

PROTON+: Contamination-Robust Prototype Banks for Test-Time Online OOD Detection with Medical VLMs

A follow-up to PROTON [★]

Abhijit Das¹, Nichula Wasalathilaka², Yifan Lu¹, Adinath Dukre¹,
Dwarikanath Mahapatra³, Shadab Khan⁴, and Imran Razzak^{1,5}

¹ Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

² University of Peradeniya, Sri Lanka

³ Khalifa University, Abu Dhabi, UAE

⁴ ADIA Lab, Abu Dhabi, UAE

⁵ MedOS, Abu Dhabi, UAE

{abhijit.das, imran.razzak}@mbzuai.ac.ae

Abstract. Medical vision-language models (VLMs) classify images from text prompts alone and must flag inputs that lie outside their prompted classes. Our detector PROTON scores each test image by its distance to an online bank of class prototypes filled from confident predictions, and fuses that distance with the softmax score, so that covariate-shifted inputs which look confident to the softmax can still be caught. Extending PROTON from retinal photographs to chest radiographs and histopathology, we find that its confidence gate admits the confidently misclassified shifted inputs it is meant to exclude, so the prototypes absorb the shift. PROTON+ anchors every class prototype to the dominant imaging-domain mode of the bank with a mean-shift step, admits only samples that are consistent with their class prototype, and fuses online z-scores instead of raw scores. With the hyper-parameters of PROTON unchanged, PROTON+ improves over MCM on the covariate and far shifts of all three modalities and on every stream order, reaching 93.9 AUROC on paediatric chest radiographs, where MCM is at chance, and 16.4 points above MCM on ultra-widefield fundus photographs.

Keywords: Out-of-distribution detection · Vision-language models · Test-time adaptation · Prototype bank.

1 Introduction

Medical VLMs such as FLAIR, UniMed-CLIP and QuiltNet classify an image from text prompts alone [20,12,7]. A screening service prompted with “normal” and

[★] Follow-up to our PROTON paper [2], MICCAI 2026 (<https://arxiv.org/abs/2606.20913>). Code, benchmark manifests, cached features and a browser demo: <https://aj-das-research.github.io/PROTON/>.

“diabetic retinopathy” also receives glaucoma cases, ultra-widefield photographs from a different camera and images that are not retinas, and each must be flagged for review instead of being assigned to a class. Static detectors read only the softmax over the class prompts, whether as its maximum (MCM [17]), as a logit or energy [6,15], with patch-level matches (GL-MCM [18]) or over augmented prompts (TAG [16]), and no static score works across semantic, covariate and far shifts on medical VLMs [9,5]. A covariate-shifted image is classified confidently, so the softmax carries no trace of the shift, while the image embedding does. PROTON [2] exploits this with an online bank of class prototypes built from confident test predictions, a prototype-distance score, and a variance-driven fusion with MCM, without labels, training or prompt engineering. Test-time learners such as TTL, TINS and STAR [24,23,4] adapt prompts or text embeddings from the stream instead and target the accuracy of the adapted classifier.

This follow-up extends PROTON to three modalities with a new codebase and examines what the bank actually holds. The confidence gate can only exclude inputs with a low maximum probability, which MCM already detects, and a covariate-shifted input passes it by definition. On the ultra-widefield stream 97% of the shifted images pass the gate, 148 of 150 are assigned to the diabetic-retinopathy class, and that class bank ends with 63% shifted members, so the prototype moves towards the shift it should detect. With an oracle that never admits a shifted sample the same prototype distance separates the shifted images at 93.1 AUROC, so the signal is there and the admission rule is the fault.

PROTON+ keeps the bank and the fusion of PROTON and replaces the admission rule by two label-free statistics of the bank itself, as Figure 1 shows. A mean-shift step over the union of all class banks finds the dominant imaging-domain mode, and each class prototype is the mean of its members weighted by proximity to that mode, so a confidently classified but shifted cluster receives almost no weight. A sample enters the bank only if its distance to the prototype of its predicted class is typical of the samples admitted before it, and the two scores are fused as online z-scores, which removes the scale mismatch between a probability and a cosine distance. Our contributions are:

1. PROTON+, three changes to PROTON that cost no label, no training and the same hyper-parameters, and that improve over MCM on every covariate and far shift of three modalities, by 16.4 to 47.3 AUROC points on covariate shift.
2. An analysis of prototype-bank contamination, with an oracle admission reference that bounds what a clean bank would give, and the finding that the prototype distance alone is the strongest single detector on covariate and far shifts.
3. A released library with three benchmarks, nine zero-shot and privileged baselines, seeded manifests, cached features and a browser demo, so that any VLM and any folder of images can be evaluated with three commands.

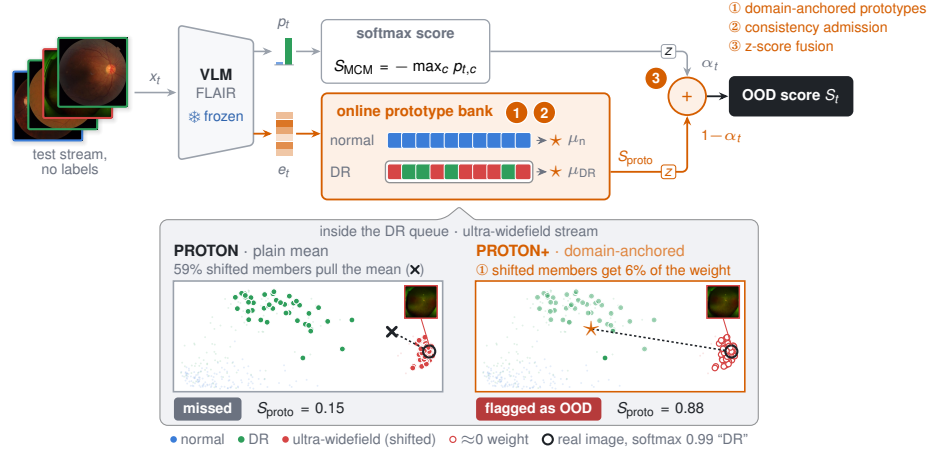


Fig. 1. A confident softmax hides the shift, the anchored prototype exposes it. Top: a frozen VLM gives each stream image a softmax p_t , read by MCM, and an embedding e_t , read by the online prototype bank. PROTON+ changes three things, in orange: ② only consistent samples enter the bank, ① class prototypes are anchored to the dominant imaging-domain mode of the bank, and ③ the two scores are fused as online z-scores. Bottom: the “diabetic retinopathy” queue at the end of an ultra-widefield stream, in a 2-D linear projection of the real FLAIR embeddings (horizontal: ID-to-ultra-widefield direction, vertical: normal-to-DR direction). PROTON’s plain mean drifts toward the 59% shifted members, so a real ultra-widefield image that the softmax labels “DR” with confidence 0.99 sits at prototype distance 0.15 and is missed. PROTON+ gives those members 6% of the anchor weight, keeps the prototype on the DR images, and the same image sits at 0.88 and is flagged.

