Selective, Regularized, and Calibrated: Harnessing Vision Foundation Models for Cross-Domain Few-Shot Semantic Segmentation

Supplementary Material

1. CD-FSS Benchmark Datasets

We conduct experiments on four target-domain datasets: FSS-1000 [7], DeepGlobe [3], ISIC2018 [2, 11], and Chest X-ray [1, 6]. Together, these datasets cover natural images, satellite imagery, dermoscopic images, and radiological scans, and thus exhibit substantial variation in appearance, scale, and imaging modality.

FSS-1000 [7] is a few-shot semantic segmentation dataset containing 1000 object categories, each associated with 10 pixel-wise annotated images. Following the standard CD-FSS protocol, we adopt the official semantic segmentation split. We report results on the designated test subset, which includes 240 categories and 2400 images and serves as one of our target domains.

DeepGlobe [3] is a satellite land-cover dataset with dense per-pixel annotations for seven semantic classes: urban, agriculture, rangeland, forest, water, barren, and unknown. Because labels are only available for the training partition, we use the official training set, comprising 803 annotated images, as the target-domain data for evaluation.

ISIC2018 [2, 11] is a dermoscopic skin-lesion segmentation benchmark designed for melanoma screening. The dataset contains high-resolution dermoscopy images, each with an expert-annotated binary mask for the primary lesion region. We follow the preprocessing and data splits commonly adopted in previous work and treat ISIC2018 as a medical target domain characterised by low contrast and irregular object boundaries.

Chest X-ray [1, 6] is a tuberculosis screening dataset composed of 566 high-resolution posterior–anterior chest radiographs (approximately 4020×4892 pixels). The images were collected from 58 tuberculosis cases and 80 normal cases. To make training feasible and ensure consistency, we resize each radiograph to 1024×1024 pixels and use this dataset as a grayscale medical target domain.

2. Details of SSP Baseline

The SSP [4] baseline is a prototype-based few-shot semantic segmentation approach that refines the support prototype by exploiting the query image. Let $\hat{F}_q \in \mathbb{R}^{c\times h\times w}$ denote the enhanced query feature map and P_s denote the foreground support prototype. SSP first produces coarse foreground and background probability maps by matching $\hat{F}_q$ to P_s using cosine similarity:

$$M_{\text{coarse}} = \text{softmax}\big(\cos(\hat{F}_q, P_s)\big), \tag{1}$$

where $\cos(\cdot,\cdot)$ computes channel-wise cosine similarity between P_s and each spatial feature of $\hat{F}q$, and the softmax is applied over the foreground/background dimension, yielding $M\text{coarse} = M^b_{\text{coarse}}, M^f_{\text{coarse}} \in \mathbb{R}^{2\times h\times w}$.

To obtain a query-adapted foreground prototype, SSP aggregates high-confidence foreground locations from $\hat{F}_q$ via masked average pooling:

$$P^f_q = \text{MAP}\big(\hat{F}_q, M^f_{\text{coarse}}(x,y) > \tau_f\big), \tag{2}$$

where $\text{MAP}(\cdot)$ denotes masked average pooling over spatial positions (x, y) whose foreground probability exceeds $\tau_f = 0.7$ and the resulting prototype satisfies $P^f_q \in \mathbb{R}^c$.

To model complex background regions, SSP constructs adaptive self-support background prototypes. We first compute a masked background feature map as:

$$\hat{F}^b_q = \hat{F}_q \odot \mathbb{I}\big(M^b_{\text{coarse}}(x,y) > \tau_b\big), \tag{3}$$

where $\odot$ denotes element-wise multiplication, $\mathbb{I}(\cdot)$ is the indicator function, and $\tau_b = 0.6$ is the confidence threshold for background pixels. As a result, $\hat{F}^b_q$ only keeps features at reliable background positions.

Next, we reshape $\hat{F}^b_q$ and $\hat{F}_q$ into matrices in $\mathbb{R}^{c\times t}$ and $\mathbb{R}^{c\times n}$, respectively, where t is the number of selected background locations and $n = h \times w$ is the total number of spatial positions. The affinity between background features and all query features is computed as:

$$A = \text{matmul}\big(\hat{F}^{b\top}_q, \hat{F}_q\big), \tag{4}$$

where $\text{matmul}(\cdot,\cdot)$ denotes matrix multiplication, producing a similarity matrix $A \in \mathbb{R}^{t\times n}$ whose $(i,j)^{\text{th}}$ entry measures the similarity between the ith background feature and the jth query location.

Adaptive background prototypes are derived by attention-style aggregation:

$$P^b_q = \text{matmul}\big(\hat{F}^b_q, \text{softmax}(A)\big), \tag{5}$$

where the softmax is applied along the first dimension of A so that each query location attends to a normalised weight distribution over the t background features. The matrix multiplication then computes the corresponding weighted

sums, which yield query-dependent background prototypes $P_q^b \in \mathbb{R}^{c\times n}$. And then they are further pooled into a compact representation.

Finally, SSP combines the original support prototype with the self-support prototypes to form the fused prototype set:

$$P = \alpha_1 P_s + \alpha_2 P_q, \qquad P_q = \{P_q^f, P_q^b\}, \tag{6}$$

where $\alpha_1 = \alpha_2 = 0.5$ are fixed fusion weights, P_s is the support-derived prototype, and P_q collects the query-derived foreground and background prototypes.

The refined segmentation for the query image is then obtained by matching $\hat{F}_q$ with the fused prototype set:

$$\bar{M}_q = \mathrm{softmax}\big(\cos(\hat{F}_q, P)\big), \tag{7}$$

where $\bar{M}_q \in \mathbb{R}^{2\times h\times w}$ denotes the final foreground/background probability maps obtained by applying a channel-wise softmax to the cosine-similarity scores.

3. Ablation on MLP Components

MLP head variants on the selected layer. With the backbone frozen and the working representation layer chosen by HLS, we attach a small MLP head at that layer to improve support-query correspondence at test time. We evaluate three variants: M0, no MLP head (apply_fc=False); M1, MLP branch present but frozen (apply_fc=True, zero_init=True; parameters fixed); and M2, a trainable MLP fine-tuned at test time on the selected layer (apply_fc=True). Only the MLP head is updated, keeping the fraction of updated parameters below 2.7%.

