PromptGate: Client-Adaptive Vision–Language Gating for Open-Set Federated Active Learning

Adea Nesturi*, David Dueñas Gaviria*, Jiajun Zeng, and Shadi Albarqouni✉

University of Bonn, University Hospital Bonn, Clinic for Diagnostic and Interventional Radiology, 53127 Bonn, Germany

Abstract. Deploying medical AI across resource-constrained institutions demands data-efficient learning pipelines that respect patient privacy. Federated Learning (FL) enables collaborative medical AI without centralizing data. Yet, real-world clinical pools are inherently *open-set*: beyond the in-distribution (ID) target classes, the unlabeled pool also holds out-of-distribution (OOD), non-target samples such as imaging artifacts, background tissue, and other pathologies. Standard Active Learning (AL) query strategies mistake this noise for informative samples, wasting scarce annotation budgets. We propose **PromptGate**, a dynamic VLM-gated framework for Open-Set Federated AL that purifies unlabeled pools before querying. PromptGate introduces a federated Class-Specific Context Optimization: lightweight, learnable prompt vectors that adapt a frozen BiomedCLIP backbone to local clinical domains and aggregate globally via FedAvg—without sharing patient data. As new annotations arrive, prompts progressively sharpen the ID/OOD boundary, turning the VLM into a dynamic gatekeeper that is *strategy-agnostic*: a plug-and-play pre-selection module enhancing any downstream AL strategy. Experiments on distributed dermatology and breast imaging benchmarks, with a random acquisition rule, show that while static VLM prompting degrades to ∼60% ID purity, PromptGate maintains >95% purity with 98% OOD recall. The source code is available at https://github.com/albarqounilab/promptgate.

Keywords: Active Learning · Federated Learning · Out-of-distribution.

1 Introduction

Deep learning has achieved impressive performance in medical image analysis, but still depends on large, carefully curated labeled datasets, which are scarce in routine clinical workflows. Active learning (AL) mitigates this by iteratively selecting informative unlabeled samples for annotation [17], and is a natural candidate to reduce pathologists' labeling effort. Most classical AL, however, assumes a *closed-set* scenario in which all pool images belong to the target classes. In contrast, hospital archives contain a heterogeneous mix of normal

* These authors equally contributed to this work.
✉ Corresponding author: Shadi Albarqouni (Shadi.Albarqouni@ukbonn.de)

tissue, artifacts, benign findings, and unrelated pathologies, so naïve AL often spends much of its budget on out-of-distribution (OOD) or non-target samples.

Open-set active learning (OAL) addresses this by separating in-distribution (ID) from OOD samples to maximize query *purity*. Recent OAL methods combine feature-based candidate sets with uncertainty sampling [13], margin-based detectors [19], or joint OOD filtering and class discovery [15]. OpenPath [23] shows that pre-trained vision-language models (VLMs) can warm up open-set AL via target/non-target text prompts in a centralized setting. These works, however, assume a single pooled dataset and treat OOD as a fixed global notion, which does not reflect real multi-institutional data.

In practice, data are fragmented across sites with different scanners and staining protocols. Federated learning (FL) [11] trains models without sharing raw data, and its combination with AL has produced federated AL (FAL) methods that reduce annotation cost under data-governance constraints [7, 1]. Extending this to *open-set federated active learning* (OS-FAL) is natural, yet existing approaches rely on task-specific representations and treat OOD geometrically in feature space, without leveraging VLM semantic priors.

Prompt learning [24] adapts frozen VLM backbones (e.g., BiomedCLIP [22]) by optimizing continuous context tokens, capturing task-specific semantics from a few labels. Concurrent federated prompt-learning methods similarly decompose prompts into shared and personalized components; e.g., FedOTP [9] aligns global and local prompts via optimal transport, FedPHA [4] reconciles heterogeneous clients through an SVD-based projection, and FOCoOp [10] adds dedicated OOD prompts with post-hoc scoring. All three, however, target *closed-set* federated classification over fixed, fully-labeled data: none provides an acquisition gate, operates on an unlabeled pool, or exploits the supervised non-target labels that OS-FAL supplies. Unlike these federated prompt-learning methods, and orthogonal to AL acquisition, we use VLM prompt learning as a *client-adaptive filtering front-end* that can precede any FAL strategy.

