specify the spatial reduction strategy. SALF explores alternative attention parameterizations and aggregation orders within the intermediate-layer attentive setting.

Source layers	Pooled reduction	Attentive reduction
Last layer	Linear probe	ATT (Chen et al., 2024)
Intermediate layers	ALF (Ciernik et al., 2026)	VQT (Tu et al., 2023), **SALF-P/SALF-A**

occupies a single GPU. For within-dataset (in-domain) evaluation, each dataset is split 60/20/20 into train/validation/test; for datasets with multiple distorted images per reference (*e.g.*, KADID10K), the split is performed at the reference-image level so that no reference image's distortions appear in more than one split, preventing content leakage. For cross-dataset evaluation, the source dataset is split 80/20 into train/validation and the full target dataset is held out as test. In both settings, the checkpoint with the best validation SRCC is selected for test-set evaluation. All experiments are repeated over 3 random seeds (8, 19, 25) following (Yadav et al., 2026), and we report the mean ± standard deviation of test SRCC/PLCC across seeds.

B Additional results on classification tasks

B.1 Per-backbone results

Tables 7 to 10 list the balanced accuracy of every probe on each of the seven backbones averaged in Table 1. SALF-P on ViT-L/16 uses a batch size of 256 on cached features. SALF-P or SALF-A has the highest mean on all seven backbones. The better SALF variant is 2.12 to 3.77 pp above ALF and 0.40 to 1.20 pp above VQT on the mean balanced accuracy. The largest gains over ALF are on GTSRB, SVHN and Aircraft, datasets whose label depends on fine-grained spatial detail in a small region of the image.

The pre-training objective decides how much the last-layer patch tokens carry. On DINOv2-S/14 and DINOv2-B/14, whose patch-level objective keeps these tokens spatially informative, ATT is 2.08 and 0.92 pp above ALF and comes within 1.2 pp of SALF-P. Most of that margin comes from GTSRB, SVHN and DMLab. On the supervised and CLIP backbones, which are trained through the CLS token alone, ATT is within 0.3 pp of ALF or below it, and SALF leads ATT by 2.45 to 4.51 pp, the largest gap on CLIP-B/16. SALF reads both the spatial and the depth axis to achieve the highest mean on all seven backbones.

B.2 Results on MAE

MAE pretrains a ViT without labels, using a lightweight decoder to reconstruct the pixels of masked patches. The loss is computed on the patch tokens only. Consequently, its CLS token receives no direct training signal. We probe the ViT-B/16 encoder pretrained this way on ImageNet-1k, denoted MAE-B/16. Table 11 lists the balanced accuracy of every probe on it for the 18 datasets. The linear probe on the last-layer CLS token averages 41.07% on these 18 datasets, against 71.38 to 77.30% on the seven backbones of Table 1. ATT reads the patch tokens of the same final layer and reaches 75.30%, so the class information is in the patch tokens rather than the CLS token. SALF-P reads the patch tokens of every layer and has the highest mean, 78.97%, 3.67 points above ATT and 2.38 points above VQT.

B.3 Results on ImageNet-1k

Table 12 lists the balanced accuracy of the linear probe, ALF, ATT, SALF-P and SALF-A on ImageNet-1k for ViT-B/16, DINOv2-B/14, CLIP-B/16, ViT-S/16 and CLIP-B/32. All five probes are trained end-to-end with the same search and retraining protocols (Supp.§A.4). SALF-P has the highest mean, 82.03%, 0.76 pp above ALF and 1.02 pp above ATT, with the largest gain on CLIP-B/16, 1.77 pp over ALF. The linear probe is the best probe on ViT-B/16 and ViT-S/16, 0.13 pp and 0.45 pp above ALF. Both of these backbones are fine-tuned on ImageNet-1k with a linear classifier on the last-layer CLS token, so that token is already trained for the evaluation classes and attending over

Table 7: **Balanced accuracy (%) on the supervised ViT-S/16 and ViT-B/16.** Bold and underline mark the best and second-best probe of each backbone. †Trained end-to-end.

	ViT-S/16						ViT-B/16					
Dataset	Linear	ALF	ATT	VQT†	SALF-P	SALF-A	Linear	ALF	ATT	VQT†	SALF-P	SALF-A
STL-10	**99.30**	99.24	99.27	98.86	99.30	98.69	**99.61**	99.53	99.58	99.59	99.60	99.55
CIFAR-10	96.47	96.82	97.06	97.38	97.34	**97.46**	97.65	98.01	98.06	98.33	98.31	**98.37**
Caltech-101	94.90	95.85	95.56	95.87	95.50	**96.54**	95.40	96.46	96.03	96.59	96.56	**96.77**
VOC2007	87.45	87.56	85.78	**88.40**	86.75	86.94	88.03	89.13	87.57	**90.15**	89.07	89.37
CIFAR-100	84.39	86.79	86.75	87.52	**87.99**	87.89	88.10	89.85	89.34	90.92	90.61	**91.17**
Country-211	15.50	**18.43**	14.66	15.09	16.90	13.75	16.11	**20.39**	16.38	18.14	20.20	17.71
Pets	93.16	92.61	92.79	92.42	**93.36**	92.58	93.81	**94.70**	94.37	94.17	94.60	94.25
Flowers	99.33	99.20	**99.53**	99.09	98.98	99.41	99.26	99.49	99.53	99.19	**99.63**	99.53
Aircraft	42.35	50.14	50.41	58.08	64.02	**66.20**	51.87	59.70	59.02	58.92	66.46	**72.06**
GTSRB	62.15	78.21	84.32	95.83	**99.07**	98.90	64.35	81.79	85.91	97.76	99.11	**99.31**
SVHN	48.52	78.12	78.75	91.90	**95.81**	94.84	56.28	82.42	87.30	93.60	**96.50**	96.06
PCAM†	83.42	87.16	**88.45**	86.41	83.62	86.99	83.99	87.31	**88.85**	87.05	86.22	87.66
EuroSAT	95.80	98.03	97.17	98.42	**98.54**	98.44	95.43	98.40	97.58	98.50	98.75	**98.78**
RESISC45	88.57	93.86	92.83	93.99	95.66	**95.77**	88.08	95.33	94.43	95.54	96.17	**97.09**
Retinopathy	46.62	51.23	48.86	50.85	51.77	**52.75**	46.57	52.08	50.26	**52.57**	51.69	50.75
DTD	75.37	76.17	73.56	**77.55**	77.02	76.86	76.81	78.19	75.80	**79.57**	78.78	79.26
FER2013	53.22	63.89	63.62	63.33	**67.72**	66.61	55.75	66.61	66.27	66.27	**68.77**	67.64
DMLab	43.36	51.62	52.90	60.24	59.96	**62.15**	44.11	52.93	56.90	**63.12**	62.39	61.76
Mean	72.77	78.05	77.90	80.62	81.63	**81.82**	74.51	80.13	80.18	82.22	82.97	**83.17**

Table 8: **Balanced accuracy (%) on the supervised ViT-L/16.** Bold and underline mark the best and second-best probe. †Trained end-to-end.

	ViT-L/16					
Dataset	Linear	ALF	ATT	VQT†	SALF-P	SALF-A
STL-10	99.69	**99.79**	99.66	99.70	99.76	99.61
CIFAR-10	98.95	98.98	98.98	99.01	98.84	**99.01**
Caltech-101	96.92	96.99	97.05	**97.88**	97.34	97.48
VOC2007	89.26	90.77	90.43	**91.72**	90.92	90.78
CIFAR-100	91.68	92.81	92.00	93.56	93.26	**93.63**
Country-211	19.98	**22.38**	19.59	21.88	20.36	19.50
Pets	94.51	94.70	94.55	94.96	**95.36**	95.21
Flowers	99.56	99.58	99.55	99.58	99.66	**99.68**
Aircraft	44.59	57.19	62.92	66.70	69.49	**70.92**
GTSRB	70.41	85.39	89.58	97.64	96.52	**99.36**
SVHN	55.60	83.98	87.49	94.00	**95.87**	95.59
PCAM†	83.24	**87.63**	87.02	86.49	85.34	87.03
EuroSAT	96.06	98.03	97.74	98.58	98.60	**98.75**
RESISC45	90.53	95.60	94.82	95.98	96.75	**97.12**
Retinopathy	47.91	51.76	49.02	52.89	**52.98**	52.77
DTD	77.71	80.53	78.51	**81.49**	79.95	80.37
FER2013	60.16	**69.34**	67.47	67.99	68.59	68.71
DMLab	46.48	54.78	59.06	62.27	63.79	**63.95**
Mean	75.74	81.12	81.41	83.46	83.52	**83.86**

patch tokens or earlier layers does not yield substantial benefit. On the three backbones pre-trained without ImageNet-1k labels, the linear probe trails the best probe by 1.15 to 5.04 pp.

Table 9: **Balanced accuracy (%) on the self-supervised DINOv2-S/14 and DINOv2-B/14.** Bold and underline mark the best and second-best probe of each backbone. †Trained end-to-end.

	DINOv2-S/14						DINOv2-B/14					
Dataset	Linear	ALF	ATT	VQT†	SALF-P	SALF-A	Linear	ALF	ATT	VQT†	SALF-P	SALF-A
STL-10	99.15	**99.30**	99.24	98.95	99.20	98.83	99.58	**99.64**	99.48	99.47	99.52	99.33
CIFAR-10	96.18	96.79	96.93	**97.64**	97.30	97.04	98.09	98.35	98.02	**98.70**	98.32	98.49
Caltech-101	95.31	96.13	95.85	96.56	**96.59**	96.26	96.39	97.16	97.07	**97.67**	97.34	96.92
VOC2007	87.81	87.94	87.03	**88.62**	87.87	86.58	89.84	90.70	90.60	**91.44**	90.08	90.76
CIFAR-100	83.91	85.89	86.07	**87.93**	87.01	86.70	89.55	90.56	89.95	**91.86**	91.39	91.12
Country-211	16.00	**19.10**	16.93	17.63	17.91	14.32	19.20	**23.13**	20.27	20.63	22.60	18.35
Pets	**95.21**	95.06	94.04	95.01	93.96	92.76	95.28	**96.25**	94.55	94.32	95.71	93.58
Flowers	99.48	98.64	**99.49**	99.30	99.00	99.03	99.68	**99.75**	99.72	99.63	99.62	99.56
Aircraft	67.57	67.78	73.84	64.21	**74.79**	73.42	69.93	76.24	76.84	69.42	**83.35**	78.72
GTSRB	67.52	79.44	93.57	96.25	97.77	**99.45**	68.97	85.75	95.15	98.33	98.96	**99.57**
SVHN	53.99	79.84	88.96	91.66	**95.98**	95.11	56.66	85.60	91.49	93.47	**96.43**	96.28
PCAM†	83.26	87.83	**89.07**	88.97	87.06	88.87	83.49	88.07	89.22	**89.50**	86.12	88.50
EuroSAT	95.06	97.68	97.10	98.10	**98.29**	98.27	95.52	98.16	97.59	98.43	**98.73**	98.65
RESISC45	89.26	94.19	94.71	94.74	95.86	**96.28**	91.53	95.27	95.33	95.79	96.65	**96.72**
Retinopathy	46.65	52.35	51.50	51.18	50.09	**53.82**	48.47	53.45	51.82	**54.65**	51.55	54.42
DTD	78.30	78.88	77.93	**79.31**	79.15	76.86	80.32	82.02	82.18	**82.61**	82.07	82.45
FER2013	54.71	63.03	64.13	61.82	**68.60**	67.07	59.27	67.57	68.38	67.16	**70.25**	67.86
DMLab	43.51	52.64	63.54	62.69	64.83	**66.68**	49.59	59.34	65.81	66.47	66.41	**66.96**
Mean	75.16	79.58	81.66	81.70	**82.85**	82.63	77.30	82.61	83.53	83.86	**84.73**	84.35

Table 10: **Balanced accuracy (%) on the language-supervised CLIP-B/32 and CLIP-B/16.** Bold and underline mark the best and second-best probe of each backbone. †Trained end-to-end.

	CLIP-B/32						CLIP-B/16					
Dataset	Linear	ALF	ATT	VQT†	SALF-P	SALF-A	Linear	ALF	ATT	VQT†	SALF-P	SALF-A
STL-10	98.05	98.30	98.35	**98.62**	98.27	98.18	99.02	99.15	98.81	**99.24**	99.19	99.04
CIFAR-10	93.63	95.66	95.29	95.91	**96.56**	96.41	93.90	96.47	96.15	96.97	97.11	**97.55**
Caltech-101	92.82	94.10	91.24	**95.94**	94.73	95.64	93.93	94.80	93.48	96.21	96.14	**96.71**
VOC2007	84.19	85.26	84.15	**87.18**	85.09	84.70	85.36	86.85	84.87	**87.78**	86.41	86.98
CIFAR-100	74.97	83.08	81.45	84.17	**85.31**	85.29	77.48	85.43	82.81	85.68	87.23	**87.68**
Country-211	23.35	**27.26**	21.96	25.35	25.57	17.29	29.34	**33.44**	24.73	31.55	28.79	23.10
Pets	89.45	90.19	**91.25**	89.45	91.01	91.08	88.14	92.78	92.75	93.06	**94.99**	94.05
Flowers	87.99	93.26	94.42	95.43	**97.10**	96.39	95.32	96.79	96.48	96.60	98.07	**98.15**
Aircraft	40.25	47.73	56.10	54.26	60.28	**65.98**	49.54	56.18	58.93	60.66	**72.65**	72.09
GTSRB	78.33	86.25	90.96	95.73	96.98	**99.06**	75.39	84.72	88.65	97.10	98.29	**99.33**
SVHN	51.52	78.08	82.03	90.90	**94.63**	93.54	59.20	81.34	83.56	92.64	**95.47**	94.60
PCAM†	78.18	86.42	85.71	87.28	86.12	**87.97**	78.66	86.07	82.80	**88.66**	87.88	88.46
EuroSAT	90.03	97.98	96.59	98.11	**98.47**	98.26	89.88	97.82	96.92	98.09	**98.51**	98.45
RESISC45	88.58	94.79	93.49	94.22	95.80	**96.02**	90.92	95.62	94.03	95.97	**96.58**	96.31
Retinopathy	40.53	49.79	46.63	**50.71**	47.80	50.14	42.79	52.56	49.97	54.03	53.12	**54.18**
DTD	72.87	**78.88**	73.46	76.91	76.54	74.79	74.41	**80.59**	74.84	78.62	79.89	78.19
FER2013	59.65	67.74	67.16	68.89	**68.97**	67.43	64.25	70.49	69.51	71.94	**72.08**	71.70
DMLab	40.41	53.46	56.14	62.67	**63.62**	63.38	42.21	55.22	57.04	65.77	65.05	**67.35**
Mean	71.38	78.23	78.13	80.65	**81.27**	81.20	73.87	80.35	79.24	82.81	**83.75**	83.55

B.4 Computational Overhead

Table 13 reports the cache size, parameter count, training GFLOPs, peak training memory and inference GFLOPs of each probing method, of VQT and of full finetuning on ViT-B/16. This section expands the cost comparison of §5.2. The accuracies of the probes and of full finetuning are in

Table 11: **Balanced accuracy (%) on the self-supervised MAE-B/16.** Bold and underline mark the best and second-best probe. †Trained end-to-end.

