PROTON: Prototype-Based Test-Time Online OOD Detection for Medical VLMs

Abhijit Das[1], Nichula Wasalathilaka[2], Yifan Lu[1], Adinath Dukre[1], Dwarikanath Mahapatra[3], Shadab Khan[4], 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) enable zero-shot clinical image classification, yet reliably detecting out-of-distribution (OOD) inputs at deployment remains an open problem. No static scoring method works across all shift types: Maximum Concept Matching (MCM) on FLAIR achieves 76.4% AUROC for far-OOD but only 42.4% for covariate shifts such as ultra-wide-field fundus images —effectively random. We trace this to a structural mismatch: covariate-shifted inputs are indistinguishable from in-distribution samples in softmax space, yet occupy distinct regions in the VLM's embedding space. To exploit this untapped signal, we propose **PROTON** (**PRO**totype-based **T**est-time **ON**line OOD detection), a lightweight post-hoc module that maintains an online **prototype bank** from high-confidence test predictions and adaptively fuses prototype distance with MCM scoring via stream-level variance statistics, requiring no model modification, training data, or prompt engineering. On the ophthalmology benchmark (FLAIR + FIVES), PROTON improves MCM by **+23.9 AUROC on covariate shift**, +8.8 on semantic, and +8.1 on far-OOD —the only zero-shot method to improve all three without hierarchical prompts or labeled data. Code is online.[6]

Keywords: Out-of-Distribution Detection · Vision-Language Models · Test-Time Adaptation · Prototype Bank

1 Introduction

Medical vision-language models (VLMs) such as FLAIR [14], UniMedCLIP [4], and QuiltNet [2] enable zero-shot clinical classification, yet deployed systems inevitably encounter out-of-distribution (OOD) inputs—misrouted pathologies, novel imaging devices, or non-clinical submissions—risking silent misdiagnosis as VLM-based screening scales. Existing OOD detectors for VLMs are *static*: MCM [11], energy [9], and GL-MCM [13] apply a fixed scoring function regardless

[6] github.com/GenMI-Lab/PROTON, Also visit the project page: PROTON.

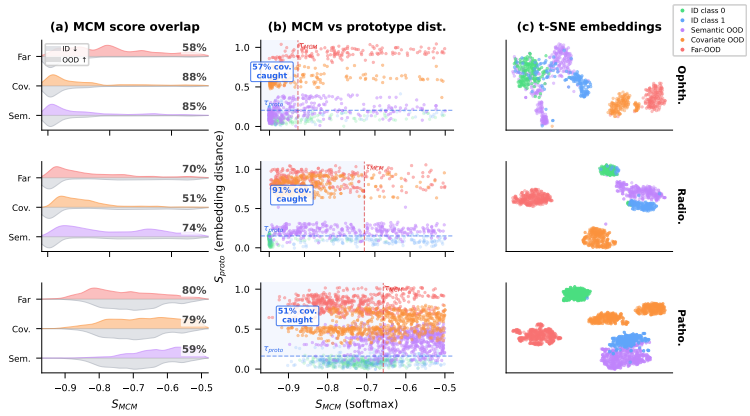

Fig. 1: **MCM's blind spot. (a)** MCM score overlap between ID and OOD (51–88% across domains). **(b)** Prototype distance (y-axis) separates covariate samples that MCM (x-axis) cannot; blue zone: 51–91% of covariate OOD caught by PROTON. **(c)** t-SNE confirms geometric separation that softmax collapses. Rows (Top to Bottom, in order): FLAIR, UniMedCLIP, QuiltNet.

of deployment context. Ju et al. [3] show that no static method works across all shift types on medical VLMs, a finding corroborated across 14 datasets by OpenMIBOOD [1]. Recent efforts address detection and adaptation separately: HVL [6] and GLAli [15] improve medical OOD detection but require labeled ID data; BCA+ [7], SCA [17], and UL-TTA [5] adapt VLMs at test time but assume a closed label set, aiming to *accept* shifted inputs rather than flag them; and OODD [8] applies test-time detection via a dynamic dictionary but requires ID training data and targets only standard classifiers on CIFAR or ImageNet.