Table 1. MLP ablation at the layer selected by HLS with the backbone frozen. Δ denotes the improvement relative to the row above.

Variant	mIoU@1	Δ	mIoU@5	Δ
M0: no MLP	65.66	–	75.20	–
M1: MLP frozen	66.33	+0.67	75.78	+0.58
M2: trainable MLP	**68.29**	**+1.96**	**77.91**	**+2.13**

Analysis of MLP ablation. Starting from M0 at 65.66 mIoU in one shot and 75.20 mIoU in 5-shot, as shown in Table 1, adding a frozen residual MLP branch (M1) raises the means to 66.33 and 75.78 mIoU, with gains of 0.67 and 0.58 over M0. This suggests that even a fixed projection stabilises channel scales and token mixing at the selected layer. Allowing this compact head to adapt at test time (M2) further increases accuracy to 68.29 and 77.91 mIoU, adding 1.96 and 2.13 over M1. Cumulatively, M2 improves over M0 by 2.63 in one shot and 2.71 in 5-shot, which correspond to relative gains of about 4.0% and 3.6%, while keeping the fraction of updated parameters under 2.7%. These gains are consistent with the Select Regularise Calibrate design. HLS provides a stable representation. The small MLP recenters and rescales features to reduce support to query mismatch, and the resulting representations interact more reliably with PGR and PAC. In practice, a single compact trainable MLP on the selected layer delivers most of the benefit with minimal overhead.

4. Investigating the Effect of the Last Layer

After HLS selects the best single-layer ℓ_{single} for each episode, we form a compact neighbourhood U anchored at ℓ_{single} and we include the last ViT layer L_{23} to mitigate fragmented shapes. We evaluate all candidates under the same episodic objective. For any U, let r_ℓ denote the single-layer ETR of layer ℓ. We compute the fusion weights and the fused representation as follows:

$$w_\ell = \frac{\exp\big(-\beta\, r_\ell - \mathrm{dist}(\ell, L_{23})/\tau\big)}{\sum_{j\in U}\exp\big(-\beta\, r_j - \mathrm{dist}(j, L_{23})/\tau\big)}$$
$$F^U = \sum_{\ell\in U} w_\ell\, F^\ell \tag{8}$$

where $\beta > 0$ controls reliance on the data evidence r_ℓ, and $\tau > 0$ is a locality bandwidth that biases the fusion toward deeper semantically aggregated layers. As $\tau \to \infty$, the locality term vanishes and the solution reduces to single layer routing, that is $\arg\min_{\ell\in U} r_\ell$. When evidence spreads across adjacent layers, a moderate τ balances data evidence and semantic aggregation and stabilises routing.

Table 2. Local fusion anchored at the routed layer. We report average mIoU for the one-shot and five-shot settings, along with the changes relative to using L_{23} alone and to excluding L_{23}.

Variant	mIoU avg.		Δ vs. L_{23}	Δ vs. no L_{23}
	1 shot	5 shot	5 shot	5 shot
F^0+L_{23}, τ=0.0	66.58	75.49	0.00	0.00
F^0+L_{23}, τ=2.0	68.29	77.85	2.36	2.36
F^0+no L_{23}, pivot=last, τ=0.0	66.45	75.29	−0.20	0.00
F^0+no L_{23}, pivot=$\ell^\star$, τ=2.0	66.83	76.34	0.85	1.05

Analysis of L23. Table 2 compares single-layer routing with local fusion. Local fusion anchored at L_{23} with τ=2.0 outperforms using L_{23} alone on both one-shot and five-shot averages. Excluding L_{23} from the candidate set reduces performance. Redirecting fusion to the routed layer $\ell^\star$ recovers part of the performance drop, yet it remains inferior to configurations that include L_{23}. By dataset, Table 3 reports larger gains on DeepGlobe [3] and ISIC2018 [2, 11], consistent with evidence drift across episodes and the need for deeper semantic aggregation. So we adopt local fusion with τ=2.0 and retain L_{23} in the candidate pool by default.

Table 3. By-dataset mIoU comparing L_{23} alone and local fusion. Including L_{23} in the candidate pool and setting τ=2.0 yields the highest averages, with the largest gains on DeepGlobe and ISIC.

Backbone (DINOv3)	DeepGlobe		ISIC		Chest X-ray		FSS-1000		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
$F^0 + L_{23}$, τ=2.0	**44.59**	**63.43**	**61.17**	**73.64**	85.80	87.88	**81.59**	**86.69**	**68.29**	**77.91**
$F^0 + L_{23}$, τ=0.0	42.90	61.49	55.17	66.53	87.06	88.29	81.20	85.63	66.58	75.49
F^0 +no L_{23}, pivot=last, τ=0.0	42.87	61.43	54.84	66.00	87.01	88.26	81.09	85.47	66.45	75.29
F^0 +no L_{23}, pivot=$\ell^\star$, τ=2.0	42.32	63.11	56.41	68.25	**87.44**	**88.41**	81.16	85.58	66.83	76.34

5. Details of Pixelwise Adaptive Calibration

Despite HLS and PGR, residual errors persist along thin boundaries, slender structures, and low contrast regions. With the backbone frozen, PAC adds three lightweight residual branches in the logit domain, coupled to the routed layer ℓ^* and to the patch attention calibrated by PGR.

Feature similarity for semantic alignment. Let $\mathbf{F}_q(x)$ denote the query feature at ℓ^*. Foreground and background prototypes, $\mathbf{P}_{\mathrm{fg}}$ and $\mathbf{P}_{\mathrm{bg}}$, are computed by masked averaging over support features at ℓ^*. We define the prototype difference logit as

$$\ell_{\mathrm{sim}}(x) = \tau_{\mathrm{sim}}\big[\cos\big(\mathbf{F}_q(x), \mathbf{P}_{\mathrm{fg}}\big) - \cos\big(\mathbf{F}_q(x), \mathbf{P}_{\mathrm{bg}}\big)\big], \tag{9}$$

where τ_{sim} is a small temperature. This branch recovers missed regions and sharpens local focus.