Dataset	Linear	ALF	ATT	VQT†	SALF-P	SALF-A
STL-10	80.92	95.90	95.86	96.09	**96.77**	96.47
CIFAR-10	50.22	87.07	90.36	91.20	94.32	**94.37**
Caltech-101	49.28	89.97	91.48	92.22	**94.77**	93.79
VOC2007	49.66	**76.58**	74.23	74.53	76.10	75.66
CIFAR-100	26.35	69.69	73.92	73.77	78.40	**79.65**
Country-211	4.55	**13.09**	10.18	11.35	11.05	10.71
Pets	32.15	81.66	87.71	88.37	**91.35**	89.18
Flowers	44.10	87.21	87.13	91.59	**93.36**	93.16
Aircraft	9.53	34.06	50.74	54.98	**65.35**	62.19
GTSRB	20.47	70.13	93.54	97.70	98.89	**99.26**
SVHN	31.98	71.82	89.14	90.87	**95.24**	94.92
PCAM†	73.27	84.47	85.84	86.46	86.31	**88.30**
EuroSAT	75.15	96.94	97.50	97.70	98.10	**98.26**
RESISC45	53.53	92.03	93.01	93.45	95.45	**95.55**
Retinopathy	36.00	50.12	51.01	**51.47**	47.73	49.18
DTD	46.81	**69.79**	63.56	67.45	69.52	68.67
FER2013	27.79	52.44	61.20	60.41	**64.87**	62.85
DMLab	27.47	45.90	58.98	59.07	63.84	**63.98**
Mean	41.07	70.49	75.30	76.59	**78.97**	78.67

Table 12: **Balanced accuracy (%) on ImageNet-1k.** All probes are trained end-to-end. Bold and underline mark the best and second-best probe.

Backbone	Linear probe	ALF	ATT	SALF-P	SALF-A
ViT-B/16	**84.97**	84.84	84.50	84.64	83.61
DINOv2-B/14	83.71	84.21	84.65	**84.86**	83.47
CLIP-B/16	76.90	79.94	79.44	**81.71**	81.27
ViT-S/16	**81.41**	80.96	80.95	80.92	80.30
CLIP-B/32	72.97	76.40	75.49	**78.01**	76.59
Mean	79.99	81.27	81.01	**82.03**	81.05

Table 14, and those of VQT on ViT-B/16 are in Table 7. Parameters and GFLOPs are counted with a 10-class head on one 224×224 image, and the probe and baseline configurations are given in Supp.§A.3. Training GFLOPs count the multiply-accumulates of every matrix product and convolution in the forward and backward pass of one training step per image in mixed precision. For the cached probes the input is the cached tokens and the one-time feature extraction (17.56 GFLOPs per image) is not included, while full finetuning take images and run the backbone at every step. Training memory is the peak GPU memory allocated during one training step at batch size 256 on an RTX 6000 Ada (fp32 input, bf16 autocast), excluding the model weights and the input batch, with random inputs and without optimizer state. For the probes we input the cached tokens, while for full finetuning the input is 224×224 images. Peak memory grows linearly with the batch size for every row.

In general, SALF-A has similar cost as ATT, using 197 cached tokens, 4.73M parameters, 18.03 GFLOPs, and 1.18 GiB of training memory. This is because it is able to compute the parameter-free layer mean once to build the cache. It adds 0.47 GFLOPs to the backbone (2.7%) and trains in 1.18 GiB at batch size 256, 7.4× less than SALF-P and 11.8× less than FLAT. SALF-P reads the same 2,364 tokens as FLAT but adds 2.83 GFLOPs instead of 5.59 and trains in 8.75 GiB instead of 13.96, because its within-layer attention projects the tokens at half the inner width.

The backward pass of every probe costs twice its forward pass, so training GFLOPs per image are three times the probe's share of inference GFLOPs: 1.40 for SALF-A and ATT, 8.47 for SALF-P and 16.76 for FLAT, on top of the one-time feature extraction of 17.56 GFLOPs per image. VQT trains

Table 13: Computational overhead of probing methods, VQT and full finetuning on ViT-B/16 (12 layers, 197 tokens per layer). Cached tokens is the number of tokens stored per image when features are precomputed; SALF-A stores the parameter-free layer mean, so its cache equals ATT's. VQT can cache the input tokens of every layer plus the final CLS token; †the VQT accuracies in Table 1 use the end-to-end setup. Training GFLOPs is the measured forward plus backward compute of one training step per image, excluding the one-time feature extraction for cached methods. Training memory is the peak GPU memory of one step at batch size 256, excluding the model weights, the input batch and the optimizer state. Inference GFLOPs are for one 224×224 image, backbone included.

Method	Trainable params	Cached tokens per image	Training GFLOPs	Training memory (GiB)	Inference GFLOPs
Cached features					
Linear probe	7.7K	1	0.00	0.02	17.56
ATT	4.73M	197	1.40	1.18	18.03
ALF	4.73M	24	0.18	0.15	17.62
VQT	0.11M	2,365	2.97	2.17	17.65
FLAT	4.73M	2,364	16.76	13.96	23.15
SALF-P	4.75M	2,364	8.47	8.75	20.39
SALF-A	4.73M	197	1.40	1.18	18.03
End-to-end					
VQT†	0.11M	N/A	17.74	4.05	17.65
Full finetuning	85.8M	N/A	52.57	16.50	17.56

only 0.11M parameters. Its queries never enter the keys and values, so the image tokens of every layer are those of the frozen backbone and can be cached. Therefore, VQT stores the input of every layer plus the final CLS token, totalling to 2,365 tokens per image. Each training step recomputes the frozen keys and values from this cache and runs one query per layer. This costs 2.97 training GFLOPs per image, 2.1× SALF-A, and 2.17 GiB of training memory, 1.8× SALF-A. Most of that memory holds the keys and values of each layer, which are kept for the backward pass. Note that our VQT accuracy runs do not use this cache. They use the model code of Tu et al. (2023), which runs the frozen backbone at every step. The end-to-end path costs 17.74 training GFLOPs per image and 4.05 GiB peak memory.

At the parameter count of ALF and ATT, SALF-P and SALF-A gain 8.82 and 8.60 pp in mean balanced accuracy over the linear probe, against 5.84 and 5.80 pp (Table 14). Full finetuning (Supp.§B.5) trains 18× more parameters, backpropagates through the backbone at 52.57 GFLOPs per image and step, 6.2× SALF-P and 37.6× SALF-A, peaks at 16.50 GiB and gains 4.38 pp.

B.5 Probing vs. full finetuning

Table 14 compares the frozen-feature probes with full finetuning of the backbone on 17 datasets, averaged over ViT-B/16, DINOv2-B/14 and CLIP-B/16. Full finetuning updates all backbone parameters (*e.g.*, 85.8M on ViT-B/16), whereas probing methods train only a lightweight head on cached features.

Setup. Full finetuning trains the backbone and a linear head on the last-layer CLS token end-to-end, with the preprocessing of end-to-end training in Supp.§A.4. Following ALF, finetuning uses fixed hyperparameters of learning rate 10^{-3} and weight decay 0.1 for the head, learning rate 10^{-5} and weight decay 10^{-6} for the backbone, 40 epochs with at least 1000 steps, a batch size of 256 per GPU, and mixed precision training. Unlike the probes, full finetuning receives no hyperparameter search, and neither uses data augmentation. FLAT (§3) uses the same search and retraining protocol as the other probes, with a batch size of 256.

Results. Full finetuning improves the linear probe by 4.38 pp on average, against 8.82 pp for SALF-P, 8.60 pp for SALF-A, 8.32 pp for FLAT and 5.84 pp for ALF. It is above the better SALF variants on only three datasets, SVHN, DMLab and EuroSAT, by 0.83, 0.28 and 0.12 pp. ALF stays well below full finetuning on several datasets, by up to 15.03 pp on GTSRB. This suggests that, under this finetuning recipe, the information that full finetuning extracts by updating the backbone is already present in the frozen features, and spatial attention recovers it at 18× fewer trainable parameters

Table 14: **Probing a frozen backbone vs. full finetuning: balanced accuracy gains (pp) over the last-layer CLS linear probe on 17 datasets.** Mean ± standard deviation over ViT-B/16, DINOv2-B/14 and CLIP-B/16. Bold and underline mark the largest and second-largest improvement.

Dataset	Linear	Δ ALF	Δ ATT	Δ FLAT	Δ SALF-P	Δ SALF-A	Δ Finetune
STL-10	$99.40_{\pm 0.27}$	**+0.04**$_{\pm 0.09}$	-0.11$_{\pm 0.07}$	-0.02$_{\pm 0.03}$	+0.03$_{\pm 0.10}$	-0.10$_{\pm 0.11}$	-2.57$_{\pm 0.91}$
CIFAR-10	$96.55_{\pm 1.88}$	+1.06$_{\pm 1.07}$	+0.86$_{\pm 1.00}$	+1.32$_{\pm 1.22}$	+1.37$_{\pm 1.32}$	**+1.59**$_{\pm 1.46}$	+1.55$_{\pm 1.85}$
Caltech-101	$95.24_{\pm 1.01}$	+0.90$_{\pm 0.12}$	+0.29$_{\pm 0.52}$	+1.36$_{\pm 0.49}$	+1.44$_{\pm 0.55}$	**+1.56**$_{\pm 0.93}$	-1.84$_{\pm 0.25}$
VOC2007	$87.74_{\pm 1.84}$	+1.15$_{\pm 0.26}$	-0.06$_{\pm 0.58}$	+1.21$_{\pm 0.27}$	+0.78$_{\pm 0.38}$	**+1.29**$_{\pm 0.29}$	-4.70$_{\pm 3.06}$
CIFAR-100	$85.04_{\pm 5.38}$	+3.57$_{\pm 3.11}$	+2.32$_{\pm 2.15}$	+4.40$_{\pm 3.12}$	+4.70$_{\pm 3.58}$	**+4.95**$_{\pm 3.76}$	+3.79$_{\pm 4.24}$
Country-211	$21.55_{\pm 5.65}$	**+4.10**$_{\pm 0.14}$	-1.09$_{\pm 2.51}$	-0.23$_{\pm 2.65}$	+2.31$_{\pm 2.04}$	-1.83$_{\pm 3.27}$	-12.63$_{\pm 2.30}$
Pets	$92.41_{\pm 3.08}$	+2.17$_{\pm 1.75}$	+1.48$_{\pm 2.28}$	+2.04$_{\pm 3.01}$	**+2.69**$_{\pm 2.95}$	+1.55$_{\pm 3.20}$	-3.24$_{\pm 3.47}$
Flowers	$98.09_{\pm 1.96}$	+0.59$_{\pm 0.63}$	+0.49$_{\pm 0.48}$	+0.99$_{\pm 1.43}$	**+1.02**$_{\pm 1.24}$	+0.99$_{\pm 1.31}$	-0.01$_{\pm 0.52}$
Aircraft	$57.11_{\pm 9.11}$	+6.93$_{\pm 0.65}$	+7.82$_{\pm 1.12}$	+16.22$_{\pm 2.09}$	+17.04$_{\pm 4.32}$	**+17.18**$_{\pm 6.01}$	-1.58$_{\pm 6.07}$
GTSRB	$69.57_{\pm 4.53}$	+14.52$_{\pm 3.68}$	+20.33$_{\pm 5.35}$	+28.87$_{\pm 4.10}$	+29.22$_{\pm 4.87}$	**+29.83**$_{\pm 4.53}$	+29.55$_{\pm 4.37}$
SVHN	$57.38_{\pm 1.30}$	+25.74$_{\pm 2.79}$	+30.07$_{\pm 4.33}$	+38.06$_{\pm 1.79}$	+38.75$_{\pm 1.77}$	+38.27$_{\pm 2.03}$	**+39.58**$_{\pm 1.15}$
EuroSAT	$93.61_{\pm 2.64}$	+4.52$_{\pm 2.42}$	+3.75$_{\pm 2.32}$	+4.83$_{\pm 2.48}$	+5.05$_{\pm 2.53}$	+5.02$_{\pm 2.51}$	**+5.17**$_{\pm 2.70}$
RESISC45	$90.18_{\pm 1.50}$	+5.23$_{\pm 1.48}$	+4.42$_{\pm 1.39}$	+6.03$_{\pm 1.35}$	+6.29$_{\pm 1.29}$	**+6.53**$_{\pm 1.76}$	+3.96$_{\pm 1.42}$
Retinopathy	$45.94_{\pm 2.36}$	+6.75$_{\pm 2.14}$	+4.74$_{\pm 1.73}$	+5.95$_{\pm 0.68}$	+6.18$_{\pm 3.05}$	**+7.17**$_{\pm 3.07}$	-1.19$_{\pm 2.14}$
DTD	$77.18_{\pm 2.43}$	**+3.09**$_{\pm 2.19}$	+0.43$_{\pm 1.17}$	+2.55$_{\pm 0.36}$	+3.07$_{\pm 1.71}$	+2.79$_{\pm 0.71}$	-6.35$_{\pm 1.71}$
FER2013	$59.76_{\pm 3.49}$	+8.47$_{\pm 1.89}$	+8.30$_{\pm 2.22}$	+9.24$_{\pm 2.14}$	**+10.61**$_{\pm 2.13}$	+9.31$_{\pm 1.88}$	+4.54$_{\pm 1.04}$
DMLab	$45.30_{\pm 3.13}$	+10.53$_{\pm 1.80}$	+14.61$_{\pm 1.41}$	+18.65$_{\pm 2.71}$	+19.31$_{\pm 2.56}$	+20.05$_{\pm 3.60}$	**+20.33**$_{\pm 3.95}$
Mean	$74.83_{\pm 1.50}$	+5.84$_{\pm 0.44}$	+5.80$_{\pm 0.34}$	+8.32$_{\pm 0.56}$	**+8.82**$_{\pm 0.90}$	+8.60$_{\pm 1.05}$	+4.38$_{\pm 1.33}$

and without a backward pass through the backbone. We also observe full finetuning performs worse than the linear probe on 9 of the 17 datasets. Seven of these are the seven smallest training sets (Flowers, DTD, Caltech-101, Pets, Aircraft, STL-10 and VOC2007, 1,020 to 7,844 images), where full finetuning trains for 50 to 334 epochs to reach the 1000 minimum step requirement and likely overfits without augmentation or a hyperparameter search.