Concretely, we propose *PromptGate*, a lightweight dynamic VLM-gated module for OS-FAL. Prompts are split into *global* tokens (shared via FedAvg) and *local* tokens (personalized per site). At each round, the frozen VLM equipped with these prompts assigns every unlabeled image to either one of the target ID classes or a non-target category; only the images predicted as ID, form a high-purity candidate set, which is then passed to any acquisition AL rule (e.g., Random, Entropy). Oracle labels obtained for the queried samples subsequently update both prompt components in a CoOp-style fashion [24], progressively aligning VLM semantics with each site's notion of target vs. non-target. Our **contributions** are threefold: we introduce PromptGate, the first learnable-prompt VLM module for OS-FAL, with global/local prompt decomposition to capture heterogeneous OOD behavior; we propose a VLM-based gating mechanism that forms a high-purity candidate pool for any downstream acquisition AL rule; and we demonstrate plug-and-play improvements in query purity and label efficiency across multiple AL strategies on two federated medical imaging benchmarks.

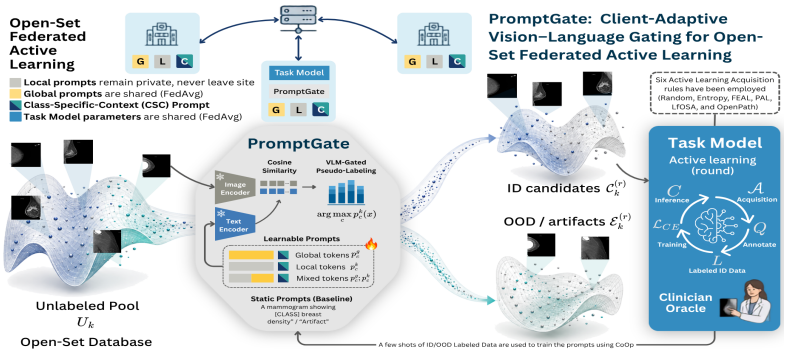

Fig. 1: Overview of PromptGate for OS-FAL.

2 Method

We now describe *PromptGate* (Fig. 1), our client-adaptive vision–language gating module for OS-FAL. We consider a C-class task with a central server and K clients. Client k holds a labeled set L_k and an unlabeled pool $U_k = U_{k,\mathrm{ID}} \cup U_{k,\mathrm{OOD}}$, where $U_{k,\mathrm{ID}}$ are target images and $U_{k,\mathrm{OOD}}$ non-target ones (artifacts, background, other pathologies). At each AL round r, the server broadcasts model and prompt parameters; each client uses a VLM to construct a *gated* ID-candidate pool from $U_k^{(r)}$, applies any acquisition strategy $\mathcal{A}$, and updates its local model and prompts with the newly acquired labels.

2.1 VLM with Global and Local Learnable Prompts

PromptGate assumes a pre-trained VLM with frozen image encoder E_{img} and text encoder E_{text}. For an image x, we obtain $z(x) = E_{\mathrm{img}}(x) \in \mathbb{R}^D$. Following CoOp [24], we model prompts as optimized continuous context tokens in the text embedding space while keeping the VLM weights fixed. We employ **Class-Specific Context (CSC)** optimization, assigning each ID class $c \in \{1, \ldots, C\}$ and a single non-target label "OOD" unique learnable tokens *factored* into: (i) global tokens $p_c^g \in \mathbb{R}^{d_p \times D}$, aggregated across clients to capture a shared semantic prior; and (ii) client-specific tokens $p_c^k \in \mathbb{R}^{d_p \times D}$, private to client k to adapt to local data heterogeneity. These matrices are concatenated to form a single context $[p_c^g; p_c^k]$. Given a class-specific template T_c, we obtain the text embedding.

