A MLP ADAPTATION ON THE SELECTED REPRESENTATION LAYER

MLP head variants. With the backbone frozen and the working representation layer chosen by HLS, we attach a small MLP head at that layer to improve support to 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 A.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. Starting from M0 at 65.66 mIoU in one shot and 75.20 mIoU in five shot, as shown in Table A, 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 stabilizes 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 five 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 Regularize 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.

B LOCAL FUSION AROUND THE ROUTED LAYER

After HLR selects the best single layer ℓ_{single} for each episode we form a compact neighborhood U centered 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 as in Sec. 3.2.1. 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:

$$w_\ell \;=\; \frac{\exp\big(-\beta\, r_\ell - \mathrm{dist}(\ell,\ell_{23})/\tau\big)}{\sum_{j\in U}\exp\big(-\beta\, r_j - \mathrm{dist}(j,\ell_{23})/\tau\big)}, \qquad F^U \;=\; \sum_{\ell\in U} w_\ell\, F^\ell, \tag{B.1}$$

Here $\beta > 0$ 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 stabilizes routing.

Table B.1. Local fusion centered 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. Table B.1 compares single layer routing with local fusion. Local fusion centered 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 B.2. 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
$F^0 + L_{23}, \tau$=2.0	**44.59**	**63.43**	**61.17**	**73.64**	85.80	87.88	**81.59**	**86.69**	**68.29**	**77.91**
$F^0 + L_{23}, \tau$=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=ℓ^*, τ=2.0	42.32	63.11	56.41	68.25	**87.44**	**88.41**	81.16	85.58	66.83	76.34

Table B.2 reports larger gains on DeepGlobe and ISIC, consistent with evidence drift across episodes and the need for deeper semantic aggregation. Therefore, we adopt local fusion with τ=2.0 and retain L_{23} in the candidate pool by default.

C Pixelwise Adaptive Calibration: Details

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, $\mathbf{P}_{\rm fg}$ and $\mathbf{P}_{\rm bg}$, are computed by masked averaging over support features at ℓ^*. We define the prototype difference logit as

$$\ell_{\rm sim}(x) = \tau_{\rm sim}\big[\cos\big(\mathbf{F}_q(x), \mathbf{P}_{\rm fg}\big) - \cos\big(\mathbf{F}_q(x), \mathbf{P}_{\rm bg}\big)\big]\,, \tag{C.1}$$

where $\tau_{\rm sim}$ 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 normalized 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

$$\ell_{\rm attn}(x) = \tau_{\rm attn}\big[(\widetilde{A}\,p_0)_x\big], \tag{C.2}$$

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 as

$$\ell_{\rm img}(x) = \tau_{\rm img}\big[\cos\big(\mathbf{v}(x), \mathbf{u}_{\rm fg}\big) - \cos\big(\mathbf{v}(x), \mathbf{u}_{\rm bg}\big)\big]\,, \tag{C.3}$$

Here $\mathbf{u}_{\rm fg}$ and $\mathbf{u}_{\rm bg}$ are image level prototypes, and $\tau_{\rm img}$ 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:

$$\ell_{\rm final}(x) = \ell_0(x) + w_{\rm sim}\,\ell_{\rm sim}(x) + w_{\rm attn}\,\ell_{\rm attn}(x) + w_{\rm img}\,\ell_{\rm img}(x), \tag{C.4}$$

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 realize a hierarchical Select, Regularize, and Calibrate pipeline that adapts at test time with a frozen backbone.

D Adaptive Gating for Pixelwise Adaptive Calibration

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 synthesize two augmented views of the support to obtain two votes.

Table D.1. 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 threshold $2/5$ for five shot.

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	-

Table D.2. 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

Analysis. 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 complementarity (see Table 2). For PAC gating, Table D.1 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 reduces the trigger rate and lowers accuracy to 77.22 to 77.85 mIoU. The per dataset study in Table D.2 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 $\ell_{\rm sim}$, the one hop attention propagation $\ell_{\rm attn}$, and the image appearance cue $\ell_{\rm img}$ contribute +0.30, +0.22, and +0.18 mIoU. Using all three reaches 77.91 mIoU, a further +0.64 (see Table 3). Together, HLS stabilizes the routed layer, PGR sharpens locality, and PAC corrects pixel level logits, yielding a cumulative gain under a frozen backbone.