The common thread across these failures is a missed signal: covariate-shifted samples produce MCM scores indistinguishable from ID yet occupy *distinct regions* in the VLM's embedding space (Fig. 1). Static methods collapse this geometry into a scalar via softmax, discarding exactly the separation that matters. Building on this insight, we propose **PROTON** (**PRO**totype-based **T**est-time **ON**line OOD detection), a lightweight post-hoc module atop a frozen VLM. PROTON maintains an **online prototype bank** of per-class centroids built from high-confidence test predictions, computes a *prototype distance score* via cosine similarity to the nearest centroid, and *adaptively fuses* this with MCM scoring using a weight derived from running MCM variance via *Welford's online algorithm*—recovering the covariate separation that softmax discards. Concretely, we make the following **contributions:**

1. We formalize the unexploited gap between test-time adaptation and static OOD detection in medical VLMs, and propose PROTON—to our knowledge, the first model-agnostic, prompt-free framework bridging both via an online prototype bank, requiring no training data or model modification.

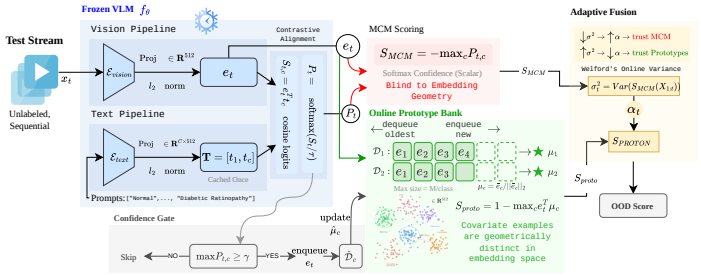

Fig. 2: **Overview of PROTON.** A frozen VLM produces embedding $\mathbf{e}_t$ and softmax probabilities $\mathbf{p}_t$ per test image. S_{MCM} scores softmax confidence; S_{proto} measures cosine distance to online class prototypes in per-class FIFO queues. Adaptive fusion weights both via MCM stream variance (α_t), and a confidence gate prevents OOD contamination of prototypes.

2. We introduce variance-driven adaptive fusion that reweights softmax and embedding-based scores via Welford's online algorithm, automatically up-weighting prototype distance when MCM becomes unreliable (covariate shift) while preserving its strength on far-OOD without any shift-type supervision.
3. PROTON is the only zero-shot method to improve all three shift types on the ophthalmology benchmark (largest gain on covariate, +**23.9** AUROC), and the effect is not MCM-specific—it lifts six base scores and transfers to radiology and pathology, robust to low ID stream fractions at ∼0.4 MB.

2 Method

Given the gap identified above, we now formalize the setting and describe PROTON's three components (Fig. 2): (1) an *online prototype bank* (Sec. 2.1), (2) a *prototype distance score* (Sec. 2.2), and (3) *adaptive fusion* (Sec. 2.3); it adds no learnable parameters and degrades gracefully to pure MCM before prototypes accumulate.

Let f_θ be a frozen medical VLM with vision encoder f_v and text encoder f_t, mapping inputs into a shared d-dimensional space (d=512 for FLAIR). Text embeddings $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_C] \in \mathbb{R}^{C \times d}$ for C in-distribution class names are computed once and cached. For each test image x_t, the vision encoder produces ℓ_2-normalized embedding $\mathbf{e}_t = f_v(x_t) \in \mathbb{R}^d$, yielding softmax probabilities $\mathbf{p}_t = \text{softmax}(\mathbf{e}_t^\top \mathbf{T}/\tau)$ and the MCM score [11] $S_{\text{MCM}}(x_t) = -\max_c\, p_{t,c}$ (higher = more OOD). Images arrive as an unlabeled stream; the detector scores each x_t using f_θ and past observations $x_1, \ldots, x_{t-1}$, with no training data, calibration labels, or weight updates permitted.