One hop attention for spatial consistency. Let $\widetilde{A}$ denote the row-normalised patch to patch attention at ℓ^* after PGR. Given the base foreground probability $p_0(x) = \sigma(\ell_0(x))$, we propagate once on the patch grid as:

$$\ell_{\mathrm{attn}}(x) = \tau_{\mathrm{attn}}\big[(\widetilde{A}\, p_0)_x\big], \tag{10}$$

This elongates responses along the object extent and suppresses spurious long-range peaks, with limited impact on the global distribution.

Image vector for appearance correction. Let $\mathbf{v}(x)$ denote a shallow appearance embedding for color and texture:

$$\ell_{\mathrm{img}}(x) = \tau_{\mathrm{img}}\big[\cos\big(\mathbf{v}(x), \mathbf{u}_{\mathrm{fg}}\big) - \cos\big(\mathbf{v}(x), \mathbf{u}_{\mathrm{bg}}\big)\big], \tag{11}$$

Here $\mathbf{u}_{\mathrm{fg}}$ and $\mathbf{u}_{\mathrm{bg}}$ are image level prototypes, and τ_{img} is a small temperature. This branch provides light global denoising and prevents over-shrinking.

The final logit is a linear combination in the logit domain:

$$\ell_{\mathrm{final}}(x) = \ell_0(x) + w_{\mathrm{sim}}\,\ell_{\mathrm{sim}}(x) + w_{\mathrm{attn}}\,\ell_{\mathrm{attn}}(x) + w_{\mathrm{img}}\,\ell_{\mathrm{img}}(x), \tag{12}$$

where $\ell_0(x)$ is the base logit from the selected representation and $w_{\cdot}$ are fixed scalar weights. A single-step refine vote gate applies residuals only when the estimated gain is positive, adding negligible overhead. Together, the three stages realise a Select-Regularise-Calibrate pipeline, adapting at test time with a frozen backbone.

6. Adaptive Behaviour of PAC

After HLS and PGR, residual errors concentrate along thin boundaries and in low-contrast regions. Pixelwise Adaptive Calibration (PAC) adds three lightweight residual branches in the logit domain, namely feature similarity, one-hop attention propagation, and image appearance, while the backbone remains frozen.

To avoid negative transfer, we enable PAC only when leave-one-out voting on the supports predicts a positive gain. Concretely, we treat each support as a pseudo query, compute the ΔmIoU with and without PAC, and enable PAC on the true query if at least T votes are positive. In the one-shot case, we synthesise two augmented views of the support to obtain two votes.

Table 4. Effect of PAC gating thresholds. We report average mIoU (%) and the trigger rate of the automatic gate. The best policy is to keep the gate always on for one shot, and to use automatic gating with a threshold $2/5$ for five shots.

Policy	1 shot	5 shot	Trigger rate (auto)
refine = off	67.54	76.67	–
auto, T=1	68.02	–	56.32
auto, T=2	–	**77.91**	**74.57**
auto, T=3	–	77.22	59.44
always on	**68.29**	77.80	–

Analysis of gate. Relative to HLS at 76.7 mIoU, PGR raises the mean to 77.3 (+0.6), PAC to 77.2 (+0.5), and using PGR together with PAC yields 77.9 (+1.2), confirming their complementarity. For PAC gating, Tab. 4 shows that in the one-shot setting, the best policy is to keep PAC on for all episodes (68.29 mIoU). In the five-shot setting, the automatic gate with threshold T=2 out of 5 achieves the highest mean mIoU (77.91) with a moderate trigger rate (74.6%), whereas T=3 out of 5 further reduces the trig-

Table 5. By-dataset mIoU and gate trigger rates. The recommended setting (one shot always on, five-shot automatic gating with threshold $2/5$) yields the highest average mIoU.

	DeepGlobe		ISIC		Chest X-ray		FSS-1000		Average		Avg. trigger rate (%)	
Setting	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
1-shot auto, 5-shot always	44.35	63.51	60.28	73.72	86.27	87.22	81.19	86.73	68.02	77.80	–	–
Trigger rate (%)	**55.83**	–	**50.00**	–	**19.50**	–	**99.95**	–	–		**56.32**	–
1 shot always, 5 shot auto 2/5	44.59	63.43	61.17	73.64	85.80	87.88	81.59	86.69	**68.29**	**77.91**	–	–
Trigger rate (%)	–	**25.67**	–	**97.83**	–	**25.67**	–	**84.46**	–		–	**74.57**
1 shot always, 5 shot auto 3/5	44.59	63.41	61.17	73.40	85.80	87.95	81.59	86.63	**68.29**	77.85	–	–
Trigger rate (%)	–	5.83	–	69.00	–	5.83	–	74.92	–		–	59.44

ger rate but lowers accuracy to 77.22–77.85 mIoU. The per-dataset study in Tab. 5 supports the same recommendation: one shot with PAC on for all episodes and five-shot with automatic gating at $T{=}2$ out of 5.

Decomposing PAC on top of HLS plus PGR at 77.27 mIoU, the similarity residual ℓ_{sim}, the one-hop attention propagation ℓ_{attn}, and the image appearance cue ℓ_{img} contribute +0.30, +0.22, and +0.18 mIoU, respectively. Using all three reaches 77.91 mIoU, a further +0.64. Together, HLS stabilises the routed layer, PGR sharpens locality, and PAC corrects pixel-level logits, yielding a cumulative gain under a frozen backbone.