2 Method

Setting and PROTON. A frozen VLM maps an image x_t to an ℓ_2 -normalised embedding $e_t \in \mathbb{R}^d$ and the C class prompts to text embeddings $T = [t_1, \dots, t_C]$, computed once. The zero-shot classifier is $p_t = \text{softmax}(e_t^\top T / \tau)$ with the temperature τ learned by the VLM, and MCM scores $S_{\text{MCM}}(x_t) = -\max_c p_{t,c}$, higher meaning more OOD. Images arrive one at a time without labels, and the detector may use only the images it has already seen. PROTON keeps one first-in-first-out queue $\mathcal{D}_c$ of at most M embeddings per class, and after scoring an image enqueues it into the queue of its predicted class $\hat{c} = \arg \max_c p_{t,c}$ if $\max_c p_{t,c} \geq \gamma$. The prototype of class c is the normalised mean of its queue, $\mu_c = \bar{e}_c / \|\bar{e}_c\|_2$, the prototype distance is

$$S_{\text{proto}}(x_t) = 1 - \max_c e_t^\top \mu_c, \quad (1)$$

and once every queue holds $K_{\min}$ members the score is

$$S_t = \alpha_t S_{\text{MCM}}(x_t) + (1 - \alpha_t) S_{\text{proto}}(x_t), \quad \alpha_t = \alpha_{\max} - \sigma(100(\sigma_t^2 - \sigma_0^2)) \Delta\alpha, \quad (2)$$

where $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$, σ_t^2 is the running variance of S_{MCM} (Welford [22]), σ_0^2 a reference variance and $\sigma(\cdot)$ the logistic function, so that α_t moves from $\alpha_{\max}$ towards $\alpha_{\min}$ when the variance of MCM rises. Before calibration the score is S_{MCM} alone.

Why the bank gets contaminated. The gate excludes only inputs with a low maximum probability, which are the inputs MCM already detects, and a covariate-shifted input is one the model classifies confidently for the wrong reason. Such inputs land in the queue of the class they were assigned to, the prototype becomes the mean of the ID cluster and the shifted cluster and lies between them, and Equation (1) no longer separates the two. Equation (2) adds a second defect, because S_{MCM} lies in $[-1, -1/C]$ while S_{proto} lies in $[0, 2]$, so α_t does not control the relative influence of the two signals.

PROTON+: domain-anchored prototypes. All ID classes of a benchmark share one imaging domain, so the union of all queues has one dominant mode whenever ID members outnumber the shifted members, which holds for streams with an ID fraction of at least 0.7 in our data. A mean-shift step [1] finds that mode: starting from the normalised mean μ of all members $\mathcal{D} = \bigcup_c \mathcal{D}_c$ we iterate three times

$$\mu \leftarrow \frac{\sum_{e \in \mathcal{D}} w(e, \mu) e}{\left\| \sum_{e \in \mathcal{D}} w(e, \mu) e \right\|_2}, \quad w(e, \mu) = \exp((e^\top \mu - 1)/h), \quad (3)$$

with a bandwidth h on the cosine scale, and call the result μ_{dom} . Each class prototype is then the domain-weighted mean of its own queue, $\mu_c \propto \sum_{e \in \mathcal{D}_c} w(e, \mu_{\text{dom}}) e$ normalised to unit length. A member in a shifted cluster at a cosine distance of several times h from the anchor therefore contributes almost nothing, and Equation (1) is unchanged.

Consistency admission. After calibration a sample enters the queue of its predicted class only if its confidence passes the gate and its distance $d_t = 1 - e_t^\top \mu_{\hat{c}}$ is typical of the samples admitted before it, $d_t \leq m_t + \kappa s_t$, where m_t and s_t are the running mean and standard deviation of the distances of samples admitted after calibration and κ sets the tolerance. The rule becomes active once $K_{\min}$ such distances exist, costs one dot product and three scalar updates, and lets the FIFO eviction flush early contamination.

Scale-free fusion. We keep Equation (2) with the same weight α_t but replace each score by its online z-score, with the statistics of S_{MCM} running over every sample seen and those of S_{proto} over the samples seen since calibration. Before calibration the z-scored MCM is emitted, so the score keeps one scale across the calibration boundary. Two variants are evaluated alongside: the maximum of the two z-scores instead of the weighted sum, and a weight driven by the rejection rate of the admission rule instead of the variance of MCM. Per sample PROTON+ computes $C + 1$ dot products, and each admission triggers the anchor update, two matrix-vector products per iteration over the $CM \times d$ bank, and the state is 0.4 MB for $C = 2$, $M = 100$, $d = 512$.

3 Experiments

Benchmarks. Three benchmarks, each with two ID classes, three OOD splits and one VLM, follow the construction of Ju et al. where it exists [9]. *Ophthalmology* uses FIVES [8] normal and diabetic retinopathy as ID (150 each), FIVES glaucoma and macular degeneration as semantic shift (150 each), DeepDRiD ultra-widefield photographs [14] as covariate shift (150) and ImageNet images [3] as far shift (150). *Radiology* uses RSNA [19] normal and lung-opacity radiographs as ID (300 each), NIH ChestX-ray14 radiographs with findings outside the ID classes as semantic shift (300, de-duplicated against RSNA, which is an NIH subset), paediatric radiographs [11] of both classes as covariate shift (150) and ImageNet (150). *Pathology* uses PatchCamelyon [21] benign and malignant lymph-node patches as ID (300 each), five colorectal tissue types from NCT-CRC-HE [10] as semantic shift (300), both classes from the held-out hospital of Camelyon17-WILDS [13] as covariate shift (150) and ImageNet (150). Every image list is a seeded, released manifest, and a disjoint labelled ID split (50 per class in ophthalmology, 300 otherwise) serves only the two privileged baselines.