Algorithm 1 PROTON: Online OOD Detection at Test Time

Require: Frozen VLM f_θ, prompts $\{y_c\}_{c=1}^{C}$, hyperparams $\gamma, M, K_{\min}, \alpha_{\min}, \alpha_{\max}, \sigma_0^2$
$\mathbf{T} \leftarrow [f_t(y_1), \ldots, f_t(y_C)]$; $\mathcal{D}_c \leftarrow \emptyset\ \forall c$; $(n, \bar{S}, M_2) \leftarrow 0$ ▷ *cache text; init bank+Welford*
for each test image x_t in stream **do**
  $\mathbf{e}_t \leftarrow f_v(x_t)$; $\mathbf{p}_t \leftarrow \mathrm{softmax}(\mathbf{e}_t^\top \mathbf{T}/\tau)$; $S_{\mathrm{MCM}} \leftarrow -\max(\mathbf{p}_t)$
  update $(n, \bar{S}, M_2)$ via Welford $\Rightarrow \sigma_t^2$ ▷ *$O(1)$ stream variance*
  if $|\mathcal{D}_c| \geq K_{\min}\ \forall c$ **then** ▷ *calibrated*
    $S_{\mathrm{proto}} \leftarrow 1 - \max_c \mathbf{e}_t^\top \boldsymbol{\mu}_c$ ▷ *prototype distance*
    $\alpha_t \leftarrow \alpha_{\max} - \sigma\big(100(\sigma_t^2 - \sigma_0^2)\big)(\alpha_{\max} - \alpha_{\min})$
    $S_t \leftarrow \alpha_t S_{\mathrm{MCM}} + (1 - \alpha_t) S_{\mathrm{proto}}$ ▷ *adaptive fusion*
  else
    $S_t \leftarrow S_{\mathrm{MCM}}$ ▷ *pure-MCM fallback*
  if $\max(\mathbf{p}_t) \geq \gamma$ **then** ▷ *confidence gate*
    enqueue $\mathbf{e}_t \rightarrow \mathcal{D}_{\hat{c}}$ ($\hat{c} = \arg\max \mathbf{p}_t$), evict oldest if full; update $\boldsymbol{\mu}_{\hat{c}}$ (1)
  emit S_t ▷ *per-sample OOD score*

2.1 Online Prototype Bank

For each class $c \in \{1, \ldots, C\}$, the bank maintains a FIFO queue $\mathcal{D}_c$ of maximum size M. After scoring x_t, a *confidence gate* determines whether to update: if $\max_c p_{t,c} \geq \gamma$, embedding $\mathbf{e}_t$ is enqueued into $\mathcal{D}_{\hat{c}}$ ($\hat{c} = \arg\max_c p_{t,c}$); otherwise x_t is scored but excluded from the bank, preventing low-confidence —potentially OOD —embeddings from contaminating prototypes. When $|\mathcal{D}_c|$ exceeds M, the oldest entry is evicted so prototypes track recent deployment statistics.

The class prototype is the ℓ_2-normalized mean of stored embeddings:

$$\boldsymbol{\mu}_c = \frac{\bar{\mathbf{e}}_c}{\|\bar{\mathbf{e}}_c\|_2}, \quad \bar{\mathbf{e}}_c = \frac{1}{|\mathcal{D}_c|} \sum_{\mathbf{e} \in \mathcal{D}_c} \mathbf{e}. \tag{1}$$

Normalization ensures cosine similarity via dot product, consistent with the VLM's contrastive training. The detector is *calibrated* once every class accumulates $K_{\min}$ samples ($K_{\min}$=5); before this, PROTON defaults to MCM.

2.2 Prototype Distance Score

Once calibrated, PROTON scores each sample by its embedding-space proximity to the nearest prototype:

$$S_{\mathrm{proto}}(x_t) = 1 - \max_c \ \mathbf{e}_t^\top \boldsymbol{\mu}_c. \tag{2}$$

Since both vectors are ℓ_2-normalized, this equals one minus the maximum cosine similarity: low for ID samples near their prototype, high for OOD samples distant from all prototypes—capturing the embedding-space separation that MCM's softmax collapses.

2.3 Adaptive Score Fusion

PROTON fuses both signals into a single score:

$$S_{\mathrm{PROTON}}(x_t) = \alpha_t \cdot S_{\mathrm{MCM}}(x_t) + (1 - \alpha_t) \cdot S_{\mathrm{proto}}(x_t), \tag{3}$$

where $\alpha_t \in [\alpha_{\min}, \alpha_{\max}]$ adapts online rather than requiring per-deployment tuning. We track the running variance σ_t^2 of S_{MCM} via Welford's algorithm ($O(1)$ memory, exact single-pass) and set:

$$\alpha_t = \alpha_{\max} - \sigma\big(100 \cdot (\sigma_t^2 - \sigma_0^2)\big) \cdot (\alpha_{\max} - \alpha_{\min}), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid and σ_0^2 a reference threshold corresponding to nominal ID-only variance. Intuitively, a homogeneous (mostly-ID) stream yields low MCM variance, pushing $\alpha_t \to \alpha_{\max}$ (trust MCM); covariate samples produce confidently-wrong softmax that spikes variance, downweighting MCM in favor of S_{proto}. The sigmoid gives smooth, outlier-robust interpolation, and $\alpha_t = 1$ before calibration recovers pure MCM. Under default hyperparameters, PROTON consistently improves over MCM across all shift types (Table 2), though extreme gate settings (γ>0.9 or γ<0.5) can degrade performance, as shown in Fig. 3(b).
Complexity. PROTON adds negligible overhead: one dot product per class for S_{proto} ($O(Cd)$), one $O(1)$ queue operation, and three scalar Welford updates per sample. Memory is bounded by $C{\times}M{\times}d$ floats ($2{\times}100{\times}512 \approx 0.4$ MB for FLAIR). No gradients, augmentation, or backpropagation are required. Algorithm 1 summarizes this.

3 Experiments

Benchmark. We adopt the ophthalmology OOD benchmark of Ju et al. [3], built on FIVES: ID classes are Normal and Diabetic Retinopathy (150 each); OOD comprises Semantic (Glaucoma + AMD, N=300), Covariate (UWF fundus, N=150), and Far-OOD (ImageNet, N=150). We evaluate on FLAIR [14] (ResNet-50, d=512, Bio_ClinicalBERT) with frozen weights and vanilla class-name prompts.
Baselines. We compare against all zero-shot methods from Ju et al. [3] Table 2: Max-Logits, Energy [9], GL-MCM [13], MCM [11], MCM with hierarchical prompts (L=1, 5), and the prompt-augmentation method TAG-MCM [10], which like PROTON requires no training data, model modification, or prompt trees. To contextualize the upper bound, we additionally report two *privileged* feature-bank detectors that consume labeled ID features (like kNN-ID and Mahalanobis-ID) which PROTON is not entitled to use. We restrict the main comparison to zero-shot methods because PROTON operates without any labeled data; few-shot approaches (CoOp [16], LoCoOp [12]) occupy a strictly more privileged setting with access to ID training samples, making direct comparison inequitable. We exclude methods that need labeled ID data (HVL [6], GLAli [15]) or target accuracy rather than detection (BCA+ [7], SCA [17], UL-TTA [5]).
Metrics and hyperparameters. AUROC, AUPR, and FPR@95 per shift type [11,3]. PROTON hyperparameters are fixed across all experiments: γ=0.7, M=100, $K_{\min}$=5, $\alpha_{\min}$=0.3, $\alpha_{\max}$=0.7, σ_0^2=0.02. The reference variance σ_0^2 is *not* tuned per-VLM: it is read off the early-stream ID variance from a brief warm-up, and sweeping $\sigma_0^2 \in [0.005, 0.08]$ leaves mean AUROC flat (69.8–71.2) due to sigmoid saturation. We analyze sensitivity to γ and M in Sec. 4.

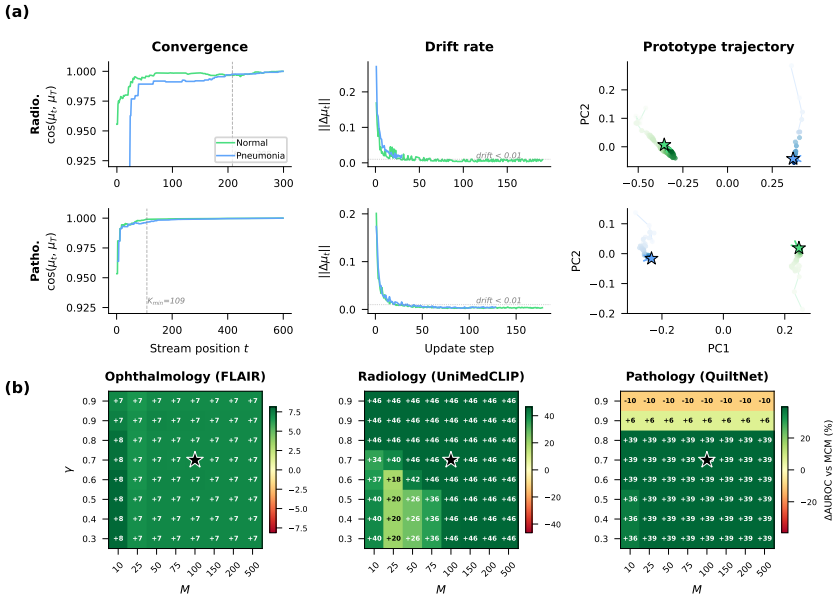

Fig. 3: **PROTON analysis. (a)** Prototype convergence (cosine similarity to final prototype, drift, and PCA trajectories; $\star$ = final). Dashed lines mark the stream index at which all classes reach $K_{\min}$. **(b)** $\gamma \times M$ sensitivity (ΔAUROC over MCM, covariate shift; $\star$ = default) across three modalities.