7. Episode-wise Layer Selectors

7.1. Episode notation and setting

Let $\ell \in \mathcal{C}$ index a ViT layer, and let $\mathbf{F}_q^\ell(x) \in \mathbb{R}^{d_\ell}$ denote the query feature at pixel x from layer ℓ. Support features are pooled using masks to form foreground and background prototypes $\mathbf{P}_{\text{fg}}^\ell$ and $\mathbf{P}_{\text{bg}}^\ell$. Given a baseline foreground probability $p_0(x) \in [0,1]$ for the query, we build soft masked query prototypes as:

$$\mathbf{Q}_{\text{fg}}^\ell = \frac{\sum_x p_0(x)\,\mathbf{F}_q^\ell(x)}{\sum_x p_0(x)}, \qquad \mathbf{Q}_{\text{bg}}^\ell = \frac{\sum_x (1-p_0(x))\,\mathbf{F}_q^\ell(x)}{\sum_x (1-p_0(x))}. \tag{13}$$

Unless noted otherwise, all scalar layer scores are range normalised *within each episode* across $\mathcal{C}$, so different selectors are comparable:

$$\tilde{s}_\ell = \frac{s_\ell - \min_{j\in\mathcal{C}} s_j}{\max_{j\in\mathcal{C}} s_j - \min_{j\in\mathcal{C}} s_j + \text{Var}\,epsilon}, \qquad \text{Var}\,epsilon = 10^{-8}. \tag{14}$$

7.2. Selectors other than HLS

We group the non-episodic selectors into two families: a heuristic static rule built from prototype and mask scores, and gradient-based proxies. Unless noted, all scalar layer scores are range-normalised across the candidate set $\mathcal{C}$ within each episode. Prototypes and the baseline mask p_0 follow the definitions in Sec. 7.1.

Table 6. Notation for layer selection in the episodic setting. All scalar layer scores are range-normalised across the candidate set $\mathcal{C}$ unless noted.

Symbol	Description
$\ell \in \mathcal{C}$	Candidate ViT layer index
$\mathbf{F}_q^\ell(x) \in \mathbb{R}^{d_\ell}$	Query feature at pixel x from layer ℓ
$\mathbf{P}_{\text{fg}}^\ell,\ \mathbf{P}_{\text{bg}}^\ell$	Supp foreground/background prototypes at layer ℓ
$\mathbf{Q}_{\text{fg}}^\ell,\ \mathbf{Q}_{\text{bg}}^\ell$	Soft masked query prototypes (see Eq. (13))
$p_0(x) \in [0,1]$	Baseline foreground probability on the query
$\text{mIoU}_{\text{sup}}(\ell)$	Support-only pseudo-query mIoU at layer ℓ

Static heuristic selector (Static-Max). This rule blends three normalised scores, namely semantic agreement, structure separation, and a complexity term combining texture and uncertainty, and selects the layer with the largest weighted sum:

$$\ell^\star_{\text{static}} = \arg\max_{\ell\in\mathcal{C}} \big[\,\alpha' S_{\text{sem}}(\ell) + \beta' S_{\text{str}}(\ell) + \gamma' C(\ell)\,\big], \qquad \alpha',\beta',\gamma' \geq 0,\ \alpha'+\beta'+\gamma' = 1, \tag{15}$$

where weights are task-specific, and the objective is a surrogate not directly tied to episode-level mIoU risk. And the Static-Max score is composed of three terms: a semantic agreement term, a structure separation term, and a texture–uncertainty complexity term.

Semantic agreement measures how well the support and query prototypes align in feature space at layer ℓ:

$$S_{\text{sem}}(\ell) = \alpha\,\cos\big(\mathbf{P}_{\text{fg}}^\ell, \mathbf{Q}_{\text{fg}}^\ell\big) + (1-\alpha)\,\cos\big(\mathbf{P}_{\text{bg}}^\ell, \mathbf{Q}_{\text{bg}}^\ell\big), \tag{16}$$

where $\alpha \in [0,1]$ balances the contribution of foreground and background prototypes, $\mathbf{P}_{\text{fg}}^\ell, \mathbf{P}_{\text{bg}}^\ell$ are the support prototypes, and $\mathbf{Q}_{\text{fg}}^\ell, \mathbf{Q}_{\text{bg}}^\ell$ are the soft query prototypes at layer

Table 7. Layer selection ablation with DINOv2 (5-shot mIoU↑). Rule lists the per episode selector and notation includes $\mathbf{g}_\ell$ gradient w.r.t. features of layer ℓ; S_{sem}, S_{str}, C normalized semantic, structure, and complexity scores; and $\text{mIoU}_{\text{sup}}(\ell)$ support only mIoU.

Method	Rule	DeepGlobe	ISIC	Chest X-ray	FSS-1000	Avg.	Δ
Static-Max	$\arg\max_\ell \left(\alpha\, S_{\text{sem}}(\ell) + \beta\, S_{\text{str}}(\ell) + \gamma\, C(\ell)\right)$	<u>50.8</u>	<u>58.5</u>	<u>79.0</u>	<u>67.5</u>	<u>64.0</u>	<u>0.0</u>
Grad-Max	$\arg\max_\ell \|\mathbf{g}_\ell\|$	48.3	58.2	50.2	66.2	55.7	-8.2
GradΔ-Max	$\arg\max_\ell \|\mathbf{g}_\ell - \mathbf{g}_{\ell-1}\|$	48.5	58.3	50.1	66.3	55.8	-8.2
HLS (ETR)	$\arg\min_\ell \left(1 - \text{mIoU}_{\text{sup}}(\ell)\right)$	**56.3**	**68.1**	**87.1**	**79.0**	**72.6**	**+8.6**

ℓ. This term encourages higher agreement between support and query prototypes, but it depends on the baseline mask p_0, which can be biased under domain shift.

Structure separation quantifies how well foreground and background features are separated in both the query and support spaces. It is defined as:

$$S_{\text{str}}(\ell) = 1 - \tfrac{1}{2}\left[\cos\left(\mathbf{Q}^\ell_{\text{fg}}, \mathbf{Q}^\ell_{\text{bg}}\right) + \cos\left(\mathbf{P}^\ell_{\text{fg}}, \mathbf{P}^\ell_{\text{bg}}\right)\right], \quad (17)$$

where the two cosine similarities measure the alignment between foreground and background features for the query and the supports, respectively. Large values of $S_{\text{str}}(\ell)$ indicate better foreground–background separation, but this term mainly captures feature-space geometry and is only indirectly related to the final mask quality.

Texture and uncertainty complexity captures both the variability of foreground features and the ambiguity of the baseline mask. It is defined as:

$$\begin{aligned} C(\ell) &= \text{Var}\left(\mathbf{Q}^\ell_{\text{fg}}\right) + \text{Ent}(p_0), \\ \text{Ent}(p_0) &= -\tfrac{1}{|\Omega|}\sum_x \big[p_0(x)\log p_0(x) \\ &\quad + (1-p_0(x))\log\left(1-p_0(x)\right)\big], \end{aligned} \quad (18)$$

where $\text{Var}(\cdot)$ denotes the per-dimension variance of query features relative to the corresponding foreground prototype, weighted by $p_0(x)$ over the pixel set Ω, and $\text{Ent}(p_0)$ is the average Bernoulli entropy of the baseline foreground probability. This term is an indirect proxy for semantic and appearance complexity and may penalise layers that are both confident and correct.