Models, prompts and protocol. The frozen VLMs are FLAIR (ResNet-50, $d = 512$) [20], UniMed-CLIP (ViT-B/16) [12] and QuiltNet (ViT-B/32) [7], each with its learned temperature and with one prompt policy, the modality-prefixed class prompt of the public benchmark code (“A fundus photograph of {class}”, “x-ray of {class}”, “histopathology image of {class}”). FLAIR averages text-encoder states over all tokens of a padded batch, so its class prompts are encoded as one batch, as its forward pass does. Zero-shot ID accuracies are 84.0%, 69.2% and 59.3%. A stream contains all ID test images and all images of one OOD split in a random order (ID fraction 0.50 to 0.80), every online detector runs over five random orders, and all its numbers are the mean $\pm$ SD over these orders, while static scores do not depend on the order. We report AUROC, the area under the precision-recall curve with OOD as the positive class (AUPR) and the false-positive rate at 95% OOD recall (FPR@95), in percent. PROTON uses $\gamma = 0.7$, $M = 100$, $K_{\min} = 5$, $\alpha \in [0.3, 0.7]$ and $\sigma_0^2 = 0.02$ as in its paper, and PROTON+ adds $h = 0.1$ and $\kappa = 2$, fixed before the experiments from the within-class cosine spread and the two-sigma rule. Features are extracted once per benchmark on one GPU in under a minute, and every detector runs on a CPU.

Baselines. Max-Logits, Energy, MCM, GL-MCM, TAG with the class prompt and ten random word shuffles of it, and the hierarchical prompts of Ju et al. at one and five descriptions per class. TAG reduces to MCM for single-word prompts, and Entropy equals MCM for two classes and is omitted. kNN and Mahalanobis distances to the labelled split are privileged upper references.

3.1 Does the bank help, and where does PROTON+ add?

Table 1 lists every zero-shot score on the ophthalmology benchmark, and Table 2 the three modalities. A row above the MCM row beats the static score

Table 1. Zero-shot OOD detection with FLAIR on the ophthalmology benchmark. AUROC and AUPR higher is better, FPR@95 lower. Online detectors report the mean over five stream orders (SD at most 2.0 AUROC points except 7.8 on the covariate shift of PROTON+, full table with SDs in the supplement). [†] hierarchical prompts of Ju et al. Best per column bold, second best underlined.

Method	Semantic			Covariate			Far-OOD		
	AUC↑	APR↑	FPR↓	AUC↑	APR↑	FPR↓	AUC↑	APR↑	FPR↓
Max-Logits	54.0	54.1	<u>79.3</u>	27.9	23.3	88.0	83.9	76.4	63.0
Energy	51.8	51.5	79.7	27.2	23.2	88.0	83.9	<u>76.5</u>	62.7
GL-MCM	75.8	75.6	86.0	68.9	45.3	69.3	78.5	62.4	75.0
MCM	<u>75.8</u>	<u>75.6</u>	86.0	68.9	45.2	69.7	78.5	62.4	75.0
TAG-MCM	68.4	70.5	88.7	30.1	23.8	93.7	72.2	47.2	60.3
MCM (L=1) [†]	66.5	65.0	84.7	<u>84.2</u>	66.3	<u>47.3</u>	51.8	38.4	98.3
MCM (L=5) [†]	58.3	56.5	87.0	66.5	42.5	65.7	54.4	33.5	74.3
PROTON [2]	68.0	72.5	92.7	46.7	29.4	83.6	<u>86.8</u>	76.2	55.0
PROTON+ (ours)	67.3	69.0	75.8	85.3	<u>63.9</u>	31.5	93.4	81.8	15.4
Δ PROTON+ vs. MCM	-8.4	-6.5	-10.2	+16.4	+18.6	-38.1	+14.9	+19.4	-59.6

Table 2. PROTON+ improves over MCM on every covariate and far shift and on every stream order. AUROC for the three benchmarks, mean $\pm$ SD over five stream orders (AUPR and FPR@95 in the supplement).

Benchmark (VLM)	Method	Semantic	Covariate	Far-OOD
Ophthalmology (FLAIR)	MCM	75.8	68.9	78.5
	PROTON	68.0 $\pm$ 0.9	46.7 $\pm$ 1.8	86.8 $\pm$ 3.6
	PROTON+ (ours)	67.3 $\pm$ 0.7	85.3 $\pm$ 7.8	93.4 $\pm$ 0.5
Radiology (UniMed-CLIP)	MCM	54.2	46.6	27.6
	PROTON	55.3 $\pm$ 0.9	61.2 $\pm$ 1.7	93.9 $\pm$ 2.3
	PROTON+ (ours)	61.4 $\pm$ 0.7	93.9 $\pm$ 0.6	91.5 $\pm$ 1.4
Pathology (QuiltNet)	MCM	59.8	66.1	67.1
	PROTON	64.9 $\pm$ 1.3	73.0 $\pm$ 3.7	75.2 $\pm$ 3.7
	PROTON+ (ours)	82.4 $\pm$ 0.3	85.6 $\pm$ 0.9	92.1 $\pm$ 0.7

the online detectors build on. **PROTON improves on far shifts in every modality and on covariate shift in radiology and pathology, but loses 22.1 AUROC points to MCM on the ultra-widefield covariate shift.** Its variance-driven weight stays between 0.42 and 0.65 on every stream, so the adaptive fusion is a constant in practice, and the gate costs rather than helps in radiology, where the bank without a gate scores 6.5 points higher on covariate shift. **PROTON+ beats MCM on the covariate and far shifts of every modality and on all five orders of every such stream, by 16.4 to 47.3 AUROC points on covariate shift and by 14.9 to 63.9 points on far shift.** MCM is below chance on both radiology shifts, so against the best static score of each split the margins are 1.1, 15.5 and 18.4 points on covariate shift and 9.4, 7.0 and 24.9 points on far shift. On semantic shift PROTON+ gains 7.2 points over MCM in radiology and 22.5 points in pathology, and on the ophthalmology semantic split, where glaucoma and macular-degeneration images lie inside the ID imaging domain and MCM already separates them at 75.8, the softmax