B.6 Additional ablations

Protocol of the ablation rows. Every ablation row in the paper (Table 15, Table 16 and Table 17) reuses the learning rate and weight decay that the search selected for its control, the default SALF-P or SALF-A on the same dataset and backbone, and runs only the retraining stage of Supp.§A.4.

Layer fusion. Table 15 varies how the L layers are combined. SALF-P fuses after the spatial attention step, over each layer's CLS token and K summaries. The default cross-attention lets a single learned query attend over all $L(K+1)$ tokens, which we compare with mean pooling and concatenation. SALF-A fuses before the spatial attention step, position by position over the full token grid. The default is mean pooling. Every ablation row reuses the hyperparameters selected for the default variant. We test three other variants: *depth attention* replaces the uniform mean by cross-attention over the L layer tokens at each position, with one learned query per position; *linear on CLS* and *linear on AP* score every layer from its own CLS token or its average-pooled patch token, and a softmax over these scores gives per-image layer weights.

Two observations stand out. First, for SALF-P, cross-attention gives the best overall performance, with mean pooling achieving similar mean accuracy (82.5 vs. 82.8). Concatenation performs the worst (81.8), with the largest drops on Aircraft (-5.3 pp) and GTSRB (-2.8 pp). We hypothesize that concatenation gives each layer its own block of classifier weights, so the head learns the class evidence separately for each of the $2L$ tokens. Cross-attention and mean pooling share one map across layers. On Aircraft, the concatenated head has $24\times$ the weights of the shared one (1.84M against 77K) and learns them from about 33 training images per class, likely too few samples to fit without overfitting. Second, the fusion method matters little for SALF-A. The four variants are within 0.3 pp of each other on the mean, and all of them are about 3 pp above ALF and ATT. The gain comes from the spatial attention, not from the fusion method, consistent with §3.

Table 15: Layer-fusion ablations on ViT-B/16 (balanced accuracy, %). Each block varies how the layers are combined and keeps everything else at the default (shaded); SALF-P fuses after the spatial attention step, SALF-A before it. All rows are retrained at the default's per-dataset hyperparameters. Mean is over the 17 cached datasets shown. Bold marks the best per block and column.

	Natural Multi-Domain						Natural Single-Domain					Specialized			Structured			
Method	STL-10	CIFAR-10	Caltech-101	VOC2007	CIFAR-100	Country-211	Pets	Flowers	Aircraft	GTSRB	SVHN	EuroSAT	RESISC45	Retinopathy	DTD	FER2013	DMLab	Mean
Linear probe	99.6	97.6	95.4	88.0	88.1	16.1	93.8	99.3	51.9	64.4	56.3	95.4	88.1	46.6	76.8	55.7	44.1	74.0
ALF	99.5	98.0	96.5	89.1	89.8	20.4	94.7	99.5	59.7	81.8	82.4	98.4	95.3	52.1	78.2	66.6	52.9	79.7
ATT	99.6	98.1	96.0	87.6	89.3	16.4	94.4	99.5	59.0	85.9	87.3	97.6	94.4	50.3	75.8	66.3	56.9	79.7
(a) Layer fusion in SALF-P (after spatial attention)																		
Cross-attention	**99.6**	98.3	96.6	**89.1**	90.6	**20.2**	94.6	**99.6**	66.5	**99.1**	**96.5**	**98.7**	**96.2**	**51.7**	**78.8**	**68.8**	62.4	**82.8**
Mean	**99.6**	**98.6**	**96.9**	88.9	**91.2**	17.5	**94.7**	99.5	**66.7**	98.9	96.2	**98.7**	**96.2**	49.0	78.7	68.4	62.2	82.5
Concat	**99.6**	98.3	96.3	88.5	91.0	19.1	93.7	97.7	61.2	96.3	95.9	98.6	95.7	49.5	77.4	68.6	**62.6**	81.8
(b) Layer fusion in SALF-A (before spatial attention)																		
Uniform mean	99.5	98.4	96.8	**89.4**	**91.2**	17.7	94.3	99.5	**72.1**	99.3	96.1	**98.8**	**97.1**	50.8	79.3	67.6	61.8	82.9
Depth attention	99.6	98.4	96.9	88.9	91.1	17.8	94.5	**99.6**	69.5	**99.5**	**96.5**	**98.8**	95.2	50.5	78.2	**68.9**	**67.4**	**83.0**
Linear on CLS	**99.7**	**98.5**	**97.0**	89.1	91.0	18.0	**94.8**	**99.6**	71.2	98.7	95.6	**98.8**	96.3	**50.9**	79.7	67.9	62.4	82.9
Linear on AP	99.6	98.3	96.9	**89.4**	**91.2**	**18.5**	94.5	**99.6**	71.6	97.5	94.1	**98.8**	95.5	50.6	**79.8**	68.4	62.2	82.7

Capacity of SALF-P. Table 16 varies the number of queries per layer, the head widths of the within-layer and across-layer attention, and whether the summarizer is shared across layers. Every variant has a mean between 82.3% and 83.1%, within 0.5 pp of the default and at least 2.6 pp above ALF and ATT (79.7%). The gain of SALF-P over the baselines does not depend on a particular capacity setting. Specifically, more queries per layer degrades the overall performance by 0.2 pp for $K{=}2$ and 0.5 pp for $K{=}4$, potentially due to overfitting. Second, the head width of the across-layer attention changes the mean by at most 0.1 pp. This agrees with Finding 3 in §3: once each layer is read with spatial attention, the depth fusion needs little capacity. Table 15 shows the same, where replacing the across-layer attention with a mean costs 0.3 pp for SALF-P and the four fusion methods of SALF-A are within 0.3 pp of each other. Third, a within-layer head width of 64 and per-layer summarizers raise the mean by 0.2 pp and 0.3 pp. Both gains come with more parameters: head width 64 raises the probe from 4.75M to 7.11M parameters (1.5×) and per-layer summarizers to 30.76M (6.5×), at unchanged FLOPs. The default keeps a single shared summarizer with head width 32.

Layer subsets. Table 17 changes which layers SALF-P and SALF-A read. Every 2nd (6 layers) and Every 3rd (4 layers) stay within 0.1 pp of all 12 layers on the mean for both methods. Last 6 reads as many layers as Every 2nd but loses 0.3 pp for SALF-P and 0.5 pp for SALF-A, of which GTSRB alone accounts for 6.1 pp and 5.0 pp, suggesting that this dataset depends on features of the early layers. This supports Findings 2 and 3 in §3: with spatial attention, fewer layers need to be aggregated across depth (Finding 3), but the layers that are kept must span the entire backbone depth (Finding 2).

C Additional motivation analysis

C.1 Layer sweep on all datasets and backbones

Figures 5 to 7 repeat Fig. 2 on ViT-B/16, DINOv2-B/14 and CLIP-B/16 for the 16 cached datasets of Table 1, in the same order (PCAM has no cached features). Accuracies are on the 15% holdout split. The pooling gap is widest on SVHN, DMLab, FGVC Aircraft and FER2013 on all three backbones, and narrowest on Country-211 and DTD, where the two curves cross. The best single all-token layer is rarely the last one: it lies between layers 5 and 11 on most of the datasets, and as early as layer 5 or 6 on SVHN, EuroSAT and DMLab. With pooled tokens, the all-layer ALF line sits above the best single-layer point on 10 to 13 of the 16 datasets, highlighting the importance of cross-layer fusion for pooled features. With all tokens, the FLAT line sits at or below the best single-layer point on approximately half of the 16 datasets, so once spatial tokens are kept, a single well-chosen layer already matches the full-depth probe.

Table 16: SALF-P capacity ablations on ViT-B/16 (balanced accuracy, %). Each block varies one component and keeps the others at the default (shaded); the baselines at the top are context for the scale of the differences and are not ranked. All SALF-P rows are retrained at the default's per-dataset hyperparameters, so they differ only in the varied component. Mean is over the 17 cached datasets shown. Bold marks the best per block and column. In (d), giving each layer its own summarizer raises the probe from 4.75M to 30.76M parameters at unchanged FLOPs.

	Natural Multi-Domain						Natural Single-Domain					Specialized			Structured			
Method	STL-10	CIFAR-10	Caltech-101	VOC2007	CIFAR-100	Country-211	Pets	Flowers	Aircraft	GTSRB	SVHN	EuroSAT	RESISC45	Retinopathy	DTD	FER2013	DMLab	Mean
Baselines																		
Linear probe	99.6	97.6	95.4	88.0	88.1	16.1	93.8	99.3	51.9	64.4	56.3	95.4	88.1	46.6	76.8	55.7	44.1	74.0
ALF	99.5	98.0	96.5	89.1	89.8	20.4	94.7	99.5	59.7	81.8	82.4	98.4	95.3	52.1	78.2	66.6	52.9	79.7
ATT	99.6	98.1	96.0	87.6	89.3	16.4	94.4	99.5	59.0	85.9	87.3	97.6	94.4	50.3	75.8	66.3	56.9	79.7
(a) Queries per layer K																		
1	**99.6**	**98.3**	**96.6**	**89.1**	90.6	**20.2**	94.6	**99.6**	**66.5**	**99.1**	96.5	**98.7**	96.2	**51.7**	78.8	**68.8**	62.4	**82.8**
2	**99.6**	**98.3**	96.4	88.8	**90.8**	19.1	94.7	**99.6**	65.7	97.5	**96.6**	**98.7**	96.2	51.3	78.8	68.6	**62.5**	82.5
4	99.5	98.2	96.4	88.4	90.6	19.0	**94.8**	99.5	65.2	98.8	96.3	98.6	**96.3**	48.4	**79.1**	68.4	61.9	82.3
(b) Within-layer head dim																		
16	**99.6**	**98.4**	**96.6**	88.9	90.7	19.5	**95.2**	99.4	64.9	98.7	96.4	**98.8**	96.1	**53.5**	**79.0**	68.7	60.9	82.7
32	**99.6**	98.3	**96.6**	**89.1**	90.6	**20.2**	94.6	**99.6**	66.5	**99.1**	96.5	98.7	**96.2**	51.7	78.8	**68.8**	**62.4**	82.8
64	**99.6**	**98.4**	**96.6**	**89.1**	**90.9**	19.4	94.8	**99.6**	**70.2**	98.9	**96.6**	98.7	**96.2**	52.9	78.8	68.3	62.2	**83.0**
(c) Across-layer head dim																		
16	**99.6**	**98.4**	**96.9**	**89.1**	**90.9**	18.5	**94.9**	99.5	**67.9**	**99.1**	**96.6**	**98.8**	**96.2**	50.9	78.9	**69.5**	62.0	82.8
32	**99.6**	98.3	96.6	**89.1**	90.6	20.2	94.6	**99.6**	66.5	**99.1**	96.5	98.7	**96.2**	51.7	78.8	68.8	62.4	82.8
64	**99.6**	98.3	96.6	88.9	**90.9**	**20.8**	**94.9**	**99.6**	66.6	98.7	96.2	98.7	**96.2**	**52.4**	**80.2**	68.6	**62.5**	**82.9**
(d) Summarizer sharing																		
Shared	**99.6**	98.3	**96.6**	89.1	90.6	20.2	94.6	**99.6**	66.5	**99.1**	**96.5**	**98.7**	96.2	51.7	78.8	68.8	**62.4**	82.8
Per layer	**99.6**	**98.4**	**96.6**	**89.3**	**91.4**	**20.4**	**94.9**	**99.6**	**70.5**	98.7	96.0	98.5	**96.6**	**51.9**	**79.5**	**68.9**	61.5	**83.1**

Table 17: Layer-subset ablation on ViT-B/16 (balanced accuracy, %). Each block keeps the aggregator fixed and changes only which layers it reads; the final layer is always included. Layers are numbered 1 to 12, with 12 the final layer: All 12 reads layers 1-12, Every 2nd reads 2, 4, . . . , 12, Every 3rd reads 3, 6, 9, 12, and Last 6 reads 7-12. All subset rows are retrained at the all-layer default's per-dataset hyperparameters. Mean is over the 17 cached datasets shown. Bold marks the best per block and column.