Gradient-based proxies. These rules, including Grad-Max and GradΔ-Max, preferentially select layers that show strong loss sensitivity or pronounced changes relative to neighbouring layers.

Gradient magnitude (Grad-Max) measures how sensitive the base loss is to perturbations at layer ℓ. It is defined as:

$$\ell^\star_{\text{grad}} = \arg\max_{\ell\in\mathcal{C}} \left\|\nabla_{\mathbf{F}^\ell_q}\mathcal{L}_{\text{base}}\right\|_2, \quad (19)$$

where $\nabla_{\mathbf{F}^\ell_q}\mathcal{L}_{\text{base}}$ denotes the gradient of the base loss with respect to the query features at layer ℓ, and the ℓ_2-norm measures its magnitude. Larger values indicate stronger loss sensitivity at that layer.

Interlayer gradient change (GradΔ-Max) measures sharp changes in gradient magnitude across adjacent layers. It is defined as:

$$\ell^\star_{\Delta\text{grad}} = \arg\max_{\ell\in\mathcal{C}} \left|\left\|\nabla_{\mathbf{F}^\ell_q}\mathcal{L}_{\text{base}}\right\|_2 - \left\|\nabla_{\mathbf{F}^{\ell-1}_q}\mathcal{L}_{\text{base}}\right\|_2\right|_2, \quad (20)$$

where the difference of gradient magnitudes captures transitions between consecutive layers, and the outer ℓ_2-norm measures the strength of this change.

Implementation notes. All gradient-based selectors reuse one forward pass of backbone features and require an additional backward pass *without* parameter updates. The computational cost per episode is thus dominated by a single backpropagation through the frozen backbone.

7.3. Task-aligned HLS (ETR)

We select the routed layer by minimising an episode-level selection risk:

$$\begin{aligned} R_{\text{layer}}(\ell) &= 1 - miou_{\text{sup}}(\ell), \\ \ell^\star_{\text{HLS}} &= \arg\min_{\ell\in\mathcal{C}} R_{\text{layer}}(\ell) = \arg\max_{\ell\in\mathcal{C}} miou_{\text{sup}}(\ell), \end{aligned} \quad (21)$$

where $miou_{\text{sup}}(\ell)$ is computed within the episode by a leave-one-out procedure at layer ℓ. Each support image is treated as a pseudo query and segmented using prototypes formed from the remaining supports, and the result is averaged over the K supports.

The criterion in Eq. (21) directly measures episode-level matching risk at the representation to be adapted, rather than optimising a handcrafted surrogate. This makes it robust to layer-level transfer variability and domain shift. In practice, HLS is parameter-free, reuses the same forward features, and adds negligible overhead.

7.4. Selector analysis and takeaway

Limitations of the non-episodic selectors. We compare per-episode selectors for test-time adaptation with a frozen backbone, as shown in Table 7. The *Static-Max* rule blends

Table 8. Generalisation of HERA across different ViT-based VFMs on the CD-FSS benchmark. "ONLY" denotes directly using the frozen VFM within SSP, while "HERA" adds our three-stage select-regularise-calibrate VFM-based segmentation framework. The last two columns report the average mIoU gains of HERA over the corresponding VFM-ONLY baselines.

Backbone	DeepGlobe		ISIC2018		Chest X-ray		FSS-1000		Average		Gain w/ HERA	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SSP	40.5	49.6	35.5	48.2	74.2	74.5	79.0	80.2	57.3	63.1	0.0	0.0
DINOv2-ONLY (Large)	43.9	52.9	50.7	56.6	51.2	49.1	67.4	67.9	53.3	56.6		
DINOv2-HERA (Large)	41.2	57.8	55.6	68.7	83.2	86.9	70.2	80.3	62.6	73.4	+9.3	+16.8
DINOv3-ONLY (Large)	**47.1**	58.4	56.3	61.7	56.4	60.2	75.8	76.4	58.9	64.2		
DINOv3-HERA (Large)	44.6	**63.4**	**61.2**	**73.6**	**85.8**	**87.9**	**81.6**	**86.7**	**68.3**	**77.9**	+9.4	+13.7
CLIP-L/14-ONLY (OpenAI)	20.4	24.4	30.1	37.4	48.6	48.4	53.8	54.5	38.2	41.2		
CLIP-L/14-HERA (OpenAI)	24.3	38.1	37.8	53.0	68.4	82.2	61.8	74.2	48.1	61.9	**+9.9**	**+20.7**
CLIP-L/14-ONLY (DataComp-XL)	23.3	27.5	39.9	46.2	59.2	59.4	58.4	59.5	45.2	48.2		
CLIP-L/14-HERA (DataComp-XL)	27.2	38.5	40.1	53.1	78.1	83.8	66.9	73.4	53.1	62.2	+7.9	+14.0

three normalized cues and selects the layer with the largest $\alpha' S_{\text{sem}}(\ell)+\beta' S_{\text{str}}(\ell)+\gamma' C(\ell)$ (see Eqs. (16) to (18)). These scores measure representation quality in feature space, including semantic agreement, structure separation, and texture or uncertainty, but they do not measure *task fit* for the episode. They lack episode-level feedback and are therefore unstable across domains. Specifically, S_{sem} inherits bias from the baseline mask p_0, S_{str} rewards orthogonality that does not guarantee correct masks, and $C(\ell)$ can penalise layers that are confident and correct. The mixture weights α', β', γ' are domain-specific. Consequently, Static-Max averages 64.0 mIoU.