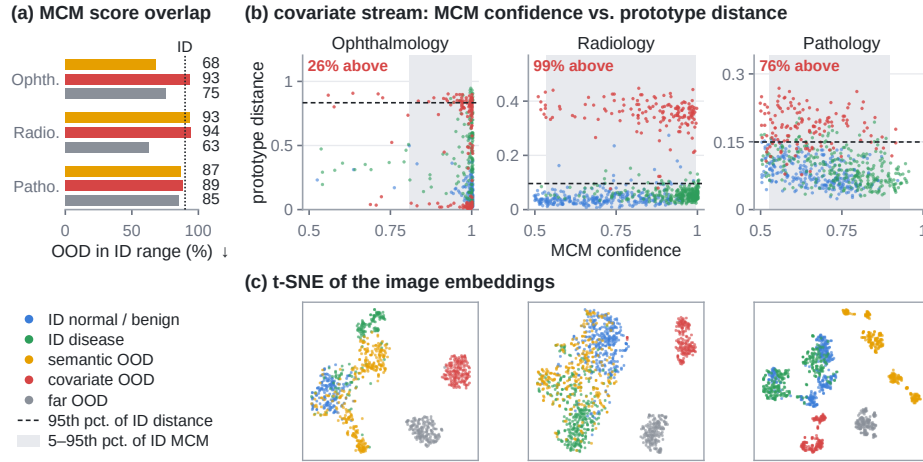


Fig. 2. Shifted inputs fall inside the ID range of the softmax score but form separate clusters in embedding space. (a) Share of OOD MCM scores inside the 5th to 95th percentile range of the ID scores (dotted line: the ID scores themselves, 90% by construction). (b) MCM confidence against the PROTON+ prototype distance on the covariate stream after calibration. The dashed line is the 95th percentile of the ID distance, the grey band spans the 5th to 95th percentile of ID MCM confidence, and the red label gives the share of covariate samples above the line. (c) t-SNE of all image embeddings. Columns: FLAIR, UniMed-CLIP, QuiltNet.

remains the signal to trust and the fused score stays at 67.3. FPR@95 on the ultra-widefield shift drops from 69.7 (MCM) and 83.6 (PROTON) to 31.5, and on the paediatric shift from 87.3 and 67.6 to 26.7. The AUROC standard deviation over orders is at most 2.0 points except on the ultra-widefield shift (7.8), where the 11 to 21 samples before calibration decide how much of the shifted cluster enters before the anchor forms. The pathology gains hold under the other prompt wordings: with the literal class names PROTON+ reaches 75.8, 72.9 and 92.5 against 64.3, 53.4 and 72.5 for MCM. Figure 2 shows why the bank path sees what the softmax cannot: covariate and far samples form clusters away from the ID classes in every modality, while most of their MCM scores fall inside the ID range.

3.2 Which change matters?

Table 3 adds the three changes one at a time for the three modalities. Z-scoring alone lifts covariate shift by 4 to 17 points, because the two raw scores have different variances and the weight cannot compensate for that. Consistency admission cannot repair a bank whose seed is already mixed, and on the ultra-widefield shift it lowers the AUROC to 40.1. **Domain-anchored prototypes carry the covariate gain, from 40.1 to 85.3 on the ultra-widefield shift and from 78.4 to 93.9 on the paediatric shift.** The prototype distance alone, from the

Table 3. Domain anchoring carries the covariate gain, and the prototype distance alone is the strongest signal on covariate and far shifts. AUROC, mean over five stream orders (SD at most 2.9 points except on the ultra-widefield shift, up to 12.6, with every SD in the supplement). The three changes are added cumulatively, then variants of PROTON+. Oracle admission never admits a shifted sample.

Variant	Ophthalmology			Radiology			Pathology		
	Sem	Cov	Far	Sem	Cov	Far	Sem	Cov	Far
MCM	75.8	68.9	78.5	54.2	46.6	27.6	59.8	66.1	67.1
PROTON	68.0	46.7	86.8	55.3	61.2	93.9	64.9	73.0	75.2
+ z-scored fusion	69.4	50.7	89.0	61.6	78.1	90.5	77.7	84.7	82.8
+ consistency admission	70.6	40.1	92.6	61.9	78.4	92.8	82.1	85.6	88.2
+ domain anchoring (= PROTON+)	67.3	85.3	93.4	61.4	93.9	91.5	82.4	85.6	92.1
S_{proto} alone, PROTON+ bank	57.0	90.4	97.9	62.6	98.8	96.2	94.9	93.7	98.7
PROTON+, oracle admission	73.6	90.4	93.5	64.9	92.4	91.3	82.1	85.8	91.4
PROTON+, max-fusion	65.7	86.6	93.2	56.0	92.2	90.8	84.8	87.9	95.2
PROTON+, rejection-driven α	64.4	86.9	93.6	60.8	95.8	93.8	88.3	88.0	96.6

PROTON+ bank, is the strongest single signal on covariate and far shifts (90.4, 98.8 and 93.7 on the three covariate shifts, 96 to 99 on the far shifts) and beats the fused score in pathology on every shift. Fusion with MCM pays only on the ophthalmology semantic shift (57.0 alone against 67.3 fused). Oracle admission is a reference rather than a bound, because it calibrates later and PROTON+ exceeds it on four pairs by up to 1.5 points, and PROTON+ stays within 6.3 points of it on every pair. The maximum of the two z-scores changes the nine pairs by -5 to $+3$ points, and the rejection-driven weight beats the variance-driven one on seven of nine pairs by up to 6 points while losing 2.9 points on the ophthalmology semantic shift.

3.3 Sensitivity, stream composition and cost

Figure 3 varies the ID fraction of the stream by sub-sampling the smaller pool, sweeps the gate, and measures the bank. **Both detectors assume an ID majority: PROTON+ beats MCM on covariate and far shifts at ID fractions of 0.7 and 0.9 in every modality, and inverts once the shifted cluster dominates the bank.** The inversion starts at 0.5 on the ophthalmology shifts and at 0.3 elsewhere, where the ultra-widefield AUROC of PROTON+ is 22.3 at 0.5 against 69 for MCM, which is order-free. PROTON+ moves by at most 3.3 points for $\gamma \in [0.5, 0.7]$ and 2.7 points for $\kappa \in [1, 4]$, and a stricter gate starves the bank where the ID images are less confident: $\gamma = 0.9$ drops the radiology covariate and far shifts by 16 and 20 points and the pathology shifts to the MCM level. The bandwidth $h \leq 0.1$ works in every modality, $h = 0.2$ loses 13 points on the ultra-widefield shift, and $h = 0.5$ no longer separates that cluster from the ID mode. Prompt wording moves the static baseline more than any detector choice: a trailing period in the FLAIR prompt moves the MCM covariate AUROC from 68.9 to 84.2 and the far AUROC from 78.5 to 51.8, while PROTON+ stays between 85.3 and 94.0 on covariate and 88.1 and 97.0 on far