	Natural Multi-Domain						Natural Single-Domain					Specialized			Structured			
Method	STL-10	CIFAR-10	Caltech-101	VOC2007	CIFAR-100	Country-211	Pets	Flowers	Aircraft	GTSRB	SVHN	EuroSAT	RESISC45	Retinopathy	DTD	FER2013	DMLab	Mean
Baselines																		
Linear probe	99.6	97.6	95.4	88.0	88.1	16.1	93.8	99.3	51.9	64.4	56.3	95.4	88.1	46.6	76.8	55.7	44.1	74.0
ALF	99.5	98.0	96.5	89.1	89.8	20.4	94.7	99.5	59.7	81.8	82.4	98.4	95.3	52.1	78.2	66.6	52.9	79.7
ATT	99.6	98.1	96.0	87.6	89.3	16.4	94.4	99.5	59.0	85.9	87.3	97.6	94.4	50.3	75.8	66.3	56.9	79.7
(a) SALF-P																		
All 12	**99.6**	98.3	96.6	89.1	90.6	**20.2**	94.6	**99.6**	66.5	**99.1**	**96.5**	98.7	96.2	51.7	78.8	68.8	62.4	82.8
Every 2nd	**99.6**	98.3	**96.8**	89.1	90.6	19.6	94.8	99.5	66.1	98.6	**96.5**	98.6	96.2	53.4	**79.8**	**69.4**	61.6	**82.9**
Every 3rd	**99.6**	**98.4**	96.5	89.0	**90.8**	19.2	**95.1**	**99.6**	65.5	98.3	96.2	**98.8**	96.3	52.3	79.3	69.0	**62.6**	82.7
Last 6	**99.6**	**98.4**	96.6	**89.3**	90.7	19.9	94.8	**99.6**	**67.1**	93.0	95.8	98.5	**96.4**	**53.5**	79.0	68.8	61.8	82.5
(b) SALF-A																		
All 12	99.5	98.4	96.8	**89.4**	**91.2**	17.7	94.3	99.5	**72.1**	**99.3**	**96.1**	98.8	**97.1**	50.8	79.3	67.6	61.8	**82.9**
Every 2nd	99.6	**98.5**	**96.9**	89.1	91.1	17.9	94.4	99.5	70.9	**99.3**	**96.1**	**98.9**	96.7	50.8	79.3	68.0	**62.8**	**82.9**
Every 3rd	99.6	98.3	96.8	89.2	91.1	18.3	**94.9**	**99.6**	69.6	99.2	**96.1**	98.8	96.6	**50.9**	79.4	**68.7**	62.4	**82.9**
Last 6	**99.7**	**98.5**	96.8	89.3	**91.2**	**18.5**	94.8	**99.6**	70.7	94.3	94.7	98.5	96.7	50.1	**79.9**	68.1	60.2	82.4

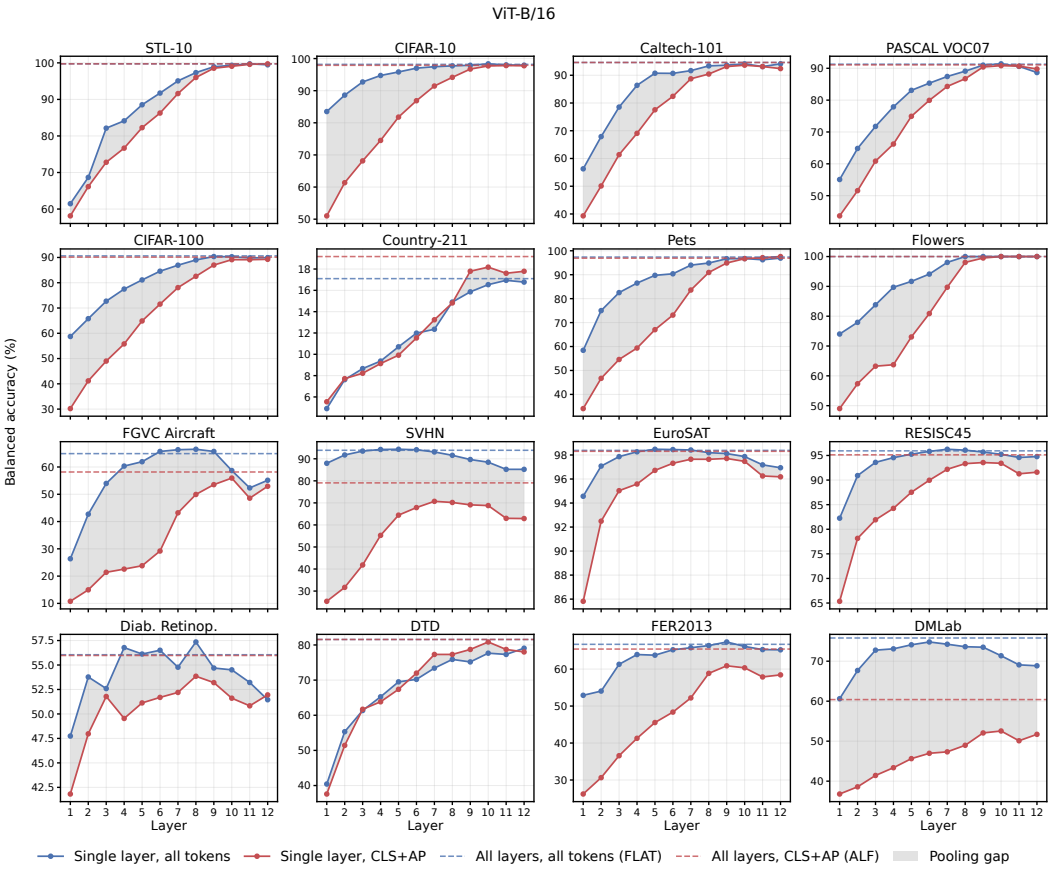

Figure 5: **Single-layer probes across depth on ViT-B/16, 16 datasets.** Same layout as Fig. 2: balanced accuracy (%) on the 15% holdout of a probe trained on one layer, attending over all tokens (blue) or over the CLS and average-pooled tokens (red), with the all-layer FLAT (blue dashed) and ALF (red dashed) probes as references. Datasets follow the order of Table 1.

C.2 One spatial map for all layers

SALF-A averages the tokens over depth before a single spatial attention (Eq. (4)), so it can select only one set of spatial positions per image, shared by every layer. This section tests that assumption on the unconstrained FLAT probe of §3, which is free to attend to a different set of positions in every layer.

Setup. We take the trained FLAT probe of each dataset and backbone and change only its attention weights at inference; no parameter is retrained. FLAT attends with one query over the $L(N+1)$ tokens of all layers, so its post-softmax weights split into a class-token weight and a patch map for every layer ℓ. In Table 18, *Mass-weighted* and *plain mean* replace every layer's patch map with the same map, computed per image and per head: the average of the L per-layer maps weighted by the attention mass each layer receives, or their unweighted mean. The mass of each layer and its class-token weight are kept, so only the spatial distribution within a layer changes. *Uniform* is the control that spreads each layer's mass evenly over its N patches. Accuracies are on the 15% holdout split of the training set (§3) on the 17 datasets with cached features, for ViT-B/16, DINOv2-B/14 and CLIP-B/16.

Results. Table 18 lists the balanced accuracy of the trained probe and the change under each of the three interventions. The mass-weighted map changes the 17-dataset mean by 0.01 pp on ViT-B/16, 0.05 pp on DINOv2-B/14 and 0.16 pp on CLIP-B/16, and stays within 0.5 pp of the trained probe on most of the 17 datasets; the largest change is -2.45 pp on Flowers for CLIP-B/16, whose holdout split has about 150 images. The uniform map loses 13.0, 18.8 and 21.2 pp on the mean, with 28 to 34 pp on GTSRB, 42 to 72 pp on SVHN and 50 to 58 pp on DMLab. It stays within

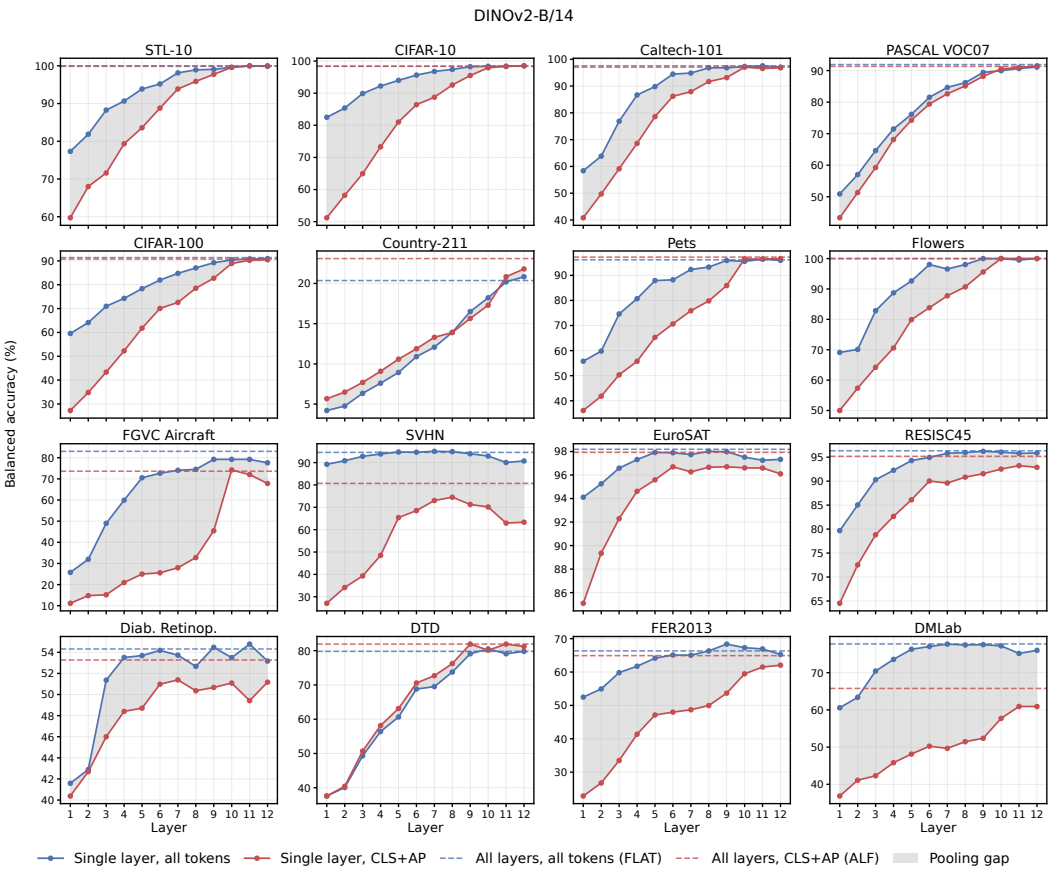

Figure 6: **Single-layer probes across depth on DINOv2-B/14, 16 datasets.** Same layout as Fig. 5.

1.5 pp on every backbone only on STL-10, VOC2007 and DTD. The plain mean sits between the mass-weighted and the uniform map, matching the probe on most datasets but falls well below it on Aircraft, DMLab, FER2013 and SVHN. These are the datasets where the probe reads mostly the early and middle layers (Fig. 2), and the maps of the low-mass layers pull an unweighted mean away from those positions. Spreading each layer's attention evenly over its patches (uniform map) degrades performance drastically, while sharing one mass-weighted selection across all layers maintains a similar level of accuracy. SALF-A has this structure by design: it averages the tokens over depth before its single spatial attention and selects one set of patches for all layers, while performing on par with SALF-P, which selects them separately in each layer.

D Additional analysis on attention maps

D.1 Cross-layer attention per dataset on additional backbones

Fig. 8 shows how the cross-layer query of SALF-P distributes its attention across layers, over the CLS and summary tokens of each layer, for each of the 17 datasets on ViT-B/16, DINOv2-B/14 and CLIP-B/16 (top), and on three additional backbones, ViT-S/16, DINOv2-S/14 and CLIP-B/32 (bottom). On the base backbones, attention rises with depth and the first four layers are nearly ignored, but the final layer is rarely the most-attended one. The summary token holds more mass than CLS on most datasets, and its share is largest on the Structured group, whose labels depend on texture or layout (DTD, FER2013, DMLab). Object-centric datasets such as Pets, Flowers and Caltech-101 read CLS in the last few layers on ViT-B/16 and DINOv2-B/14, while CLIP-B/16 reads the summary token from the mid-to-deep layers on nearly every dataset. The datasets whose label depends on a small region of the image (*e.g.*, Aircraft, GTSRB, DMLab) place higher emphasis on summary tokens from the middle of the network. In the middle layers, ViT attention still includes local heads, while

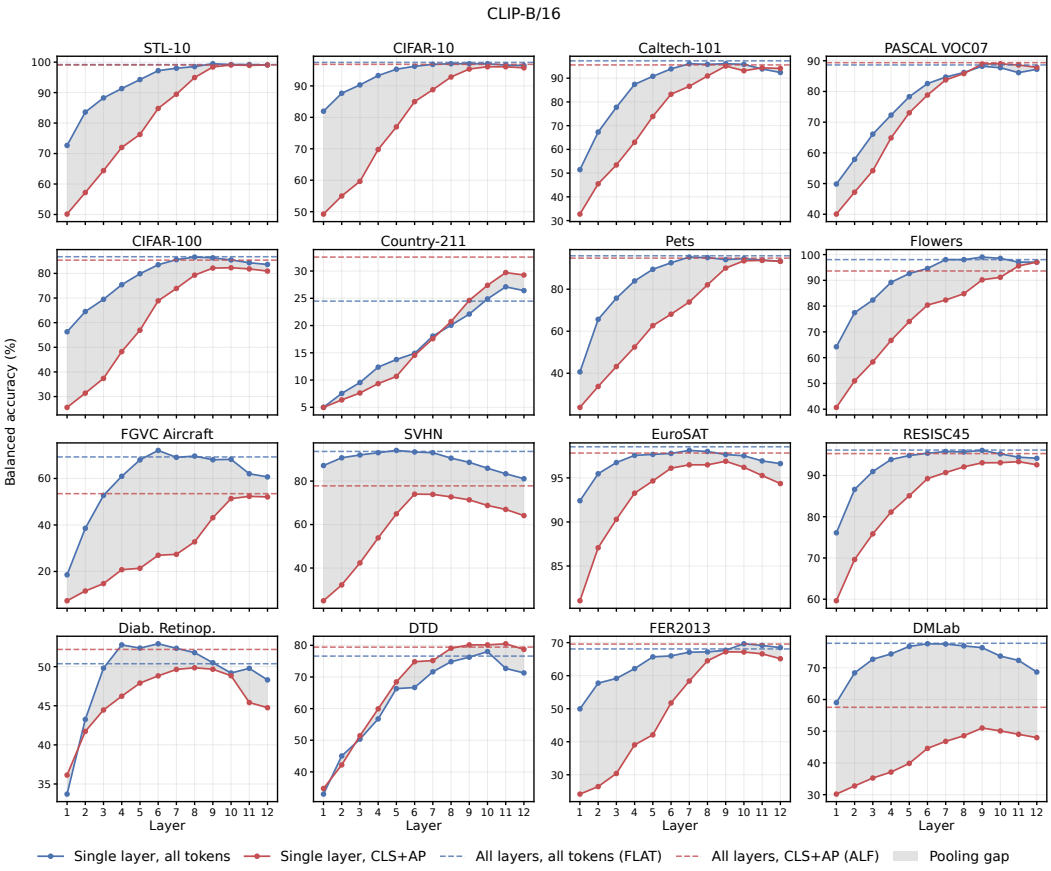

Figure 7: **Single-layer probes across depth on CLIP-B/16, 16 datasets.** Same layout as Fig. 5.

in the last layers every token attends to the whole image. The patch tokens at the relevant region therefore describe that region more specifically in the middle layers, which is where the summarizer finds the class evidence. The small backbones follow the same pattern.