Table 1: **Robustness and cross-domain transfer. (a)** Per-shift AUROC vs. stream ID-fraction ρ. **(b)** Absolute covariate AUROC across three VLMs/modalities (Radiology: Normal/Pneumonia, RSNA, 300 ea.; Pathology: Benign/Malignant, PCam-derived, 300 ea.).

(a) Stream ID-fraction ρ				**(b)** Cross-domain (Cov. AUROC)			
ρ	Sem	Cov	Far	Domain (VLM)	MCM	PROTON	Δ
90%	63.2	67.4	84.8	Ophth. (FLAIR)	42.4	66.3	+23.9
70% (default)	62.7	66.3	84.5	Radio. (UniMedCLIP)	44.7	90.4	+45.7
30%	58.8	58.5	80.4	Patho. (QuiltNet)	45.3	84.2	+38.9
10%	55.1	50.9	77.2				
MCM (ref.)	53.9	42.4	76.4				

4 Results and Analysis

PROTON achieves 62.7 / 66.3 / 84.5% AUROC on semantic / covariate / far-OOD (Table 2), improving vanilla MCM by +8.8, +23.9, and +8.1 points respectively—the only zero-shot method to improve all three shifts without hierarchical prompts. The +23.9 covariate gain (42.4%→66.3%) confirms the core hypothesis: UWF images produce confident softmax yet lie far from ID prototypes in embedding space. At 84.5% far-OOD, PROTON exceeds even hierarchical MCM L=5 (82.4%), showing that prototype distance complements MCM even where it already performs well.

Table 2: Zero-shot OOD detection on FLAIR + FIVES. L: hierarchical prompt levels. Δ: gain over vanilla MCM. Best per column in **bold**, second-best underlined.

		Semantic			Covariate			Far-OOD		
	Method	AUC↑	APR↑	FPR↓	AUC↑	APR↑	FPR↓	AUC↑	APR↑	FPR↓
Static	Max-Logits	40.3	35.0	97.0	42.8	30.0	94.2	73.8	59.2	58.2
	Energy [9]	39.8	35.0	97.0	43.8	30.0	93.1	71.4	56.6	61.9
	GL-MCM [13]	52.1	47.6	92.4	43.5	30.0	93.4	74.1	59.5	57.8
	MCM [11]	53.9	49.7	90.1	42.4	30.0	94.6	76.4	61.9	54.2
Hr.	MCM (L=1)	61.6	58.4	80.4	65.6	48.2	70.0	52.2	36.3	91.6
	MCM (L=5)†	**66.7**	**63.6**	**74.0**	**87.7**	**70.4**	**46.7**	82.4	68.3	44.9
	PROTON (Ours)	62.7	59.6	79.0	66.3	49.0	69.3	**84.5**	**70.5**	**41.7**
	Δ vs MCM	▲+8.8	▲+9.9	▼11.1	▲+23.9	▲+19.0	▼25.3	▲+8.1	▲+8.6	▼12.5

†Requires domain-expert + ChatGPT-generated hierarchical prompt trees.

Table 3: Ablation on the ophthalmology benchmark. **(a)** Effect of each PROTON component (AUROC across all shift types). **(b)** Sensitivity to confidence threshold γ (covariate shift, M=100). **(c)** Sensitivity to bank size M (at γ=0.7).