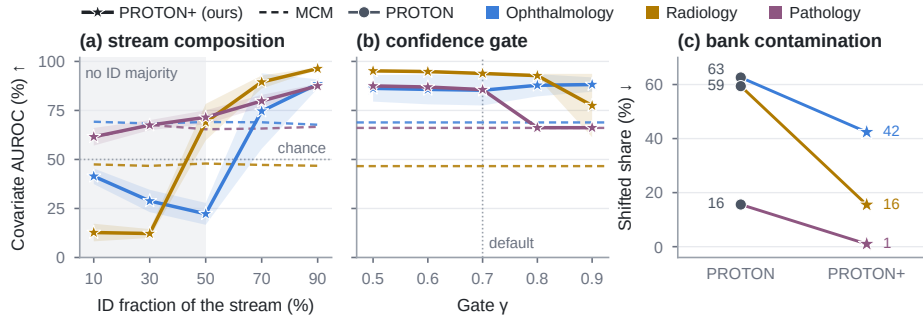


Fig. 3. PROTON+ beats MCM whenever ID images are the stream majority, holds across gates, and keeps the bank clean. (a) Covariate AUROC against the ID fraction of the stream (solid with stars: PROTON+, dashed: MCM; shaded: no ID majority). (b) Covariate AUROC of PROTON+ against the gate γ (dashed: MCM, dotted: default). (c) Shifted share of the most contaminated class bank on the covariate stream, PROTON against PROTON+. Means over five stream orders, bands $\pm$ SD.

shift across the three single-prompt wordings. With Max-Logits or Energy as the base score, which score 27 to 28 on the ultra-widefield shift, PROTON+ reaches 87 on covariate and 99 on far shift. The bank of PROTON+ ends with 42%, 16% and 1% shifted members on the three covariate streams against 63%, 59% and 16% for PROTON. One PROTON step takes 0.02 ms and one PROTON+ step 0.10 to 0.28 ms on a single CPU core in NumPy, with 0.41 MB of state, and calibration takes 14 to 20 samples in ophthalmology, 30 to 74 in radiology and 51 to 116 in pathology. Five of the six PROTON prototypes stay above a cosine of 0.99 to their final value after at most 29 admissions, the contaminated retinopathy prototype only after 123, and PROTON+ raises the cosine of the most contaminated prototype to the clean class mean from 0.57 to 0.84 in ophthalmology and from 0.88 to 0.99 in radiology (Figure 4). Over a grid of gates and bank sizes, PROTON+ beats PROTON in all 168 settings and MCM in 165 of them, while PROTON stays below MCM in all 56 ophthalmology settings (Figure 5).

4 Conclusion

PROTON+ is the recommended successor of PROTON: the same frozen VLM, the same bank and the same hyper-parameters, with prototypes anchored to the dominant imaging-domain mode, admission gated by prototype consistency and fusion on z-scores. It improves over MCM on the covariate and far shifts of three modalities, by up to 47.3 and 63.9 AUROC points over MCM and by up to 15.5 and 24.9 points over the best static score of a split, at 0.1 to 0.3 ms per sample. The prototype distance alone is the strongest detector on those shifts, and the fusion with the softmax score keeps the semantic shifts covered where the softmax carries information the embedding lacks. Three directions

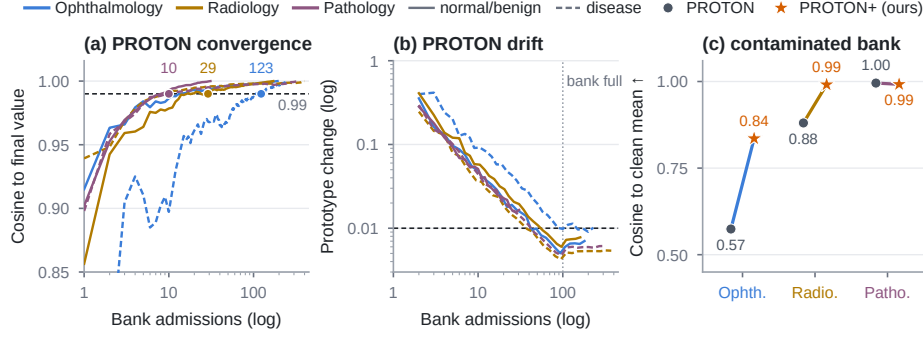


Fig. 4. Prototypes settle within thirty admissions in five of six banks, and PROTON+ keeps the most contaminated bank close to the clean class. (a) Cosine of the PROTON prototype after k admissions to its final value (solid: normal or benign class, dashed: disease class), with a dot where both prototypes of a modality stay above 0.99. (b) Change of the PROTON prototype per admission, with 0.01 dashed and the full bank ($M = 100$) dotted. (c) Cosine of the prototype of the most contaminated bank to the class mean of the labelled ID split, averaged after ten admissions, PROTON (circle) and PROTON+ (star). Covariate streams, means over five stream orders.