The CLS profile of DINOv2 differs from the other models, with the CLS mass concentrated on the last two layers instead of spreading over the deeper half of the network. DINOv2 trains its patch tokens with a patch-level objective in addition to the image-level objective on the CLS token, so the patch tokens carry the image content at every depth. We hypothesize that its CLS token therefore gathers an image-level summary only in the last layers, where the image-level objective acts, while the summary tokens take the weight earlier.

D.2 Spatial attention maps

All backbones, with probabilities. Fig. 9 repeats Fig. 4 on ViT-B/16, DINOv2-B/14 and CLIP-B/16, with the input shown separately, a fourth example from Aircraft, and the probability p each probe assigns to the ground-truth class. On ViT-B/16, ALF assigns the true class 0.12 to 0.33, while both SALF variants assign it 0.95 or more. The test examples are picked separately for each backbone, showing cases where ALF misclassifies and both SALF variants classify correctly. Similar to ViT-B/16, the maps concentrate on the few patches that carry the class evidence, and the SALF-P and SALF-A maps select similar patches. The exception is SVHN on CLIP-B/16, where the most-attended layer of SALF-P is the final one and its map is nearly uniform instead of focusing on the class of interest. Therefore, a natural question to ask is what the attention map for other layers look like.

All layers. Fig. 10 shows the SALF-P summarizer map at every layer for the same twelve images, together with the weight w assigned to each layer's summary token. In the first layers, the attention map spreads over most of the image and w is small. From the middle layers on, the map concentrates

Table 18: One spatial map for all layers (balanced accuracy on the 15% holdout split, %). Each block is the trained FLAT probe of one backbone. The shaded row is the probe as trained; the rows below give the change in pp when its attention maps are replaced at test time, no parameter retrained. *Mass-weighted* gives every layer the same patch map, the average of its per-layer maps weighted by each layer's attention mass; *plain mean* uses their unweighted mean. Both are computed per image and head. *Uniform* spreads each layer's attention mass evenly over its patches. All keep every layer's mass and class-token weight. Mean is over the 17 datasets shown.

	Natural Multi-Domain						Natural Single-Domain					Specialized			Structured			
Attention map	STL-10	CIFAR-10	Caltech-101	VOC2007	CIFAR-100	Country-211	Pets	Flowers	Aircraft	GTSRB	SVHN	EuroSAT	RESISC45	Retinopathy	DTD	FER2013	DMLab	Mean
(a) ViT-B/16																		
As trained	99.7	98.2	94.6	91.3	90.6	17.1	97.4	100.0	64.9	100.0	93.9	98.4	95.9	56.1	81.6	66.7	75.8	83.7
Mass-weighted	+0.0	+0.1	-0.2	+0.2	+0.1	+0.0	-0.5	+0.0	+0.6	+0.0	+0.1	+0.0	+0.2	+0.3	-0.7	+0.1	-0.4	+0.0
Plain mean	-2.1	-0.2	+0.2	-0.2	-0.3	-0.2	+0.4	+0.0	-7.0	+0.0	-2.1	+0.0	-0.3	+0.2	-1.8	-2.2	-4.7	-1.2
Uniform	-0.1	-1.4	-0.5	-0.3	-5.0	-2.3	-0.7	+0.0	-45.3	-27.8	-41.8	-8.9	-16.7	-2.2	-1.1	-16.0	-50.1	-13.0
(b) DINOv2-B/14																		
As trained	99.9	98.4	97.1	91.9	91.4	20.4	96.2	100.0	83.0	100.0	94.6	98.2	96.3	54.3	79.8	66.3	77.7	85.0
Mass-weighted	+0.0	+0.0	+0.2	-0.3	+0.0	-0.4	+0.0	+0.0	+0.6	+0.0	-0.1	+0.0	+0.1	+0.4	+0.0	-0.4	-1.0	+0.0
Plain mean	+0.0	-0.2	+0.5	-0.5	-0.6	-0.2	-0.5	-0.5	-13.2	+0.0	-1.2	-0.5	-0.8	-0.9	+1.1	-2.6	-6.7	-1.6
Uniform	+0.0	-1.4	-1.5	-0.8	-6.1	-6.0	-10.9	-0.5	-69.1	-34.2	-72.0	-13.4	-19.7	-8.9	+1.4	-20.2	-56.5	-18.8
(c) CLIP-B/16																		
As trained	99.2	97.5	97.3	88.6	86.8	24.5	95.9	98.0	69.3	99.8	93.6	98.5	96.2	50.4	76.6	68.1	77.7	83.4
Mass-weighted	+0.0	+0.0	-0.2	+0.1	-0.1	+0.3	+0.2	-2.5	+0.0	+0.0	+0.0	-0.1	-0.2	-0.2	-0.7	+0.6	-0.1	-0.2
Plain mean	+0.0	-0.5	-1.5	+0.4	-1.7	+1.5	-1.6	-1.0	-16.9	-0.1	-2.9	-0.7	-0.7	-5.4	-1.1	-2.1	-6.3	-2.4
Uniform	-0.8	-7.0	-14.7	-0.3	-31.5	-1.8	-26.3	-6.9	-59.1	-29.6	-68.8	-13.0	-17.0	-18.9	+1.4	-8.0	-58.1	-21.2

Table 19: **In-domain results with the SigLIP2 backbone.** We report the mean and standard deviation of the Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) over 3 random seeds. The best performance for each metric is highlighted in bold.

	KADID10K		**KonIQ10K**		**SPAQ**	
Method	SRCC (↑)	PLCC (↑)	SRCC (↑)	PLCC (↑)	SRCC (↑)	PLCC (↑)
B_Gated	$0.841_{\pm 0.039}$	$0.840_{\pm 0.035}$	$0.928_{\pm 0.011}$	$0.943_{\pm 0.008}$	$0.914_{\pm 0.001}$	$0.918_{\pm 0.000}$
ALF	$0.910_{\pm 0.013}$	$0.912_{\pm 0.011}$	$0.944_{\pm 0.004}$	$0.954_{\pm 0.001}$	$0.925_{\pm 0.005}$	$0.928_{\pm 0.006}$
SALF-P	$\mathbf{0.924_{\pm 0.011}}$	$\mathbf{0.924_{\pm 0.011}}$	$\mathbf{0.948_{\pm 0.001}}$	$0.956_{\pm 0.001}$	$\mathbf{0.928_{\pm 0.001}}$	$\mathbf{0.932_{\pm 0.001}}$
SALF-A	$0.917_{\pm 0.022}$	$0.921_{\pm 0.022}$	$0.948_{\pm 0.002}$	$\mathbf{0.957_{\pm 0.002}}$	$0.923_{\pm 0.003}$	$0.925_{\pm 0.002}$

on the object and adjacent layers select nearly the same patches. Layers 5 to 10 receive a larger share of the attention than layers 1 to 4 or layers 11 and 12. These findings suggest that the cross-layer query puts most of its weight on intermediate layers rather than the last one, as in Finding 2. Furthermore, once the map has formed, it selects the same positions at every depth, so one spatial map can serve all layers. This is the structure SALF-A imposes by averaging over depth before a single spatial attention, and it explains why that average loses little accuracy (Finding 3, Supp.§C.2).

E ADDITIONAL RESULTS ON IQA TASKS

As shown in Table 19, we evaluate our methods using the SigLIP2 backbone across three in-domain NR-IQA benchmarks. SALF-P consistently outperforms the B_Gated and ALF baselines on nearly all metrics and achieves the highest overall correlations on KADID10K and SPAQ. Meanwhile, the highly efficient SALF-A variant remains strictly competitive, surpassing the ALF baseline on KonIQ10K and obtaining the highest PLCC for that dataset. These in-domain results clearly validate two critical architectural designs. First, aggregating features across multiple layers significantly outperforms single-layer readouts, as evidenced by the consistent gains over the B_Gated baseline. Second, actively preserving fine-grained spatial representations provides a distinct performance advantage over the static average pooling utilized by ALF.

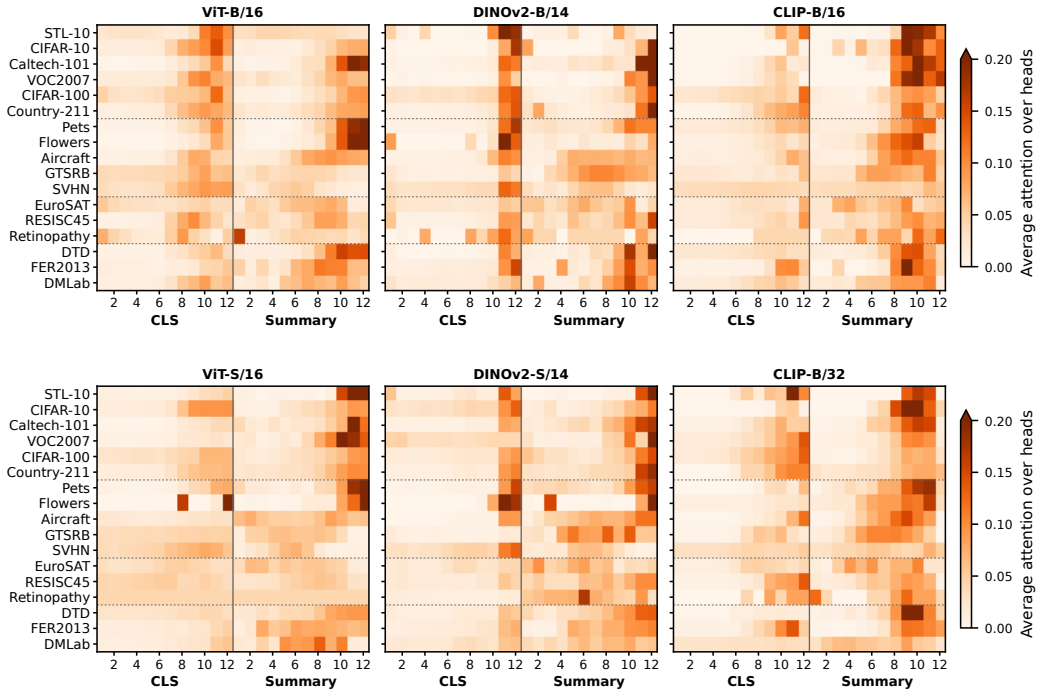

Figure 8: **Where SALF-P's cross-layer query looks, per dataset.** Attention weights of the cross-layer query over the CLS and summary token of each layer, averaged over heads and test images, on ViT-B/16, DINOv2-B/14 and CLIP-B/16 (top) and on ViT-S/16, DINOv2-S/14 and CLIP-B/32 (bottom).

Table 20: **Balanced-accuracy (BA) gains over ALF.** Mean ± standard deviation. n: model-task (dataset-backbone) pairs; SALF-A: SA; SALF-P: SP.

Group	n	$\mathcal{A}_b(\text{ALF}, \text{SA})$	$\mathcal{A}_b(\text{ALF}, \text{SP})$
HAA	21	$+13.5_{\pm 4.5}$	$+12.9_{\pm 4.5}$
All η_S<0.5	26	$+11.6_{\pm 6.3}$	$+11.4_{\pm 5.6}$
Non-HAA	81	$+0.4_{\pm 3.3}$	$+0.8_{\pm 2.4}$

F AVERAGING EFFICIENCY OF THE FROZEN FEATURES

This appendix gives the full definition of the averaging-efficiency statistic η used in §5.3, starting with a more complete overview, then details on its estimation, its limitations, and the complete per-dataset and per-family results.