(a) Components				**(b)** Threshold γ				**(c)** Bank size M			
Variant	Cov	Sem	Far	γ	AUC	APR	F95	M	AUC	APR	F95
MCM baseline	42.4	53.9	76.4	0.5	66.2	48.9	69.4	10	67.1	49.6	68.4
+ Proto. (α=.5)	57.9	59.2	80.9	0.6	66.3	49.0	69.3	25	66.1	48.9	69.5
+ Adapt. fusion	62.7	61.4	82.9	**0.7**	**66.3**	**49.0**	**69.3**	50	66.4	49.1	69.2
+ Conf. gate (**Full**)	**66.3**	**62.7**	**84.5**	0.8	66.2	48.9	69.4	**100**	**66.3**	**49.0**	**69.3**
				0.9	65.5	48.4	70.3	200	66.2	48.9	69.4

Why does PROTON work? Fig. 1(c) provides a direct answer: t-SNE projections show that ID classes form tight clusters while UWF covariate samples occupy geometrically distinct regions—despite producing nearly identical MCM scores (Fig. 1(a), 88% overlap). PROTON's prototype bank captures exactly this structure. Fig. 3(a) confirms that prototypes converge rapidly on the primary ophthalmology benchmark: cosine similarity to the final FLAIR centroids exceeds 0.99 within ∼55 samples and drift falls below 0.01 after ∼80 updates, requiring no dedicated calibration phase, with radiology and pathology following the same trajectory.

Adaptive fusion behavior. The variance-driven weight α_t responds to stream composition automatically: it drops sharply after covariate-shifted samples enter (MCM variance spikes, upweighting S_{proto}) but remains high for far-OOD where MCM is already discriminative—demonstrating shift-dependent adaptation from statistics alone, without any shift-type labels.

Robustness to stream composition and classifier accuracy. PROTON assumes a non-zero ID fraction to seed prototypes; we make this explicit and stress-test it. Table 1(a) varies the stream ID-fraction ρ: PROTON stays above MCM on all shifts down to ρ=10%, because the confidence gate plus the $K_{\min}$ pure-MCM fallback (Alg. 1) keep prototypes uncontaminated and recover the baseline in the worst case, so PROTON never degrades below MCM. Under forced misrouting, it remains above MCM through 30% label errors (65.5 vs.

57.6 mean covariate AUROC). Results are order-stable: over five stream-order permutations the per-shift AUROC is 62.7±1.1 / 66.3±1.8 / 84.5±0.9. Table 1(b) reports absolute covariate AUROC across three VLMs, replacing the earlier Δ-only heatmap: gains are largest on covariate shift everywhere (+23.9 / +45.7 / +38.9), and semantic and far-OOD also improve in all domains.

Ablations. Table 3 decomposes PROTON's gains. Prototype distance with fixed fusion (α=0.5) provides the largest single improvement (+15.5 covariate AUROC); adaptive fusion adds +4.8 by dynamically reweighting across shift types; and the confidence gate contributes a further +3.6 by preventing OOD contamination of prototypes. Hyperparameters are robust: performance is stable across $\gamma \in [0.5, 0.8]$ and $M \in [25, 200]$, degrading only at extremes (γ>0.9 starves the bank; γ<0.5 admits OOD noise); Fig. 3(b) shows this stability holds across the $\gamma \times M$ grid for radiology and pathology too.

Generalization beyond MCM. PROTON requires only a scalar base score and an embedding, so MCM is one instantiation rather than a dependency. Table 4(a) instantiates PROTON on six base scores: every one improves on covariate shift (mean +11–15 AUROC), with semantic and far-OOD following the same pattern, confirming that the gain stems from image-embedding geometry that complements any softmax/logit score. Table 4(b) further shows PROTON (66.3) matches privileged feature-bank detectors that consume labeled ID features (kNN-ID 70.8, Mahalanobis-ID 68.6) while using none, and composes with hierarchical prompts (Hier.MCM + PROTON reaches 88.4, exceeding either alone).

5 Discussion

Asymmetric gains and complementarity. The covariate gain dwarfs the semantic and far-OOD ones because it is the failure mode PROTON targets: UWF images depict the same pathologies as standard fundus (high softmax confidence that blinds MCM) yet differ visually—wider field, different illumination, peripheral structures—placing them in separable embedding regions (Fig. 1(c)) that early-stream prototypes capture. For semantic and far-OOD shifts MCM already separates well, so prototype distance acts as a complement rather than the primary signal.

Hierarchical MCM (L=5) achieves a higher covariate AUROC of 87.7% by enriching *what* the model represents through multi-level clinical text embeddings, whereas PROTON improves *how* OOD scores are derived from existing representations. These two strategies are complementary, not competing: combining them (Hier.MCM + PROTON) reaches 88.4% covariate AUROC, above either alone, and PROTON already exceeds Hier.MCM on far-OOD (84.5 vs. 82.4); unlike hierarchical prompting, PROTON requires no domain expertise or LLM-generated prompt trees and is deployable out of the box with vanilla class names.