E Alternative Layer Selection Criteria and DINOv2 Results

E.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 $\mathbf{P}_{\rm fg}^\ell$ and $\mathbf{P}_{\rm bg}^\ell$. Given a baseline foreground probability $p_0(x) \in [0,1]$ for the query, we build soft masked query prototypes as

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

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

$$\tilde{s}_\ell = \frac{s_\ell - \min_{j\in\mathcal{C}} s_j}{\max_{j\in\mathcal{C}} s_j - \min_{j\in\mathcal{C}} s_j + \varepsilon}, \quad \varepsilon = 10^{-8}. \tag{E.2}$$

E.2 Selectors other than HLS: definitions, intuition, and caveats

We group the non episodic selectors in Table 4 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 normalized across the candidate set $\mathcal{C}$ within each episode*. Prototypes and the baseline mask p_0 follow the definitions in Sec. E.

Table E.1. Notation for layer selection in the episodic setting. All scalar layer scores are range normalized 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
$\mathbf{P}_{\text{fg}}^\ell,\ \mathbf{P}_{\text{bg}}^\ell$	Support foreground and background prototypes at layer ℓ
$\mathbf{Q}_{\text{fg}}^\ell,\ \mathbf{Q}_{\text{bg}}^\ell$	Soft masked query prototypes (see Eq. equation E.1)
$p_0(x) \in [0,1]$	Baseline foreground probability on the query
$\mathrm{mIoU}_{\text{sup}}(\ell)$	Support only pseudo query mIoU at layer ℓ (risk proxy)

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

$$\ell^\star_{\text{static}} = \arg\max_{\ell \in \mathcal{C}} \left[\alpha' S_{\text{sem}}(\ell) + \beta' S_{\text{str}}(\ell) + \gamma' C(\ell) \right], \qquad \alpha', \beta', \gamma' \geq 0,\ \alpha' + \beta' + \gamma' = 1. \tag{E.3}$$

Caveat: weights are domain and task specific, and the objective is a surrogate not directly tied to episode level mIoU risk.

Component scores of Static-Max.

- **Semantic agreement**
$$S_{\text{sem}}(\ell) = \alpha\ \cos\big(\mathbf{P}_{\text{fg}}^\ell, \mathbf{Q}_{\text{fg}}^\ell\big) + (1-\alpha)\ \cos\big(\mathbf{P}_{\text{bg}}^\ell, \mathbf{Q}_{\text{bg}}^\ell\big), \quad \alpha \in [0,1]. \tag{E.4}$$
Intuition: encourages higher agreement between support and query prototypes. *Caveat*: depends on the baseline mask p_0, which can be biased under shift.
- **Structure separation**
$$S_{\text{str}}(\ell) = 1 - \tfrac{1}{2}\left[\cos\big(\mathbf{Q}_{\text{fg}}^\ell, \mathbf{Q}_{\text{bg}}^\ell\big) + \cos\big(\mathbf{P}_{\text{fg}}^\ell, \mathbf{P}_{\text{bg}}^\ell\big)\right]. \tag{E.5}$$
Intuition: encourages foreground and background orthogonality in the query and support spaces. *Caveat*: measures feature geometry rather than final mask quality.
- **Texture and uncertainty complexity**
$$C(\ell) = \mathrm{Var}\big(\mathbf{Q}_{\text{fg}}^\ell\big) + \mathrm{Ent}\big(p_0\big), \quad \mathrm{Ent}(p_0) = -\tfrac{1}{|\Omega|} \sum_x \left[p_0 \log p_0 + (1-p_0)\log(1-p_0)\right]. \tag{E.6}$$
Here $\mathrm{Var}(\cdot)$ denotes the per dimension variance of query features relative to the corresponding prototype, weighted by p_0. *Caveat*: an indirect proxy that may penalize layers that are confident and correct.

Gradient based proxies. These rules favor layers with large loss sensitivity or sharp changes across adjacent layers.

Gradient magnitude (Grad-Max).

$$\ell^\star_{\text{grad}} = \arg\max_{\ell \in \mathcal{C}} \left\| \nabla_{\mathbf{F}_q^\ell} \mathcal{L}_{\text{base}} \right\|_2. \tag{E.7}$$

Intuition: select the layer to which the base loss is most sensitive. *Caveat*: residual paths and normalization in ViTs can amplify gradients in later layers, biasing the choice.