F.1 OVERVIEW

Can We Predict The Cost of Spatial Averaging from Feature Statistics? We hypothesize that SALF outperforms ALF when averaging destroys task-relevant fine-grained spatial cues. We ask whether this can be predicted from feature statistics, without training probes. For features F and an axis $a \in \{L, S\}$ (layers or space), let $w \in \mathbb{R}^{N_a}$ linearly aggregate F along a. Following the classical Fisher discriminant (Fisher, 1936), let $\text{SNR}_a(w)$ be the ratio of inter-to-intra-class variation after aggregation (Supp.§F.3). The *averaging efficiency*, $\eta_a = \text{SNR}_a(\vec{1})/(\max_w \text{SNR}_a(w)) \in [0, 1]$, compares *uniform* averaging with the best w. Thus, low η_a indicates that averaging substantially reduces SNR, whereas high η_a suggests little benefit from reweighting. Let $\mathcal{A}_b(X, Y)$ denote the balanced-accuracy (BA) gain of Y over X. We find that η_S *correlates strongly with* $\mathcal{A}_b(\text{ALF}, \text{SALF-P})$ (Pearson $r = -0.84$). Further, since SALF-A averages uniformly over layers, its largest gains should occur when spatial averaging is inefficient but layer averaging is not. We therefore define *high averaging*

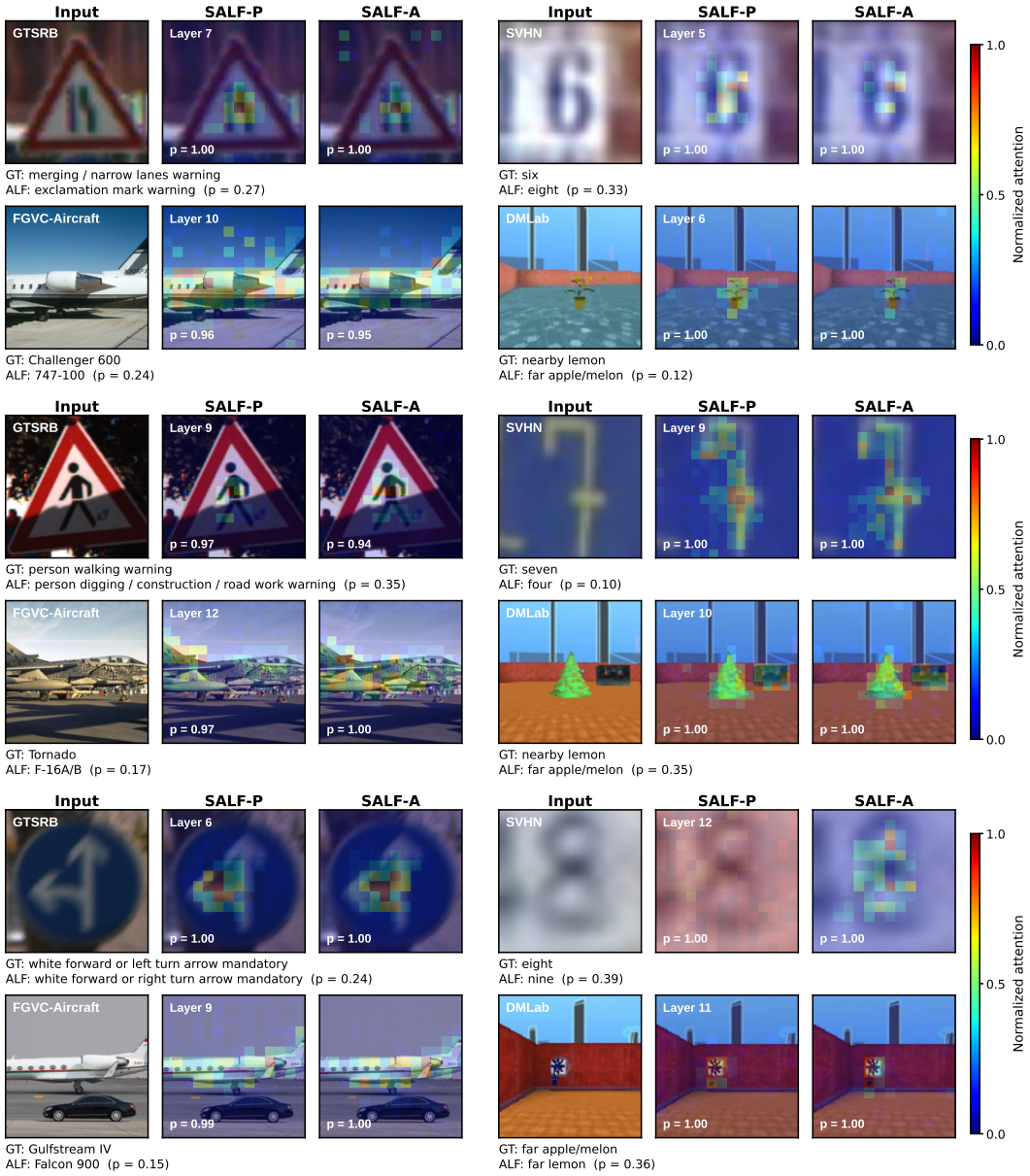

Figure 9: **Where the spatial query looks on ViT-B/16 (top), DINOv2-B/14 (middle) and CLIP-B/16 (bottom).** Test images that ALF misclassifies and both SALF variants classify correctly on each backbone, with the input, the SALF-P summarizer map at the most-attended layer and the SALF-A map over the fused tokens. Maps are averaged over heads and normalized; each square is one 16×16 patch for ViT-B/16 and CLIP-B/16 and one 14×14 patch for DINOv2-B/14. p is the probability assigned to the ground-truth class; ALF's prediction and p are given under each image.

asymmetry (HAA) tasks by $\eta_S < 0.5$ and $\eta_L > 0.5$. Although only a fraction of dataset–backbone pairs are HAA, their mean BA gain is more than $10\times$ that of non-HAA pairs (Table 20).

F.2 Setting and Notation

For image i with label $y_i \in \mathcal{Y}$, transformer layer $\ell \in \{1, \ldots, N_L\}$ and spatial token position $s \in \{1, \ldots, N_S\}$, let $h_{i\ell s} \in \mathbb{R}^d$ denote the set of feature tokens used to construct the statistic per dataset. In practice, (i) the CLS token is included (though note ALF gives special treatment to it) and

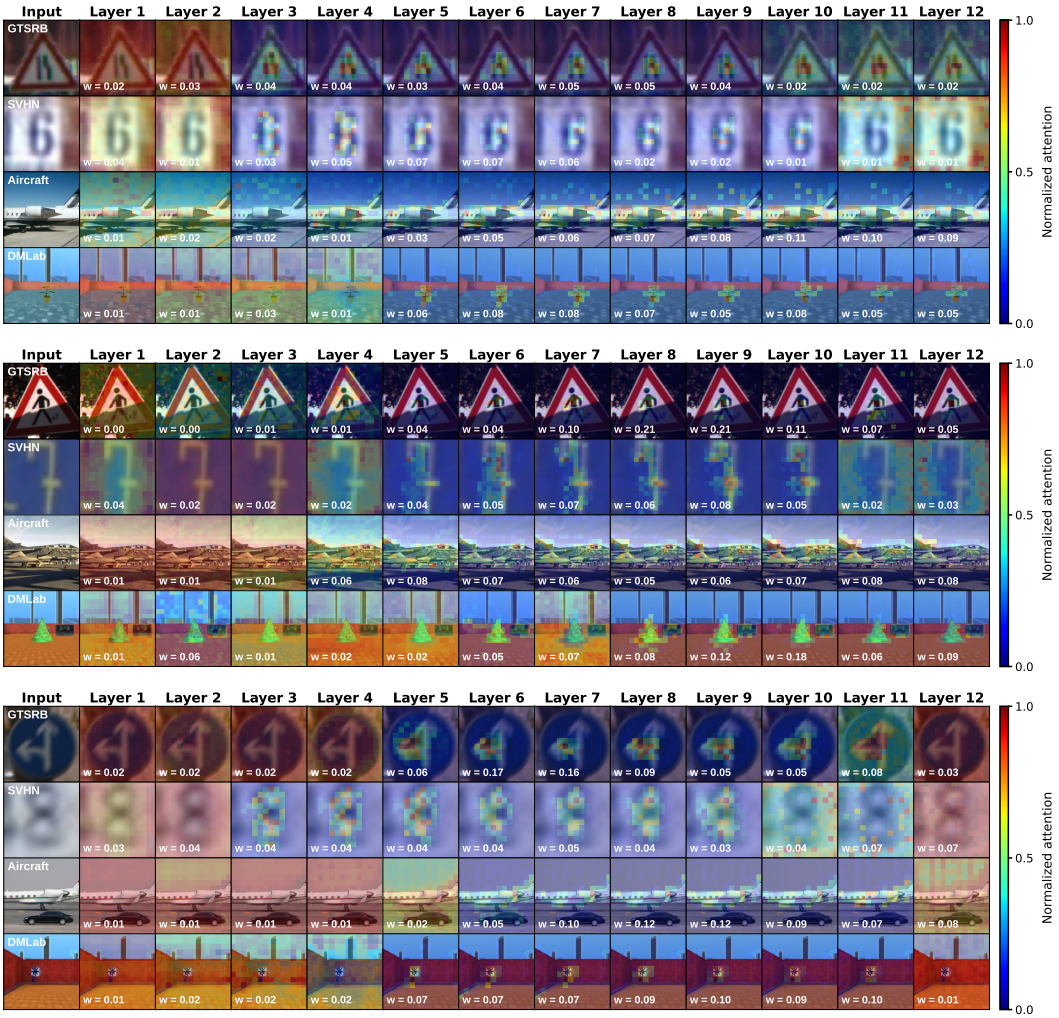

Figure 10: **SALF-P summarizer map at every layer** on ViT-B/16 (top), DINOv2-B/14 (middle) and CLIP-B/16 (bottom), for the images of Fig. 4 and Fig. 9. w is the attention weight of the cross-layer query on that layer's summary token; the remaining weight is on the CLS tokens. Maps are averaged over heads and divided by their own maximum, so a nearly uniform map appears mostly red.

(ii) the probe configuration omits block 0 and normally includes the backbone's final representation; the OpenCLIP exception is detailed in Supp.§F.6.

Two of the three probes impose uniform reductions on complementary axes. ALF uniformly averages the patch tokens within each layer, retains CLS separately, and then learns to fuse across depth (§4.1). SALF-A uniformly averages the token grid across depth (layers) and subsequently applies learned spatial attention (Eq. (4)). SALF-P instead attends spatially within each layer and then attends over the stacked layer summaries (Eq. (3)), making both stages nonlinear, learned, and input-dependent.

Our goal is to understand *when uniform averaging over layers and space is likely to be damaging*, without the expensive hyper-parameter searching process required to train probes, but merely via the statistics of the dataset and features. This should be related to *when SALF-A (layerwise mean) outperforms ALF (spatial mean)*. The simple second-order statistic below asks whether a fixed linear weighting of the collapsed axis could better preserve class-discriminative variation.

F.3 A Signal-to-Noise Test for Uniform Averaging

Let $\bar{h}_{c\ell s}$ be the mean feature of class c (averaged over images, for layer ℓ and spatial position s) and $\bar{h}_{\ell s} = |\mathcal{Y}|^{-1} \sum_c \bar{h}_{c\ell s}$ the equally weighted mean of the class means, so that the between-class term weights classes equally. Define the class deviation and the within-class residual via

$$d_{c\ell s} = \bar{h}_{c\ell s} - \bar{h}_{\ell s} \qquad \text{and} \qquad r_{i\ell s} = h_{i\ell s} - \bar{h}_{y_i \ell s}. \tag{5}$$

To analyse the layer axis, we take the coordinate index j to be a layer, whose associated vector is the concatenation of all $N_S \times d$ token-feature entries at that layer; to analyse the spatial axis, j is a token position and its vector is the concatenation of all $N_L \times d$ layer-feature entries at that position. For an axis with K coordinates, class separation and within-class variation are summarised by two $K \times K$ Gram matrices,

$$(A_a)_{jj'} = \mathbb{E}_c\big\langle d_{cj}, d_{cj'} \big\rangle, \qquad \text{and} \qquad (C_a)_{jj'} = \mathbb{E}_i\big\langle r_{ij}, r_{ij'} \big\rangle, \tag{6}$$

where the inner products run over the axis that is *not* being analysed as well as over the feature dimension. Concretely, the layer matrix is $(A_L)_{\ell\ell'} = |\mathcal{Y}|^{-1} \sum_c \sum_s \langle d_{c\ell s}, d_{c\ell' s} \rangle$, which sums over space, and the spatial matrix is its analogue summed over layers. Intuitively, A_a captures how the inter-class spread aligns over a (i.e., $(A_L)_{ii}$ measures variance in layer i, while $(A_L)_{ij}$ captures covariance between layers i and j); thus, it is a measure of *signal* for classification. In contrast, C_a analogously measures the second-order statistics of intra-class spread, and thus is a measure of "noise" within a category.

For a fixed weighting $w \in \mathbb{R}^K$ of that axis, $w^T A_a w$ and $w^T C_a w$ measure the total inter-class and intra-class variance induced by contraction with w. Thus, the class signal-to-noise ratio of the resulting weighted average using w along a is the Rayleigh quotient

$$\mathrm{SNR}_a(w) = \frac{w^\top A_a \, w}{w^\top C_a \, w}, \tag{7}$$

which is of the same form as the classical linear Fisher discriminant (Fisher, 1936), but over aggregation axis weights, not over feature space. Its maximum $\max_w \mathrm{SNR}_a(w)$ is the largest generalised eigenvalue of $A_a w = \lambda \, C_a w$, and uniform averaging corresponds to $w = \vec{1}$. The *averaging efficiency* of axis a is their ratio:

$$\eta_a = \frac{\mathrm{SNR}_a(\,\vec{1}\,)}{\max_w \mathrm{SNR}_a(w)}. \tag{8}$$

Larger values indicate that the uniform mean is closer to the optimised fixed weighting; smaller values indicate a larger gap under this criterion. In other words, larger η_a suggests uniform averaging over a might not be too damaging (though, this may *not* be the case for a nonlinear aggregation), while smaller η_a indicates that *even among linear contractions*, the uniform average harms SNR.

F.4 Estimation

We estimate our statistics from the training splits of the datasets (no test data is ever used). Notice that η_a is computed in a parameter-free manner (i.e., without learning a model).

To compute the optimal weighting $w^* = \arg\max_w \mathrm{SNR}a(w)$, we solve the generalized eigenvalue problem $A_a w = \lambda C_a w$. Before solving this, we regularize C_a to prevent numerical instability: we add a small ridge term to form $Ca, \tau = C_a + \tau I$, where $\tau = 10^{-3} \cdot \mathrm{tr}(C_a)/K$. We then perform a Cholesky decomposition of the regularized matrix to obtain a standard eigenvalue problem (i.e., whitening it).

The layer and spatial axis eigensystem dimensions are, respectively, 12×12 and 257×257 for DINOv2-S/14 and DINOv2-B/14; 11×11 and 197×197 for OpenCLIP-B/16; 11×11 and 50×50 for OpenCLIP-B/32; and 12×12 and 197×197 for supervised ViT-S/16 and ViT-B/16.

No probe is fit for this diagnostic; the ridge, subsample size, random seed and HAA threshold are fixed globally rather than tuned per dataset.

F.5 Limitations

A *low* η_a indicates that a fixed linear weighting improves on the uniform mean for this second-order criterion; it does not establish that a downstream nonlinear probe will suffer. A *high* η_a indicates little

headroom for fixed linear reweighting, but again places no constraint on what a nonlinear readout could recover from the uncollapsed axis. Further, SNR itself does not have a trivial relationship with accuracy (i.e., improving SNR does not guarantee an increase in classifier performance).

There are two other limitations regarding the statistic and the probes. First, η_S scores a uniform mean over all N_S positions, whereas ALF concatenates CLS with the mean of the patch tokens (§4.1). It therefore does not perfectly model ALF because of the special treatment of CLS. Second, η_a compares the uniform mean only with fixed linear reweightings of the same axis. In addition to nonlinearity, a probe that also changes the reduction on the other axis is separate from this comparison, such as between SALF-A and SALF-P considered in §F.10.