Gradient-based proxies capture loss sensitivity rather than alignment. *Grad-Max* selects the layer with the largest gradient norm (see Eq. (19)), and *Grad*Δ*-Max* looks for sharp inter-layer gradient changes (see Eq. (20)). In ViT backbones such as DINOv2 [8] and DINOv3 [10], blocks are architecturally homogeneous and connected by residual paths and layer normalisation. This can cause gradients to grow toward the last blocks, so both rules tend to collapse to deep layers irrespective of the episode semantics. This Grad-CAM-style assumption therefore fails, and the selected layer often has the largest perturbation rather than being the most suitable for segmentation. These proxies correlate weakly with support and query matching quality and yield 55.7 and 55.8 mIoU on average.

Why HLS (ETR) is better. Our *HLS* uses a task-aligned criterion that directly minimises the episode-level selection risk $\ell^\star_{\text{HLS}} = \arg\min_{\ell \in \mathcal{C}} \left(1 - miou_{\text{sup}}(\ell)\right)$ (see Eq. (21)). It performs a self-prediction evaluation within the episode. Each support is treated as a pseudo query and is segmented using prototypes from the remaining supports, and the score is the support-only mIoU at layer ℓ. This provides dynamic, episode-aware feedback aligned with the target objective, with no extra parameters, and negligible overhead. HLS reaches 72.6 mIoU, which is +8.6 over Static-Max and +16.8 over the best gradient proxy. The gain is especially large on Chest X-ray [1, 6] (from 50.1 to 87.1 mIoU, +37.0), and the gap widens on other VFM backbones.

8. Generalising HERA across ViT-based VFMs

In this section, we evaluate the generalisation of HERA across different ViT-based visual foundation models (VFMs). We use SSP as a common few-shot segmentation framework and treat several representative ViT-based VFMs as frozen backbones. Specifically, we consider the self-supervised DINOv2 [8] and DINOv3 [10], and the contrastively trained CLIP-L/14 (OpenAI) [9] and CLIP-L/14 (DataComp-XL) [5]. On the CD-FSS benchmark, we keep all training and evaluation settings identical to those in the main experiments. The only degrees of freedom are the choice of backbone and whether HERA is applied. In Table 8, the "ONLY" rows correspond to directly using the frozen VFM within SSP, whereas the "HERA" rows apply our three-stage select–regularise–calibrate adaptation framework on top of the same VFM. The last two columns report the average mIoU gains achieved by HERA over the corresponding VFM-ONLY baselines.

HERA yields consistent and substantial relative improvements across the ViT-based VFMs. For self-supervised VFMs, DINOv2-HERA increases the average mIoU from 53.3/56.6 to 62.6/73.4, corresponding to gains of **+9.3** and **+16.8** in the 1-shot and 5-shot settings, respectively. DINOv3-HERA improves over DINOv3-ONLY from 58.9/64.2 to 68.3/77.9, corresponding to **+9.4** and **+13.7** mIoU, and surpasses 80% mIoU on several targets. On CLIP-based VFMs, HERA still shows clear and consistent gains. For OpenAI CLIP-L/14, the average mIoU

increases from 38.2/41.2 to 48.1/61.9, i.e. gains of **+9.9** and **+20.7**. For DataComp-XL CLIP-L/14, HERA delivers improvements of **+7.9** and **+14.0** in the 1-shot and 5-shot regimes, respectively. These results demonstrate that, under a unified hyperparameter setting and without any source-domain retraining, HERA systematically enhances the performance of diverse ViT-based VFMs on CD-FSS.

In CD-FSS, CLIP is primarily pre-trained for image-text alignment and image-level recognition, so its patch-level representations are naturally more global and less tailored to precise segmentation boundaries. This mismatch makes CD-FSS a particularly challenging downstream task for CLIP. Nevertheless, HERA even achieves some of the largest relative gains on CLIP backbones, especially in the 5-shot setting. This indicates that our select–regularise–calibrate mechanism effectively mitigates the mismatch between pre-training objectives and dense prediction. When the underlying VFM is more aligned with dense recognition, as in the case of DINOv3, HERA further amplifies its potential and achieves strong, often state-of-the-art performance on CD-FSS. Overall, these observations suggest that HERA can serve as a generic adaptor for ViT-based VFMs and is well-positioned to benefit future VFMs that are even better suited to class-disjoint few-shot segmentation.

9. Component Ablation of DINOv2

The sequence Select-Regularise-Calibrate yields monotonic improvements. HLS provides the dominant gain by stabilising the chosen adaptation layer for each episode. PGR reduces attention noise, such as spurious far-field peaks, while preserving global coverage. PAC then corrects residual artifacts along thin boundaries and in low-contrast regions. Gains are larger in the one-shot regime, where supervision is scarcer, which is consistent with the design intent. These results show that the hierarchical refinements generalise from DINOv3 [10] to DINOv2 [8] and to other VFMs, indicating effectiveness that is agnostic to the backbone.

Table 9. Component ablation on DINOv2 (average mIoU). Δ_{V0} denotes the improvement over the V0 baseline, and Δ_{prev} denotes the improvement relative to the row above. Best scores in bold.

Setting	Avg. 1-shot	Avg. 5-shot	Δ_{V0} (1s / 5s)	Δ_{prev} (1s / 5s)
V0 baseline	57.03	68.49	0.00 / 0.00	0.00 / 0.00
+ HLS	60.34	72.64	+3.31 / +4.15	**+3.31 / +4.15**
+ PGR	61.10	73.28	+4.07 / +4.79	+0.76 / +0.64
+ PAC	**62.58**	**73.42**	**+5.55 / +4.93**	+1.48 / +0.14

All selectors reuse cached features. HLS uses pseudo-query scoring on the support only and therefore adds negligible overhead. PGR has no trainable parameters. PAC operates as a lightweight residual fusion and is gated automatically in five-shot episodes. Consequently, the overall parameter and runtime budgets remain low while providing improvements that are aligned with the task.