Interlayer gradient change (GradΔ-Max).

$$\ell^\star_{\Delta\text{grad}} = \arg\max_{\ell \in \mathcal{C}} \left\| \left\| \nabla_{\mathbf{F}_q^\ell} \mathcal{L}_{\text{base}} \right\|_2 - \left\| \nabla_{\mathbf{F}_q^{\ell-1}} \mathcal{L}_{\text{base}} \right\|_2 \right\|_2. \tag{E.8}$$

Intuition: detect transition points across adjacent layers. *Caveat*: still a gradient scale proxy, only weakly coupled to episode level decisions.

Implementation notes. All rules reuse a single forward pass of backbone activations. Gradient based proxies require one backward pass *without* parameter updates. The per episode cost is dominated by a single backpropagation through the frozen backbone.

E.3 TASK ALIGNED HLS (ETR)

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

$$R_{\text{layer}}(\ell) \;=\; 1 - miou_{\text{sup}}(\ell), \qquad \ell^{\star}_{\text{HLS}} \;=\; \arg\min_{\ell \in \mathcal{C}} R_{\text{layer}}(\ell) \;=\; \arg\max_{\ell \in \mathcal{C}} miou_{\text{sup}}(\ell). \tag{E.9}$$

Here $miou_{\text{sup}}(\ell)$ 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.

Rationale. The criterion in equation E.9 directly measures episode level matching risk at the representation to be adapted, rather than optimizing 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.

E.4 SELECTOR ANALYSIS AND TAKEAWAY

Why the three non episodic selectors underperform. Table 4 compares per episode layer selectors with a frozen backbone. The *Static Max* rule blends 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. equation E.4 to equation E.6). 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 penalize layers that are confident and correct. The mixture weights α', β', γ' are domain specific. Consequently, Static Max averages 71.9 mIoU.

Gradient based proxies capture loss sensitivity rather than alignment. *Grad Max* selects the layer with the largest gradient norm (see Eq. (E.7)), and *GradΔ Max* looks for sharp inter layer gradient changes (see Eq. (E.8)). In ViT backbones such as DINOv2 and DINOv3, blocks are architecturally homogeneous and connected by residual paths and layer normalization. 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 73.1 and 73.2 mIoU on average.

Why HLS (ETR) is better. Our *HLS* uses a task aligned criterion that directly minimizes the episode level selection risk $\ell^{\star}_{\text{HLS}} = \arg\min_{\ell \in \mathcal{C}} \big(1 - miou_{\text{sup}}(\ell)\big)$ (see equation E.9). 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 low variance, no extra parameters, and negligible overhead. HLS reaches 76.7 mIoU, which is +4.8 over Static Max and +3.5 over the best gradient proxy. The gain is especially large on ISIC (from 48.2 to 73.6 mIoU, +25.4), and the gap widens on other VFM backbones such as DINOv2.

E.5 DINOV2: COMPONENT ABLATION (1-/5-SHOT) AND TAKEAWAYS

Table E.2. 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 (fusion=off, refine=off)	57.03	68.49	0.00 / 0.00	0.00 / 0.00
+ HLS (enable fusion and routing)	60.34	72.64	+3.31 / +4.15	**+3.31 / +4.15**
+ PGR (Gaussian prior for attention)	61.10	73.28	+4.07 / +4.79	+0.76 / +0.64
+ PAC (auto refine)	**62.58**	**73.42**	**+5.55 / +4.93**	+1.48 / +0.14

Analysis. The sequence *Select* $\rightarrow$ *Regularize* $\rightarrow$ *Calibrate* yields monotonic improvements. HLS provides the dominant gain by stabilizing 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 generalize from DINOv3 to DINOv2 and to other VFMs, indicating effectiveness that is agnostic to the backbone.

Practical remarks. 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.

F DISCLOSURE OF LARGE LANGUAGE MODEL (LLM) USAGE

We used large language models (LLMs) only to assist with writing. Specifically, LLMs were employed to polish wording, improve clarity, and refine the presentation (grammar, coherence, and flow) of certain sections. All scientific ideas, methodology, experiments, analyses, and conclusions were conceived and executed exclusively by the authors. LLM assistance was limited to language-related edits and suggestions. All outputs were reviewed and revised by the authors. The use of LLMs did not contribute to the research design, data collection, data analysis, or the intellectual content of the findings.