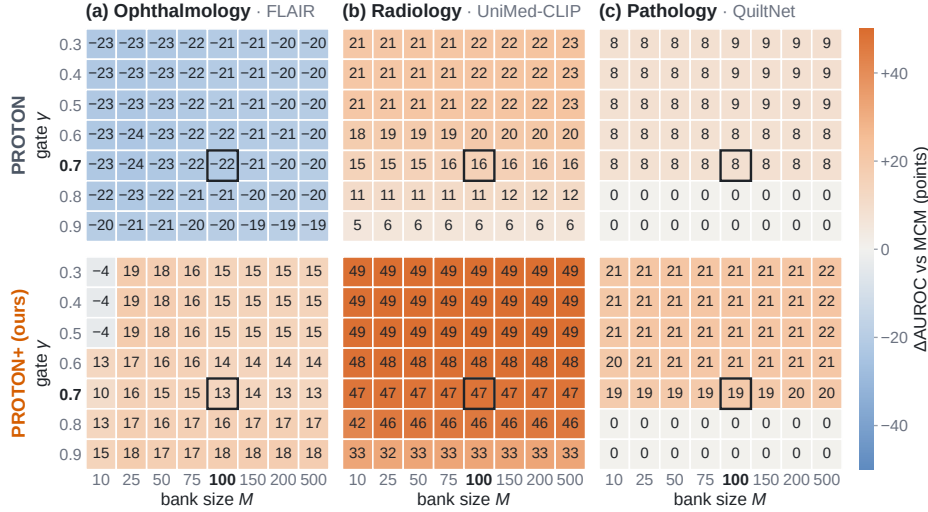


Fig. 5. PROTON+ beats PROTON in all 168 gate and bank-size settings and stays above MCM in all but three of them. Change in covariate AUROC over MCM (points, three stream orders) for each gate γ and per-class bank size M , PROTON (top) and PROTON+ (bottom) on one colour scale, with the defaults $\gamma = 0.7$ and $M = 100$ boxed. At the defaults PROTON+ gains 13.5, 47.4 and 19.5 points against -21.6 , 15.8 and 8.2 for PROTON. PROTON+ drops below MCM only with the smallest ophthalmology bank ($M = 10$, $\gamma \leq 0.5$), and $\gamma \geq 0.8$ starves the pathology bank. Rows with $\gamma \leq 0.5$ coincide because with two classes the top probability is never below 0.5.

follow. Choosing the fusion weight from the stream without labels is open, since neither the variance of MCM nor the rejection rate of the admission rule recovers that semantic loss. Streams with an ID minority invert both detectors once the shifted cluster dominates the bank. A comparison with test-time prompt learners is made direct by the released streams and features. The library, the seeded manifests, the cached features and a browser demo with the real FLAIR weights let any VLM and any folder of images be evaluated with three commands.

References

1. Cheng, Y.: Mean shift, mode seeking, and clustering. *IEEE TPAMI* **17**(8), 790–799 (1995)
2. Das, A., Wasalathilaka, N., Lu, Y., Dukre, A., Mahapatra, D., Khan, S., Razzak, I.: Proton: Prototype-based test-time online ood detection for medical vlms. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 527–537. Springer (2026)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *CVPR* (2009)
4. Fu, Y., Wan, H., Luo, X., Deng, M.: STAR: Test-time adaptation can enhance universal prompt learning for vision-language models. In: *CVPR* (2026)
5. Gutbrod, M., Rauber, D., Nunes, D.W., Palm, C.: OpenMIBOOD: Open medical imaging benchmarks for out-of-distribution detection. In: *CVPR*. pp. 25874–25886 (2025)
6. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: *ICML* (2022)
7. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1M: One million image-text pairs for histopathology. In: *NeurIPS*. vol. 36, pp. 37995–38017 (2023)
8. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data* **9**(1), 475 (2022)
9. Ju, L., Zhou, S., Zhou, Y., Lu, H., Zhu, Z., Keane, P.A., Ge, Z.: Delving into out-of-distribution detection with medical vision-language models. In: *MICCAI*. pp. 133–143 (2025)
10. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine* **16**(1), e1002730 (2019)
11. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**(5), 1122–1131 (2018)
12. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: UniMed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372* (2024)
13. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: WILDS: A benchmark of in-the-wild distribution shifts. In: *ICML* (2021)

14. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns* **3**(6), 100512 (2022)
15. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: *NeurIPS*. vol. 33, pp. 21464–21475 (2020)
16. Liu, X., Zach, C.: TAG: Text prompt augmentation for zero-shot out-of-distribution detection. In: *ECCV*. pp. 237–254 (2024)
17. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In: *NeurIPS*. vol. 35, pp. 35087–35102 (2022)
18. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: GL-MCM: Global and local maximum concept matching for zero-shot out-of-distribution detection. *IJCV* **133**(6), 3586–3596 (2025)
19. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence* **1**(1), e180041 (2019)
20. Silva-Rodríguez, J., Chakor, H., Kobbi, R., Dolz, J., Ben Ayed, I.: A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision. *Medical Image Analysis* **99**, 103357 (2025)
21. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology. In: *MICCAI* (2018)
22. Welford, B.P.: Note on a method for calculating corrected sums of squares and products. In: *Technometrics*. vol. 4, pp. 419–420 (1962)
23. Yang, Y., Feng, J., Xu, J., Wang, X., Gu, Q., Ye, N.: TINS: Test-time ID-prototype-separated negative semantics learning for OOD detection. *arXiv preprint arXiv:2605.10756* (2026)
24. Ye, J., Liao, J., Lai, R., Lu, X., Zhuang, J., Gan, Z., Wang, R.: TTL: Test-time textual learning for OOD detection with pretrained vision-language models. *arXiv preprint arXiv:2604.15756* (2026)

A Supplementary material

A.1 Benchmark construction

FIVES holds 800 fundus photographs, 200 per class, of which 150 per ID class enter the stream and the remaining 50 form the labelled split. The 150 ultra-widefield images are drawn from the 256 photographs of the three DeepDRiD ultra-widefield folders, and ImageNet images are 150 validation images of the ten-class Imagenette subset at full resolution. RSNA radiographs are converted from DICOM to 8-bit PNG with per-image min-max scaling, and 600 per class are drawn, 300 for the stream and 300 for the labelled split. The radiology semantic split is drawn from the public 5,606-image sample of NIH ChestX-ray14 after removing images whose findings include No Finding, Pneumonia, Infiltration, Consolidation, Effusion or Edema, and after removing every candidate whose 64 by 64 grayscale correlation with an RSNA image of the ID or labelled split exceeds 0.95 (22 of 1,834 candidates), the level unrelated RSNA radiographs reach among themselves. PatchCamelyon patches are 96 by 96 pixels, 600 per label from the test split (300 stream, 300 labelled). The semantic split draws 60 patches from each of adipose, debris, mucus, smooth muscle and cancer-associated stroma in NCT-CRC-HE-7K, excluding normal mucosa, lymphocytes, tumour epithelium and background. The covariate split draws 75 patches per label from the out-of-distribution test hospital of Camelyon17-WILDS.

A.2 Full tables

Table 4. Standard deviation over the five stream orders of every metric of the online detectors in Tables 1, 2, 6 and 7.