F.6 Datasets and backbones

We analyse six frozen backbones of §5: DINOv2 ViT-S/14 and ViT-B/14, OpenCLIP ViT-B/32 and ViT-B/16, and supervised ViT-S/16 and ViT-B/16. These span three pre-training paradigms with two backbone variants each. The analysis uses 17 classification datasets. Every backbone has η_L, η_S, ALF, SALF-P, and SALF-A results on all 17, yielding $6 \times 17 = 102$ backbone–dataset pairs.

For OpenCLIP, the trained probes use residual-block hooks 1–11 together with the final `visual` representation, which is zero-padded from 512 to 768 dimensions, following Ciernik et al. (2026). For simplicity, our analysis via η uses only the 11 homogeneous, pre-projection residual blocks: since the final projection changes both dimension and channel basis, it has no canonical channelwise alignment with the preceding layers, for the layer-axis Gram. Despite this mismatch, the OpenCLIP association remains strong.

F.7 More Details on the Threshold Groups

Let $\Delta_{L\text{-}S} = \eta_L - \eta_S$. Table 21 shows more complete results regarding how η_a can be used to group dataset-model pairs with high-vs-low accuracy changes. In the HAA rows, family-level mean η_S lies between 0.24 and 0.32, while mean $\Delta_{L\text{-}S}$ lies between $+0.39$ and $+0.55$. In the Non-HAA rows, mean η_S is at least 0.79, mean $\Delta_{L\text{-}S}$ is near -0.36, and the mean gains remain within 1.5 pp of zero.

More generally, define $\mathrm{HAA}_x = \{\eta_S < 0.5 - x \text{ and } \eta_L > 0.5 + x\}$ for $x \in [0, 0.5)$, so HAA is HAA_0. Table 21 also reports $\mathrm{HAA}_{0.1}$ and $\mathrm{HAA}_{0.25}$, together with the η_S-only thresholds $\eta_S < 0.4$ and $\eta_S < 0.25$. The pooled gain is stable as the threshold tightens: $+13.5_{\pm 4.5}$ pp (HAA, n=21), $+13.8_{\pm 3.7}$ pp ($\mathrm{HAA}_{0.1}$, n=17), and $+13.2_{\pm 1.8}$ pp ($\mathrm{HAA}_{0.25}$, n=5). At $x = 0.25$, supervised ViT has no pairs left at all and OpenCLIP retains only one, so those family cells are empty rather than noisy; the pattern holds wherever enough pairs remain to evaluate it. In other words, small changes to the thresholds have limited impact over the simple choice of 0.5.

F.8 Per-Dataset Results

Table 22 reports η_L, η_S, $\Delta_{L\text{-}S}$ and the balanced- and top-1-accuracy gaps for every dataset, averaged within each backbone family. Each accuracy gap is formed within a backbone before averaging across the two backbones in its family.

GTSRB, Dmlab and SVHN are HAA for all six backbones. All three datasets have double-digit across-backbone mean SALF-A gains: $+16.6$, $+10.5$ and $+14.2$ pp, respectively. This pattern is consistent with a benefit from retaining spatial variation, although η_S itself compares only fixed position weightings. STL-10, PASCAL VOC07, CIFAR-10 and Caltech-101 show the opposite ordering, with high η_S and gaps near zero. Country-211 has mean $\eta_S = 0.95$, indicating little benefit from alternative fixed spatial weighting under this criterion; ALF wins by 6.2 pp.

FGVC Aircraft is a notable exception: its mean gain is $+11.8$ pp despite family-averaged $\eta_L = 0.46$ and $\eta_S = 0.42$, and it is HAA for only two backbones. Conversely, Pets has mean $\eta_L = 0.27$ and $\eta_S = 0.54$, while its measured gap is slightly negative (-0.5 pp). These cases likely illustrate the noise/error incurred when a fixed, second-order statistic is compared with trained, input-dependent, nonlinear attention-based probe.

Table 21: **Averaging asymmetry and the gain over ALF.** Family-resolved version of Supp.Table 20, with the two frozen-feature statistics that define the groups and the SALF-P comparison included, together with the more stringent HAA_x and η_S-only thresholds defined in Supp.§F.7. $\mathcal{A}_b(\mathrm{ALF},\mathrm{SA})$ and $\mathcal{A}_b(\mathrm{ALF},\mathrm{SP})$ are balanced-accuracy gains in percentage points. Entries are means over backbone×dataset pairs, subscripted by their standard deviation (n=0 cells are empty for that family); the number of backbones per family is shown in parentheses.

Group	Family	n	η_S	$\Delta_{L\text{-}S}$	$\mathcal{A}_b(\mathrm{ALF},\mathrm{SA})$	$\mathcal{A}_b(\mathrm{ALF},\mathrm{SP})$
HAA	DINOv2 (2)	6	$0.24_{\pm 0.03}$	$+0.54_{\pm 0.06}$	$+13.57_{\pm 4.21}$	$+12.96_{\pm 3.97}$
$\eta_S<0.5$	OpenCLIP (2)	9	$0.32_{\pm 0.09}$	$+0.39_{\pm 0.19}$	$+12.62_{\pm 4.91}$	$+11.73_{\pm 4.55}$
$\eta_L>0.5$	ViT (2)	6	$0.32_{\pm 0.04}$	$+0.55_{\pm 0.07}$	$+14.65_{\pm 4.49}$	$+14.62_{\pm 4.94}$
$\mathrm{HAA}_{0.1}$	DINOv2 (2)	6	$0.24_{\pm 0.03}$	$+0.54_{\pm 0.06}$	$+13.57_{\pm 4.21}$	$+12.96_{\pm 3.97}$
$\eta_S<0.4$	OpenCLIP (2)	5	$0.27_{\pm 0.06}$	$+0.54_{\pm 0.11}$	$+12.99_{\pm 2.17}$	$+12.17_{\pm 2.86}$
$\eta_L>0.6$	ViT (2)	6	$0.32_{\pm 0.04}$	$+0.55_{\pm 0.07}$	$+14.65_{\pm 4.49}$	$+14.62_{\pm 4.94}$
$\mathrm{HAA}_{0.25}$	DINOv2 (2)	4	$0.23_{\pm 0.02}$	$+0.57_{\pm 0.04}$	$+13.45_{\pm 1.96}$	$+13.09_{\pm 2.25}$
$\eta_S<0.25$	OpenCLIP (2)	1	$0.19_{\pm 0.00}$	$+0.69_{\pm 0.00}$	$+12.13_{\pm 0.00}$	$+9.83_{\pm 0.00}$
$\eta_L>0.75$	ViT (2)	0	–	–	–	–
$\eta_S<0.5$	DINOv2 (2)	10	$0.29_{\pm 0.07}$	$+0.33_{\pm 0.28}$	$+8.46_{\pm 7.66}$	$+9.03_{\pm 6.44}$
	OpenCLIP (2)	9	$0.32_{\pm 0.09}$	$+0.39_{\pm 0.19}$	$+12.62_{\pm 4.91}$	$+11.73_{\pm 4.55}$
	ViT (2)	7	$0.34_{\pm 0.06}$	$+0.47_{\pm 0.22}$	$+14.86_{\pm 4.13}$	$+14.52_{\pm 4.52}$
$\eta_S<0.4$	DINOv2 (2)	9	$0.28_{\pm 0.06}$	$+0.38_{\pm 0.24}$	$+9.65_{\pm 7.07}$	$+10.15_{\pm 5.70}$
	OpenCLIP (2)	7	$0.28_{\pm 0.05}$	$+0.45_{\pm 0.17}$	$+13.44_{\pm 2.08}$	$+13.06_{\pm 2.87}$
	ViT (2)	6	$0.32_{\pm 0.04}$	$+0.55_{\pm 0.07}$	$+14.65_{\pm 4.49}$	$+14.62_{\pm 4.94}$
$\eta_S<0.25$	DINOv2 (2)	4	$0.23_{\pm 0.02}$	$+0.57_{\pm 0.04}$	$+13.45_{\pm 1.96}$	$+13.09_{\pm 2.25}$
	OpenCLIP (2)	1	$0.19_{\pm 0.00}$	$+0.69_{\pm 0.00}$	$+12.13_{\pm 0.00}$	$+9.83_{\pm 0.00}$
	ViT (2)	0	–	–	–	–
Non-HAA	DINOv2 (2)	28	$0.79_{\pm 0.21}$	$-0.36_{\pm 0.20}$	$+0.11_{\pm 2.19}$	$+0.78_{\pm 2.29}$
	OpenCLIP (2)	25	$0.84_{\pm 0.14}$	$-0.38_{\pm 0.17}$	$-0.26_{\pm 3.32}$	$+0.35_{\pm 1.70}$
	ViT (2)	28	$0.85_{\pm 0.14}$	$-0.32_{\pm 0.21}$	$+1.23_{\pm 3.96}$	$+1.16_{\pm 2.93}$

Table 22: **Per-dataset averaging efficiency and accuracy gaps, averaged within each backbone family** (number of backbones in parentheses). $\Delta_{L\text{-}S}=\eta_L-\eta_S$. Writing SA for SALF-A, the last two columns are the balanced- and top-1-accuracy gain of SALF-A over ALF, $\mathcal{A}_b=\mathcal{A}_b(\mathrm{ALF},\mathrm{SA})$ and $\mathcal{A}_1=\mathcal{A}_1(\mathrm{ALF},\mathrm{SA})$, in percentage points; positive favours SALF-A. Each accuracy gap is formed within a backbone before averaging across the two backbones in its family.

	DINOv2 (2)					OpenCLIP (2)					ViT (2)				
Dataset	η_L	η_S	$\Delta_{L\text{-}S}$	$\mathcal{A}_b$	$\mathcal{A}_1$	η_L	η_S	$\Delta_{L\text{-}S}$	$\mathcal{A}_b$	$\mathcal{A}_1$	η_L	η_S	$\Delta_{L\text{-}S}$	$\mathcal{A}_b$	$\mathcal{A}_1$
Country-211	0.58	0.97	−0.39	−4.78	−4.78	0.48	0.92	−0.43	−10.15	−10.15	0.58	0.95	−0.38	−3.68	−3.68
FER2013	0.35	0.62	−0.27	+2.16	+2.48	0.64	0.52	+0.11	+0.45	+0.94	0.73	0.71	+0.03	+1.88	+3.39
FGVC Aircraft	0.45	0.35	+0.10	+4.06	+4.05	0.59	0.38	+0.21	+17.08	+17.09	0.35	0.53	−0.18	+14.21	+14.21
GTSRB	0.80	0.25	+0.55	+16.91	+11.80	0.76	0.30	+0.46	+13.71	+8.53	0.95	0.34	+0.61	+19.11	+12.40
STL-10	0.45	0.88	−0.42	−0.39	−0.39	0.45	0.77	−0.32	−0.11	−0.12	0.42	0.89	−0.48	−0.27	−0.26
PASCAL VOC07	0.43	0.97	−0.54	−0.65	−0.24	0.38	0.98	−0.60	−0.21	−0.40	0.47	0.97	−0.50	−0.19	−0.03
Caltech-101	0.39	0.76	−0.37	−0.05	+0.66	0.50	0.77	−0.27	+1.73	+0.06	0.41	0.87	−0.46	+0.50	+0.26
CIFAR-10	0.40	0.93	−0.52	+0.20	+0.20	0.35	0.84	−0.49	+0.91	+0.91	0.43	0.93	−0.50	+0.50	+0.50
CIFAR-100	0.43	0.93	−0.50	+0.69	+0.69	0.45	0.88	−0.43	+2.23	+2.23	0.48	0.93	−0.45	+1.21	+1.21
Diab. Ret.	0.35	0.87	−0.52	+1.22	+2.80	0.38	0.87	−0.50	+0.98	+2.79	0.75	0.83	−0.08	+0.10	+0.32
Dmlab	0.75	0.26	+0.49	+10.83	+11.02	0.88	0.23	+0.65	+11.02	+11.42	0.87	0.30	+0.57	+9.68	+9.75
DTD	0.50	0.97	−0.47	−0.79	−0.79	0.57	0.98	−0.41	−3.24	−3.24	0.79	0.97	−0.19	+0.88	+0.88
EuroSAT	0.54	0.77	−0.24	+0.54	+0.55	0.63	0.87	−0.24	+0.46	+0.43	0.78	0.81	−0.03	+0.39	+0.38
Flowers	0.40	0.62	−0.22	+0.10	−0.16	0.50	0.77	−0.27	+2.24	+2.00	0.39	0.84	−0.45	+0.12	+0.12
Pets	0.32	0.39	−0.07	−2.48	−2.45	0.27	0.56	−0.28	+1.08	+1.11	0.21	0.66	−0.45	−0.24	−0.23
RESISC45	0.40	0.99	−0.58	+1.77	+1.80	0.42	0.99	−0.57	+0.96	+1.00	0.66	0.98	−0.33	+1.84	+1.86
SVHN	0.81	0.21	+0.59	+12.98	+13.12	0.66	0.29	+0.37	+14.36	+14.52	0.80	0.33	+0.47	+15.18	+14.94
Mean	0.49	0.69	−0.20	+2.49	+2.37	0.52	0.70	−0.18	+3.15	+2.89	0.59	0.76	−0.16	+3.60	+3.29

F.9 Correlation with The Probe-Accuracy Gaps

Table 23 gives Pearson and Spearman correlations of η_S and $\Delta_{L\text{-}S}$ with all three pairwise accuracy gaps, pooled over datasets and backbone sizes within each family, for balanced and top-1 accuracy

Table 23: **Correlation between the frozen-feature statistics and the probe-accuracy gaps**, pooled over datasets and backbone sizes within each family (number of backbones in parentheses). SA denotes SALF-A and SP denotes SALF-P; $\mathcal{A}(X, Y)$ is the accuracy gain of Y over X. Each cell is Pearson r / Spearman ρ and n is the number of backbone×dataset pairs. η_S and $\Delta_{L\text{-}S}$ correlate with both gaps that involve ALF, whereas their correlations with $\mathcal{A}(\mathrm{SP}, \mathrm{SA})$ are weak (see Supp.§F.10).