10. 1-shot Support Augmentation Study

In the 1-shot setting, we explore a lightweight support-augmentation strategy designed to stabilise the leave-one-out scoring used by HLS. To maintain compatibility with HERA's prototype-construction and refinement pipeline, we adopt a simple *soft-copy* instance-synthesis procedure. Given a support image and its foreground mask, we paste the foreground onto a randomly sampled location within the same image to generate additional pseudo-supports while maintaining mask consistency. Formally, for a foreground instance $(I_{\mathrm{fg}}, M_{\mathrm{fg}})$ and a background view $(I_{\mathrm{bg}}, M_{\mathrm{bg}})$, the synthesized example is:

$$\begin{aligned} I'_{\mathrm{bg}} &= I_{\mathrm{bg}} + M_{\mathrm{fg}}^{(x,y)} \odot I_{\mathrm{fg}}, \\ M'_{\mathrm{bg}} &= \mathrm{clip}\big(M_{\mathrm{bg}} + M_{\mathrm{fg}}^{(x,y)},\, 0, 1\big), \end{aligned} \tag{22}$$

where $M_{\mathrm{fg}}^{(x,y)}$ denotes the foreground mask placed at a random location (x, y), and clip ensures that the resulting mask remains within $[0, 1]$. This soft-copy operation preserves foreground semantics while introducing controlled variations in appearance.

Across backbones and datasets, we have an observation about soft-copy augmentation: the synthesised pseudo-supports remain highly correlated in semantic structure, causing the leave-one-out estimates used by HLS to fluctuate only minimally. Nevertheless, HERA already maintains competitive 1-shot performance under this simple augmentation mechanism.

Importantly, this observation shows that our framework is highly compatible with more powerful support-augmentation strategies. Because HERA only requires soft masks, foreground prototypes, and within-episode consistency measures, it can seamlessly incorporate more advanced augmentations, including generative instance synthesis, structure-preserving shape deformation, and domain-aware appearance perturbations. These directions offer substantial potential for improving 1-shot stability and overall cross-domain generalisation capability.

In summary, soft-copy serves as a minimal yet stable augmentation method for 1-shot episodes, demonstrating that HERA remains competitive even under limited synthetic diversity. More expressive support-synthesis strategies remain compatible with our design and provide promising avenues for further improving robust and scalable 1-shot performance across diverse domains.

11. More Visualisation Results about HERA

We provide additional qualitative results under the 1-shot setting to support the analyses in this paper. We include (i) segmentation predictions across four target datasets and (ii) layer-wise foreground probability maps from all ViT layers, which visually illustrate the layer sensitivity phenomenon and the semantic transition exploited by HERA.

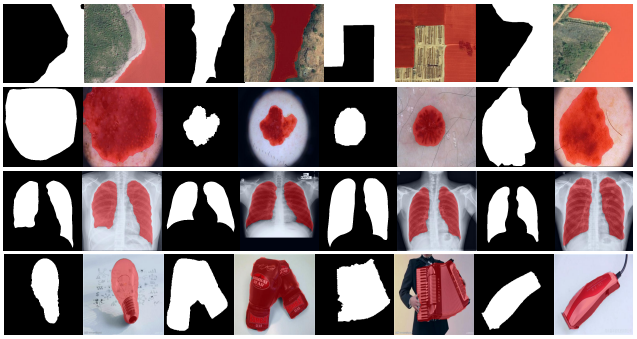

Figure 1. 1-shot segmentation results on the DeepGlobe, ISIC, Chest X-ray and FSS-1000 datasets, showing the predicted masks overlaid on the query images.

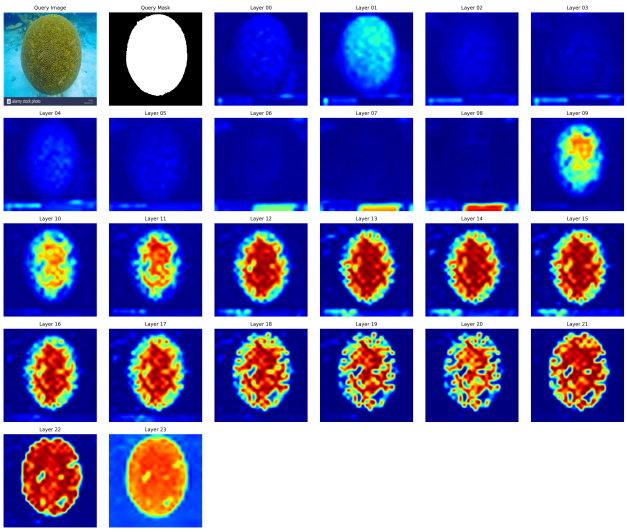

Figure 2. Layer-wise predicted foreground probability maps (Layer 00–23) showing the emergence of semantic structures across depth under the 1-shot setting.

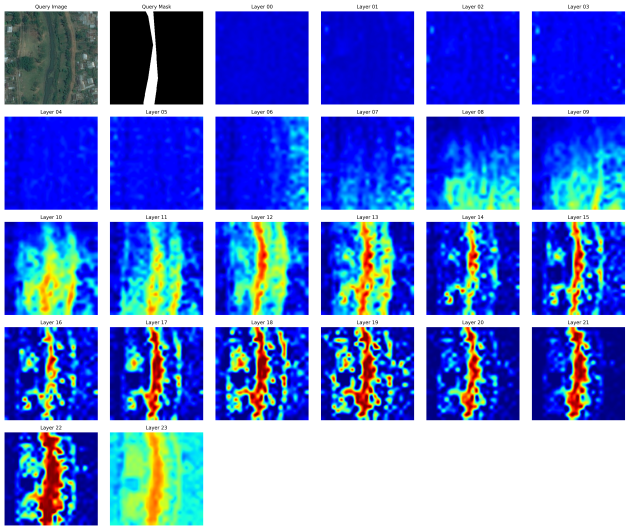

Figure 3. Additional layer-wise probability maps demonstrating consistent semantic transitions across diverse 1-shot episodes.