Scope and future work. PROTON's closed-form fusion could be replaced by a learned gating mechanism for further gains; we defer this to preserve the

Table 4: **PROTON is not MCM-specific. (a)** Covariate AUROC of six base scores alone (Base) vs. as PROTON's base signal (+PROTON). **(b)** Covariate AUROC vs. broader baselines. ‡privileged (labeled ID features); †LLM-generated prompt trees.

(a) PROTON across base scores

Base score	Base	+PROTON	Δ
Max-Logits	42.8	62.5	+19.7
Energy	43.8	62.0	+18.2
GL-MCM	43.5	65.8	+22.3
MCM	42.4	66.3	+23.9
Entropy	44.1	63.1	+19.0
TAG-MCM	45.7	67.1	+21.4

(b) Broader baselines

Method	Cov. AUC
TAG-MCM [10]	45.7
kNN-ID‡	70.8
Mahalanobis-ID‡	68.6
Hier.MCM (L=5)†	87.7
PROTON (ours)	**66.3**
Hier.MCM + PROTON	**88.4**

parameter-free design. Beyond binary detection, the two-score decomposition already supports shift-type *triage*: classifying each sample as covariate/semantic/far-OOD directly from $(S_{\text{MCM}}, S_{\text{proto}})$ reaches macro-F1 = 0.70 on the 3-way task—high S_{proto} with low S_{MCM} flags covariate shift for re-acquisition, while high S_{MCM} routes to specialist review or rejection. We will extend this triage analysis to additional modalities beyond the three studied here.

6 Conclusion

We presented PROTON, the first test-time adaptive OOD detection framework for medical VLMs. By fusing online embedding-space prototypes with MCM via stream-level variance, it lifts chance-level covariate detection while also improving semantic and far-OOD —training-free, prompt-free, and parameter-free at 0.4 MB. More broadly, PROTON demonstrates that the test stream itself contains sufficient structure to improve detection through online prototypes, encouraging further work toward multi-domain validation, shift-type triage, and integration with hierarchical prompting.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gutbrod, M., Rauber, D., Nunes, D.W., Palm, C.: Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25874–25886 (2025)
2. 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. Advances in neural information processing systems **36**, 37995–38017 (2023)

3. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 133–143. Springer (2025)
4. 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)
5. Kim, B.: Ultra-light test-time adaptation for vision–language models. arXiv preprint arXiv:2511.09101 (2025)
6. Lai, R., Lu, X., Chen, K., Chen, Q., Zheng, W.S., Wang, R.: Hierarchical vision-language learning for medical out-of-distribution detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 230–239. Springer (2025)
7. Li, X., Li, J., Li, F., Zhu, L., Yang, Y., Shen, H.T.: Generalizing vision-language models to novel domains: A comprehensive survey. arXiv preprint arXiv:2506.18504 (2025)
8. Lin, L., Bai, Y., Su, H., Zhu, C., Wang, Y., Zhou, Y., Fu, H., Chen, J.: Oodbench: Out-of-distribution benchmark for large vision-language models. arXiv preprint arXiv:2602.18094 (2026)
9. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
10. Liu, X., Zach, C.: Tag: Text prompt augmentation for zero-shot out-of-distribution detection. In: European Conference on Computer Vision. pp. 237–254. Springer (2024)
11. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Advances in neural information processing systems **35**, 35087–35102 (2022)
12. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. In: Thirty-Seventh Conference on Neural Information Processing Systems (2023)
13. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection: A. miyai et al. International Journal of Computer Vision **133**(6), 3586–3596 (2025)
14. 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 (Jan 2025). https://doi.org/10.1016/j.media.2024.103357, http://dx.doi.org/10.1016/j.media.2024.103357
15. Yan, J., Guan, X., Zheng, W.S., Chen, H., Wang, R.: Global and Local Vision-Language Alignment for Few-Shot Learning and Few-Shot OOD Detection . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15964. Springer Nature Switzerland (September 2025)
16. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
17. Zhou, Y., Levine, S., Weston, J., Li, X., Sukhbaatar, S.: Self-challenging language model agents. arXiv preprint arXiv:2506.01716 (2025)