		vs. η_S			vs. $\Delta_{L\text{-}S}$		
Family	n	$\mathcal{A}(\mathrm{SP}, \mathrm{SA})$	$\mathcal{A}(\mathrm{SP}, \mathrm{ALF})$	$\mathcal{A}(\mathrm{ALF}, \mathrm{SA})$	$\mathcal{A}(\mathrm{SP}, \mathrm{SA})$	$\mathcal{A}(\mathrm{SP}, \mathrm{ALF})$	$\mathcal{A}(\mathrm{ALF}, \mathrm{SA})$
Balanced accuracy ($\mathcal{A}_b$)							
DINOv2 (2)	34	−0.01/−0.06	+0.79/+0.53	−0.73/−0.54	+0.05/−0.06	−0.87/−0.51	+0.82/+0.42
OpenCLIP (2)	34	−0.29/−0.16	+0.86/+0.75	−0.82/−0.68	+0.24/+0.04	−0.84/−0.70	+0.78/+0.61
ViT (2)	34	−0.19/+0.07	+0.87/+0.53	−0.86/−0.50	−0.01/−0.17	−0.82/−0.65	+0.77/+0.64
Top-1 accuracy ($\mathcal{A}_1$)							
DINOv2 (2)	34	−0.16/−0.09	+0.59/+0.44	−0.72/−0.56	+0.23/+0.13	−0.64/−0.27	+0.81/+0.44
OpenCLIP (2)	34	−0.30/−0.24	+0.83/+0.68	−0.79/−0.65	+0.26/+0.15	−0.79/−0.61	+0.75/+0.56
ViT (2)	34	−0.16/+0.06	+0.86/+0.53	−0.86/−0.50	−0.05/−0.22	−0.83/−0.70	+0.75/+0.63

separately. For brevity, we abbreviate SALF-A as SA and SALF-P as SP. For balanced accuracy, η_S is negatively correlated with $\mathcal{A}_b(\mathrm{ALF}, \mathrm{SA})$ ($r = -0.73$ to -0.86): lower spatial averaging efficiency coincides with larger SALF-A gains. Second, it is positively correlated with $\mathcal{A}_b(\mathrm{SP}, \mathrm{ALF})$ ($r = +0.79$ to $+0.87$), the corresponding gap involving ALF and SALF-P. Third, its correlation with $\mathcal{A}_b(\mathrm{SP}, \mathrm{SA})$ is weak ($|r| \leq 0.29$) and changes sign across families in rank correlation. The analogous pooled correlation with η_L is also weak ($r = +0.03$), as discussed in Supp.§F.10.

F.10 Why η does not separate SALF-A from SALF-P

The primary distinction relevant to η_L is that SALF-A applies a fixed uniform mean over depth, whereas SALF-P performs spatial attention within each layer followed by learned attention across layer summaries. It is therefore natural to ask whether η_L is associated with $\mathcal{A}(\mathrm{SP}, \mathrm{SA})$. For balanced accuracy, the correlation is weak: $+0.11$, $+0.12$ and -0.22 in the three families, and $+0.03$ when pooled.

In terms of performance, the observed balanced-accuracy difference is rather small: SALF-A minus SALF-P is -0.2 ± 1.8 pp over all 102 pairs and $+0.6 \pm 1.6$ pp within HAA. More importantly, this comparison does not hold the spatial treatment fixed. η_L evaluates fixed linear reweightings of depth under a common treatment of space, while SALF-P changes both the aggregation order and the spatial readout relative to SALF-A. Hence, η_L may not be able to characterize their small gap in performance. One conjecture for their similarity in performance is that averaging may actually be helpful in avoiding overfitting and that, given access to spatial processing, the sophistication of layerwise processing becomes less impactful.

By contrast, ALF and SALF-P differ most directly in whether patches are uniformly summarised within each layer. η_S targets their primary distinction, although the separate treatment of CLS means that the correspondence is not exact. The stronger association with their accuracy gap is consistent with this interpretation. Further, SALF-A differs by a uniform mean *along the layer axis*, so a natural measure is to combine η_L with η_S, when comparing to ALF (which mean pools over space)—we do this via the HAA grouping.

G Spatial Stationarity: Static vs. Dynamic Selectivity

This appendix defines ξ from §5.3, explains the estimation, and gives complete per-backbone results. We begin with an overview that summarizes the main results.

G.1 Overview

How Do Attention Maps Explain SALF's Performance? We analyze SALF spatial attention maps (SAMs) to understand when it improves on ALF. Although non-uniform attention might be expected

to favour SALF-P, Kullback–Leibler divergence from uniformity, $\bar{\Delta}$, is not significantly correlated with the BA gap (Pearson $r = 0.13$, $p > 0.05$). Instead, we decompose non-uniformity $\bar{\Delta}$ into dataset-level (static) and image-dependent (dynamic) components. We define the spatial stationarity $\xi \in [-1, 1]$, indicating static-dominant (-1) to dynamic-dominant (+1) SAMs, which strongly predicts SALF-P's advantage over ALF ($r = -0.86$, $p < 10^{-4}$): static-dominant datasets yield larger gains. Splitting tasks at $\xi = 0$, SALF-P improves BA by 12.6 ± 3.7 pp for $\xi < 0$, versus 0.9 ± 2.5 pp for $\xi \geq 0$. We hypothesize that SALF-P can exploit *reusable* non-uniform SAMs, which ALF struggles to represent, whereas highly image-dependent attention is harder to learn. See Supp.§G for details.

G.2 Setting and notation

SALF-P produces, at inference time, per-layer spatial attention maps and cross-layer weights. For image j and backbone layer ℓ:

- $Z_{j\ell} \in \Delta^{S-1}$ is SALF-P's within-layer cross-attention map: a probability distribution over the S patch positions, obtained by attending a learned spatial-pooling query to the patch tokens of layer ℓ.
- $\alpha_{j\ell} \geq 0$, $\sum_\ell \alpha_{j\ell} = 1$, are the normalized cross-layer weights on the spatial summary of each layer for image j.

The *spatial marginal* for image j is the α-weighted mixture:

$$P_j(s) \;=\; \sum_\ell \alpha_{j\ell}\, Z_{j\ell}(s), \quad P_j \in \Delta^{S-1}.$$

For this analysis, we omit the CLS-token branch and renormalize the retained attention so that P_j sums to one over spatial positions. We define the dataset-mean map $\bar{P}(s) = \frac{1}{N}\sum_j P_j(s)$ and the uniform distribution $U(s) = 1/S$, which is the implicit spatial "attention" used by ALF's patch-token average pooling.

G.3 Static/dynamic decomposition and spatial stationarity ξ

Decomposition. Total spatial selectivity is the mean KL divergence from uniform, $\overline{\Delta} = \frac{1}{N}\sum_j \mathrm{KL}(P_j \,\|\, U)$. By the chain rule of KL divergence this decomposes exactly into:

$$\underbrace{\frac{1}{N}\sum_j \mathrm{KL}(P_j \| U)}_{\overline{\Delta}} \;=\; \underbrace{\mathrm{KL}(\bar{P} \| U)}_{\Delta_s \geq 0} \;+\; \underbrace{\left[H(\bar{P}) - \tfrac{1}{N}\textstyle\sum_j H(P_j)\right]}_{\Delta_d = I(\text{image; position}) \geq 0}.$$

$\Delta_s = \mathrm{KL}(\bar{P} \,\|\, U)$ is the divergence of the dataset-averaged attention map from uniform—the shared spatial selectivity attributable to a single, fixed "look-here" rule. $\Delta_d = I(\text{image; position})$ is the empirical mutual information between image identity and attended location—the image-dependent component of selectivity beyond the shared prior. Both are non-negative by construction.

Definition of the Spatial Stationarity ξ. Using the decomposition of the non-uniformity, we can define the following simple information-theoretic measure:

$$\xi \;=\; \frac{\Delta_d - \Delta_s}{\Delta_d + \Delta_s + \varepsilon} \;\in\; [-1, +1], \qquad \varepsilon = 10^{-4} \text{ nats}.$$

We refer to ξ as the *spatial stationarity* of the dataset's attention maps. Its sign is exactly the sign of $\Delta_d - \Delta_s$; its magnitude describes the imbalance between the two terms relative to total selectivity and ε.

Interpreting ξ.

$\xi < 0$ **(static-dominant).** $\Delta_s > \Delta_d$, so the dataset-shared component accounts for the larger share of spatial selectivity. Values near -1 additionally require $\Delta_s \gg \Delta_d$ and $\Delta_s \gg \varepsilon$. Examples: GTSRB (centred traffic signs), SVHN (centred house numbers).

$\xi > 0$ **(dynamic-dominant).** $\Delta_d > \Delta_s$, so per-image variation accounts for the larger share of spatial selectivity. Values near $+1$ additionally require $\Delta_d \gg \Delta_s$ and $\Delta_d \gg \varepsilon$.

$\xi \approx 0$ **(balanced or negligible-total).** This means $|\Delta_d - \Delta_s|$ is small relative to $\Delta_d + \Delta_s + \varepsilon$. For non-negligible total selectivity it indicates $\Delta_s \approx \Delta_d$; vanishing total selectivity also drives ξ toward zero only when $\overline{\Delta} \ll \varepsilon$. Thus $\overline{\Delta}$ reports magnitude, whereas ξ reports signed balance.

G.4 Estimation and coverage

Maps are computed on a class-stratified subsample of 500 test images per backbone×dataset pair. Once the trained probe maps are available, computing ξ requires no additional fitting. We use the base-size variant of each family (ViT-B/16, OpenCLIP-B/16, DINOv2-B/14), giving $3 \times 17 = 51$ model-dataset pairs.

G.5 Correlation with the accuracy gap

Correlation measures. *Spearman* ρ uses within-backbone ranking: each backbone's observations are ranked separately, then pooled. *Pearson* r uses raw values pooled across all backbones. Two-sided permutation p-values for Spearman correlations use 2×10^4 permutations: for pooled inference we permute the 17 dataset assignments jointly across all three backbones, giving $p < 10^{-4}$ for ξ and $p = 0.226$ for total selectivity $\overline{\Delta}$; per-backbone tests permute datasets within that backbone. The *leave-one-dataset-out (LOO) range* is the interval $[\rho_{\min}, \rho_{\max}]$ obtained by repeating the correlation with each dataset removed in turn: the pooled row omits all three corresponding backbone×dataset pairs, whereas each per-backbone row omits its single corresponding pair.

Table 24: Correlation of ξ with the full-test-set SALF-P over ALF balanced-accuracy gap.

	n	Spearman ρ	perm p	LOO range	Pearson r	Pearson p
Pooled (all 3 backbones)	51	-0.78	$< 10^{-4}$	$[-0.82, -0.74]$	-0.86	$< 10^{-4}$
ViT-B/16	17	-0.75	0.001	$[-0.84, -0.69]$	-0.79	< 0.001
OpenCLIP-B/16	17	-0.86	$< 10^{-4}$	$[-0.92, -0.83]$	-0.91	$< 10^{-4}$
DINOv2-B/14	17	-0.73	0.002	$[-0.83, -0.68]$	-0.91	$< 10^{-4}$

The correlation is significant ($p < 0.01$) in every backbone individually and is not driven by any single dataset.

Contrast: raw spatial non-uniformity. Table 25 reports the correlation of the total selectivity $\overline{\Delta} = \Delta_s + \Delta_d$ (raw non-uniformity, no static/dynamic split) with the same gap. Neither the pooled nor the per-backbone correlations reach significance ($p < 0.05$); thus, we detect no significant relationship.

Table 25: Correlation of raw spatial non-uniformity $\overline{\Delta}$ with the full-test-set SALF-P over ALF gap. Same methodology as Table 24.

	n	Spearman ρ	perm p	LOO range	Pearson r	Pearson p
Pooled (all 3 backbones)	51	$+0.24$	0.226	$[+0.17, +0.41]$	$+0.13$	0.36
ViT-B/16	17	$+0.30$	0.244	$[+0.24, +0.56]$	$+0.19$	0.46
OpenCLIP-B/16	17	$+0.33$	0.197	$[+0.20, +0.49]$	$+0.22$	0.40
DINOv2-B/14	17	$+0.09$	0.729	$[-0.02, +0.20]$	$+0.14$	0.59

G.6 Group-threshold view

Table 26 splits all 51 pairs at $\xi = 0$.

Table 26: Mean $\pm$ std full-test-set SALF-P over ALF balanced-accuracy gain (pp), split by sign of ξ. All 51 pairs across ViT-B/16, OpenCLIP-B/16, and DINOv2-B/14.

Group	n	Mean gain (pp)
Static-dominant ($\xi < 0$)	9	$+12.6 \pm 3.7$
Dynamic-dominant or balanced ($\xi \geq 0$)	42	$+0.9 \pm 2.5$

The nine static-dominant pairs are GTSRB, FGVC Aircraft, and SVHN, each paired with all three backbones. Pooling all 51 pairs, the ratio of the two mean gains is $13.9\times$ (approximately $14\times$); the static-dominant mean is also larger for each backbone, with backbone-specific ratios ranging from $12.7\times$ to $16.4\times$.

G.7 Remarks on Analysis Results

The largest accuracy gains occur for tasks whose learned SALF-P attention maps are static-dominant, meaning that their spatial selectivity contains a substantial dataset-shared component. Because ALF uniformly averages patch tokens, it does not explicitly preserve patch-level spatial structure. This architectural difference may help explain its lower performance on these tasks. In contrast, SALF-P attends over spatial tokens, allowing it to represent dataset-shared spatial preferences directly. However, the present analysis is correlational and does not establish the exact causal mechanism of SALF's gains.

Why SALF-P offers less advantage over ALF on tasks with $\xi \geq 0$ remains uncertain. In particular, our analysis does not isolate the contribution of image-conditioned weighting from that of a fixed, learned, non-uniform spatial weighting. Such a fixed weighting could potentially capture dataset-shared spatial preferences but, by construction, could not adapt its weights to individual images. We conjecture that further gains on these tasks (with substantial image-dependent selectivity) may require more expressive probes, capable of stronger image-conditioned spatial selection.