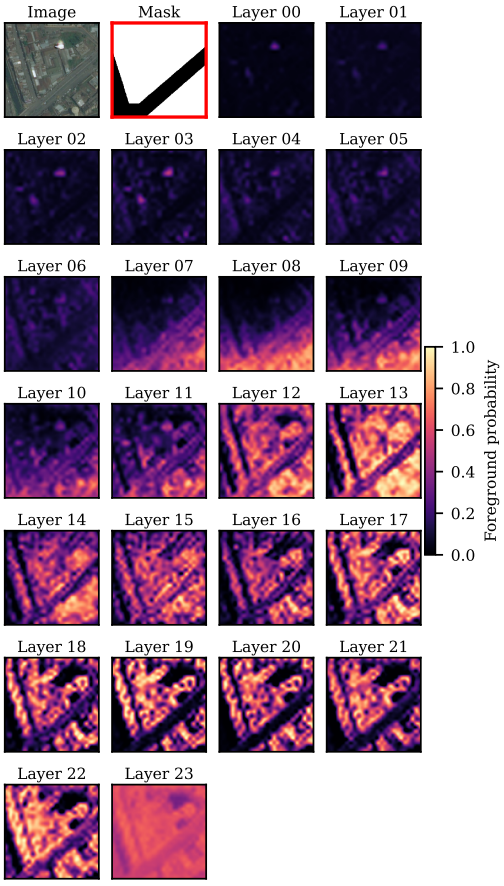

Figure 4. Per-layer probability responses for a representative 1-shot query, highlighting the mid-layer semantic turning point captured by HERA.

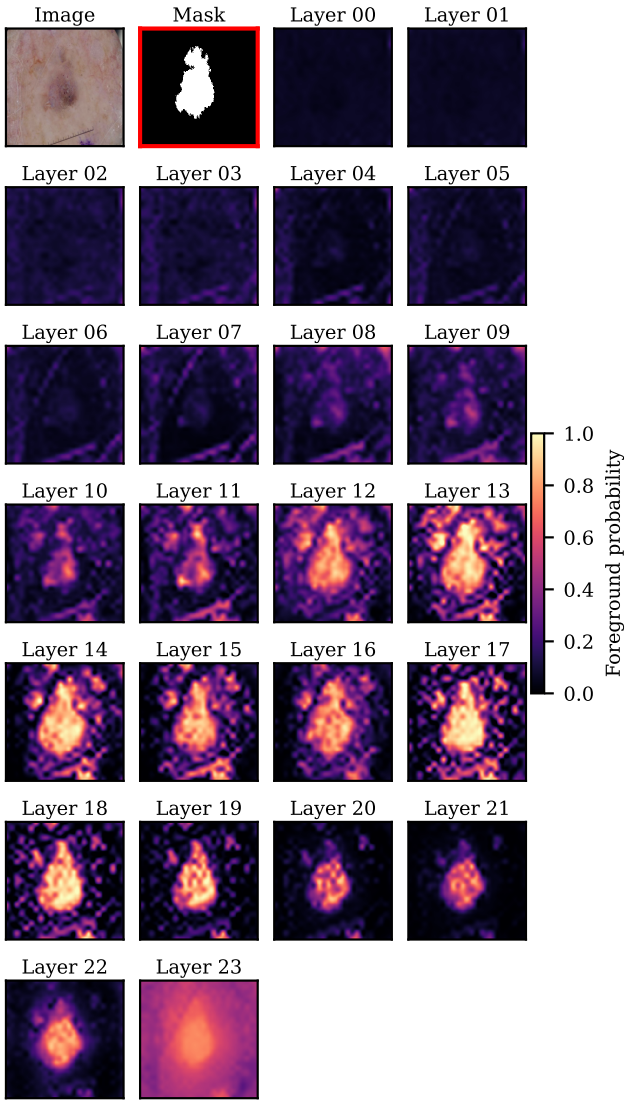

Figure 5. Layer-wise probability visualisation for another 1-shot sample, illustrating stable depth-dependent semantic strengthening across datasets.

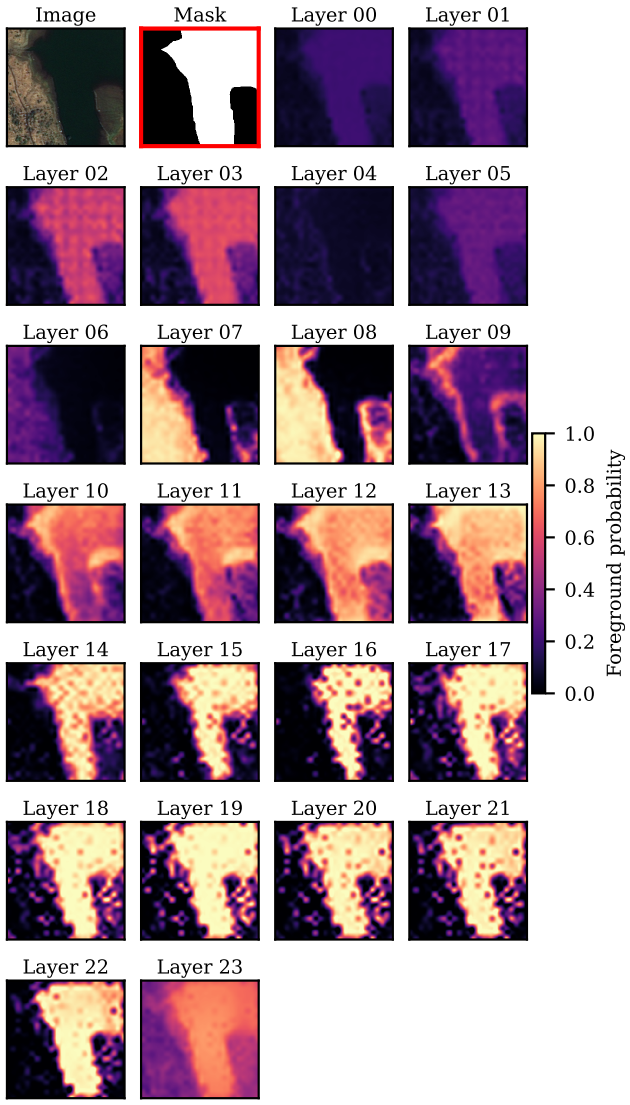

Figure 6. Foreground probability maps across all ViT layers for a third 1-shot episode, revealing the deep-layer aggregation pattern leveraged by the proposed routing mechanism.

References

[1] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013. 1, 6

[2] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 1, 2

[3] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 1, 2

[4] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *European conference on computer vision*, pages 701–719. Springer, 2022. 1

[5] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 6

[6] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013. 1, 6

[7] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020. 1

[8] 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. 6, 7

[9] 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*, pages 8748–8763. PmLR, 2021. 6

[10] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 6, 7

[11] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1, 2