Benchmark	Method	Semantic			Covariate			Far-OOD		
		AUROC	AUPR	FPR	AUROC	AUPR	FPR	AUROC	AUPR	FPR
Ophth.	PROTON	0.9	0.7	2.3	1.8	0.8	6.1	3.6	2.6	26.0
	PROTON+	0.7	0.3	1.7	7.8	6.8	26.5	0.5	1.2	1.8
Radio.	PROTON	0.9	0.5	1.2	1.7	1.4	3.7	2.3	2.1	33.5
	PROTON+	0.7	0.5	1.4	0.6	3.4	5.4	1.4	4.4	31.5
Patho.	PROTON	1.3	1.3	1.2	3.7	2.5	8.7	3.7	5.9	12.5
	PROTON+	0.3	0.8	5.4	0.9	2.5	4.1	0.7	2.1	4.3

A.3 Prompt sensitivity

“vanilla” is the bare class name, “l1” the modality template of the main tables (for FLAIR the built-in caption with a trailing period), “id_acc” the pathology wording with the best zero-shot accuracy on the disjoint labelled split (“normal” and “cancer”), and “l5” the five-description hierarchy of Ju et al. for ophthalmology, written in the same style for the other two modalities. Under the five-description

Table 5. Standard deviation over the five stream orders of the AUROC values in Table 3.

Variant	Ophthalmology			Radiology			Pathology		
	Sem	Cov	Far	Sem	Cov	Far	Sem	Cov	Far
PROTON	0.9	1.8	3.6	0.9	1.7	2.3	1.3	3.7	3.7
+ z-scored fusion	1.1	0.4	1.5	0.8	0.8	1.3	0.6	0.8	2.5
+ consistency admission	2.9	2.8	1.7	0.7	12.1	1.4	0.7	1.0	7.9
+ domain anchoring (= PROTON+)	0.7	7.8	0.5	0.7	0.6	1.4	0.3	0.9	0.7
S_{proto} alone, PROTON+ bank	1.0	12.6	0.8	0.5	0.5	2.2	0.7	0.9	0.4
PROTON+, oracle admission	2.9	1.1	0.6	0.9	1.5	1.6	0.6	0.8	1.1
PROTON+, max-fusion	0.6	6.5	0.7	0.7	1.2	1.9	0.5	1.5	1.4
PROTON+, rejection-driven α	2.4	9.0	1.9	0.8	2.3	2.1	3.5	3.5	1.7

Table 6. Zero-shot OOD detection with UniMed-CLIP on the radiology benchmark (means over five orders).

Method	Semantic			Covariate			Far-OOD		
	AUC $\uparrow$	APR $\uparrow$	FPR $\downarrow$	AUC $\uparrow$	APR $\uparrow$	FPR $\downarrow$	AUC $\uparrow$	APR $\uparrow$	FPR $\downarrow$
Max-Logits	69.4	56.1	82.3	0.4	10.8	100.0	75.7	64.6	90.2
Energy	70.2	57.4	84.8	0.4	10.8	100.0	77.0	67.4	90.5
GL-MCM	54.3	35.7	89.7	46.7	18.1	87.3	29.2	14.2	98.0
MCM	54.2	35.2	89.7	46.6	18.1	87.3	27.6	14.0	99.7
TAG-MCM	57.4	36.8	87.8	48.1	19.3	85.5	41.3	16.0	87.5
MCM (L=1) †	54.2	35.2	89.7	46.6	18.1	87.3	27.6	14.0	99.7
MCM (L=5) †	54.2	32.0	77.5	<u>78.4</u>	<u>38.2</u>	<u>46.3</u>	84.6	52.3	39.5
PROTON [2]	55.3	38.3	95.2	61.2	31.0	67.6	93.9	92.5	33.5
PROTON+ (ours)	61.4	42.4	88.1	93.9	82.1	26.7	<u>91.5</u>	<u>74.6</u>	<u>35.7</u>
Δ PROTON+ vs. MCM	+7.2	+7.1	-1.5	+47.3	+64.1	-60.7	+63.9	+60.6	-64.0

prompts the ID accuracy falls to 57% in ophthalmology, the bank fills with mis-routed samples, and both PROTON and PROTON+ lose most of their margin.

A.4 Base scores, broader baselines and sensitivity

The privileged detectors separate every covariate and far split at 99 to 100 AUROC. The reference variance σ_0^2 of the fusion weight is not inert: sweeping it from 0.005 to 0.08 moves the mean AUROC of PROTON by 2 points in ophthalmology, rising with σ_0^2 , and by 14 points in radiology and 7 points in pathology, falling with σ_0^2 . A three-way triage of the shift type from the pair of scores reaches a macro-F1 of 0.60, 0.85 and 0.32 in the three modalities with a logistic regression under five-fold cross-validation.

A.5 Text encoding in FLAIR

The text tower of FLAIR averages the hidden states of three BERT layers over all token positions of the padded batch, padding included, so the embedding of “A fundus photograph of normal” changes with the longest prompt in its batch. The

Table 7. Zero-shot OOD detection with **QuiltNet** on the pathology benchmark (means over five orders).

Method	Semantic			Covariate			Far-OOD		
	AUC $\uparrow$	APR $\uparrow$	FPR $\downarrow$	AUC $\uparrow$	APR $\uparrow$	FPR $\downarrow$	AUC $\uparrow$	APR $\uparrow$	FPR $\downarrow$
Max-Logits	2.1	19.1	100.0	21.3	12.4	97.7	25.0	15.3	99.3
Energy	1.7	19.1	100.0	19.4	12.2	99.0	23.3	15.0	99.7
GL-MCM	56.7	40.5	94.0	67.2	30.9	77.7	67.2	29.1	75.2
MCM	59.8	42.4	90.3	66.1	29.8	85.2	67.1	31.5	77.2
TAG-MCM	48.7	33.1	97.2	48.0	19.4	92.0	45.1	18.7	95.0
MCM (L=1) †	59.8	42.4	90.3	66.1	29.8	85.2	67.1	31.5	77.2
MCM (L=5) †	15.3	20.4	99.8	64.6	30.7	77.3	61.1	26.3	87.3
PROTON [2]	64.9	54.9	92.8	73.0	47.3	87.2	75.2	50.4	80.4
PROTON+ (ours)	82.4	72.5	67.7	85.6	64.6	50.7	92.1	81.6	35.0
Δ PROTON+ vs. MCM	+22.5	+30.2	-22.6	+19.5	+34.8	-34.5	+25.0	+50.1	-42.2

Table 8. Prompt sensitivity, ophthalmology (AUROC Sem / Cov / Far).

Prompt set	ID acc.	MCM (Sem/Cov/Far)			PROTON			PROTON+		
flair_default	84.0	75.8	68.9	78.5	68.0	46.7	86.8	67.3	85.3	93.4
vanilla	86.0	65.6	64.9	83.5	58.5	46.7	94.4	58.2	87.9	97.0
l1	83.7	66.5	84.2	51.8	65.0	66.6	78.2	64.6	94.0	88.1
l5	57.3	58.3	66.5	54.4	41.2	24.0	25.3	56.2	50.8	58.8

released forward pass encodes all class prompts in one batch and the benchmark notebooks of Ju et al. call it, so the main protocol encodes the class prompts jointly. Encoding each prompt alone moves the MCM covariate AUROC by 7 points with the default caption and by 22 points with the bare class names.

Table 9. Prompt sensitivity, radiology.

Prompt set	ID acc.	MCM (Sem/Cov/Far)			PROTON			PROTON+		
vanilla	73.2	41.8	57.5	55.5	44.9	78.9	94.7	53.2	95.5	96.2
l1	69.2	54.2	46.6	27.6	55.3	61.2	93.9	61.4	93.9	91.5
l5	55.3	54.2	78.4	84.6	50.3	79.1	88.3	49.9	88.5	96.8

Table 10. Prompt sensitivity, pathology.

Prompt set	ID acc.	MCM (Sem/Cov/Far)			PROTON			PROTON+		
l1	59.3	59.8	66.1	67.1	64.9	73.0	75.2	82.4	85.6	92.1
id_acc	72.7	48.8	48.0	45.1	52.9	55.6	58.2	72.5	70.1	81.1
vanilla	64.5	64.3	53.4	72.5	66.7	58.9	82.2	75.8	72.9	92.5
l5	65.5	15.3	64.6	61.1	32.6	61.5	65.7	15.4	73.4	81.5

Table 11. Base scores alone, with PROTON and with PROTON+ (AUROC Sem / Cov / Far), ophthalmology, radiology and pathology from top to bottom.

	Base			+PROTON			+PROTON+		
Base score	Sem	Cov	Far	Sem	Cov	Far	Sem	Cov	Far
Max-Logits	54.0	27.9	83.9	57.3	23.6	87.0	57.1	87.1	99.1
Energy	51.8	27.2	83.9	55.3	22.6	87.0	56.1	87.2	99.1
GL-MCM	75.8	68.9	78.5	67.6	45.7	85.0	67.3	85.3	93.4
MCM	75.8	68.9	78.5	68.0	46.7	86.8	67.3	85.3	93.4
TAG-MCM	68.4	30.1	72.2	67.9	22.6	89.4	65.2	58.7	91.5

	Base			+PROTON			+PROTON+		
Base score	Sem	Cov	Far	Sem	Cov	Far	Sem	Cov	Far
Max-Logits	69.4	0.4	75.7	67.1	4.5	80.6	69.1	92.9	98.8
Energy	70.2	0.4	77.0	67.8	4.5	81.8	69.2	92.9	98.9
GL-MCM	54.3	46.7	29.2	55.1	60.6	93.5	61.4	93.9	91.6
MCM	54.2	46.6	27.6	55.3	61.2	93.9	61.4	93.9	91.5
TAG-MCM	57.4	48.1	41.3	58.2	60.6	94.2	63.6	86.8	83.9

	Base			+PROTON			+PROTON+		
Base score	Sem	Cov	Far	Sem	Cov	Far	Sem	Cov	Far
Max-Logits	2.1	21.3	25.0	12.7	29.7	31.4	75.5	79.4	92.5
Energy	1.7	19.4	23.3	12.4	28.1	29.8	75.5	78.8	92.2
GL-MCM	56.7	67.2	67.2	62.9	73.8	75.4	83.5	88.0	94.8
MCM	59.8	66.1	67.1	64.9	73.0	75.2	82.4	85.6	92.1
TAG-MCM	48.7	48.0	45.1	58.4	61.0	59.5	76.6	74.3	81.4

Table 12. Broader baselines (AUROC Sem / Cov / Far). [†] uses labelled ID features.

Ophthalmology			
Method	Sem	Cov	Far
TAG-MCM	68.4	30.1	72.2
MCM (L=5) [†]	58.3	66.5	54.4
kNN-ID [†]	76.8	100.0	99.8
Mahalanobis-ID [†]	82.4	100.0	100.0
PROTON	68.0	46.7	86.8
MCM (L=5) + PROTON	41.2	24.0	25.3
PROTON+ (ours)	67.3	85.3	93.4

Radiology			
Method	Sem	Cov	Far
TAG-MCM	57.4	48.1	41.3
MCM (L=5) [†]	54.2	78.4	84.6
kNN-ID [†]	67.8	99.9	100.0
Mahalanobis-ID [†]	58.9	100.0	100.0
PROTON	55.3	61.2	93.9
MCM (L=5) + PROTON	50.3	79.1	88.3
PROTON+ (ours)	61.4	93.9	91.5

Pathology			
Method	Sem	Cov	Far
TAG-MCM	48.7	48.0	45.1
MCM (L=5) [†]	15.3	64.6	61.1
kNN-ID [†]	100.0	99.4	100.0
Mahalanobis-ID [†]	100.0	99.8	100.0
PROTON	64.9	73.0	75.2
MCM (L=5) + PROTON	32.6	61.5	65.7
PROTON+ (ours)	82.4	85.6	92.1

Table 13. Bandwidth h of the domain anchor (PROTON+ AUROC).

h	Ophthalmology			Radiology			Pathology		
	Sem	Cov	Far	Sem	Cov	Far	Sem	Cov	Far
0.02	66.9	82.7	89.2	58.5	94.4	91.8	81.1	83.1	90.8
0.05	68.1	86.9	92.5	60.7	94.1	91.6	82.2	85.2	91.9
0.1	67.3	85.3	93.4	61.4	93.9	91.5	82.4	85.6	92.1
0.2	66.9	72.6	93.8	61.6	94.2	91.6	82.5	85.6	89.0
0.5	68.1	40.3	93.7	61.8	86.4	92.0	82.2	85.6	88.5

Table 14. PROTON under the gate γ and bank size M , ultra-widefield covariate shift.

γ	AUC	APR	F95	M	AUC	APR	F95
0.5	46.9	29.4	82.5	10	45.3	29.3	92.0
0.6	46.7	29.4	83.8	25	44.6	28.6	89.9
0.7	46.7	29.4	83.6	50	45.4	28.9	87.2
0.8	47.8	30.1	81.3	100	46.7	29.4	83.6
0.9	49.4	30.7	79.9	200	47.9	29.9	79.5