$$t_c^k = E_{\mathrm{text}}\big([p_c^g; p_c^k], T_c\big) \in \mathbb{R}^D, \qquad c \in \{1, \ldots, C, \mathrm{OOD}\}.$$

2.2 VLM-Gated Pseudo-Labeling and Acquisition

For each $x \in U_k^{(r)}$, client k computes cosine similarities $s_c^k(x) = z(x)^\top t_c^k / \big(\|z(x)\| \, \|t_c^k\|\big)$ and converts them into pseudo-label probabilities via temperature-scaled softmax: $p_c^k(x) \propto \exp\big(s_c^k(x)/\tau_{\mathrm{VLM}}\big)$, with pseudo-label $\hat{y}^k(x)$ and define:

$$\hat{y}^k(x) = \arg\max_c p_c^k(x), \qquad p_{\max}^k(x) = \max_c p_c^k(x).$$

PromptGate then filters $U_k^{(r)}$ into an **ID-candidate pool (VLM-gated)**:

$$\mathcal{C}_k^{(r)} = \{x \in U_k^{(r)} \mid \hat{y}^k(x) \in \{1, \ldots, C\}\},$$

retaining samples whose most probable VLM class is an ID category. Here, gating is category-based on $K{+}1$ classes; we deliberately avoid a threshold in $p^k_{\max}(x)$ to avoid conflating gating with uncertainty-based AL scores. Implicit confidence filtering is achieved via the temperature-scaled softmax. The discarded pool $\mathcal{E}_k^{(r)} = U_k^{(r)} \backslash \mathcal{C}_k^{(r)}$ serves as Quality Control. Discarding is not permanent: rejected samples stay in U_k and are re-evaluated by the refreshed gate at every round, so any ID sample mistakenly excluded at round r can re-enter $\mathcal{C}_k^{(r')}$ at a later round $r' > r$ as prompt learning sharpens the ID/OOD boundary. Any AL strategy $\mathcal{A}$ then selects queries from the gated pool: $Q_k^{(r)} = \mathcal{A}\big(\mathcal{C}_k^{(r)}, L_k^{(r)}, \theta_k^{(r)}\big)$, where $\theta_k^{(r)}$ is the current client model. PromptGate then acts as a plug-in *pre-selection gate* in front of any OS-FAL rule.

2.3 Prompt Learning and Federated Aggregation

After querying, the oracle provides true labels for all $x \in Q_k^{(r)}$: either an ID class $y \in \{1, \ldots, C\}$ or a coarse non-target label. Let $\phi^g = \{p_c^g\}_c$ and $\phi^k = \{p_c^k\}_c$ denote global and client-specific tokens. We define a prompt loss:

$$\mathcal{L}^{(r)}_{\text{prompt},k}(\phi^g, \phi^k) = \frac{1}{|Q_k^{(r)}|} \sum_{x \in Q_k^{(r)}} \ell_{\text{CE}}\big(p^k(x), y(x)\big),$$

where ℓ_{CE} is cross-entropy between VLM class probabilities $p^k(x)$ and the true label $y(x)$. Each client performs gradient steps on ϕ^g, ϕ^k via SGD, warm-started from the previous round's prompt state. Each client then sends only the ϕ^g update to the server, which aggregates via FedAvg and broadcasts $\phi^{g,(r+1)}$. Local tokens $\phi^{k,(r+1)}$ remain private. Over rounds, global prompts capture a federated semantic prior for ID vs. non-target, while local prompts specialize to each site's OOD behavior, progressively improving the ID-candidate pool for any downstream OS-FAL rule. Prompts are not retrained from scratch at each round: they are warm-started from the previous round's solution and refined *incrementally* for a defined number of epochs on the round's oracle-labeled queries.

3 Experimental Design and Results

Baselines. We evaluate four configurations: (i) **Coldstart**, the standard federated AL baseline querying all unlabeled samples (ID and OOD) without filtering; (ii) **UpperB**, an oracle with only ID samples in the pool, representing an upper bound on AL performance; (iii) **Baseline** (VLM-Static), where a zero-shot BiomedCLIP partitions the pool into a Gated (predicted ID) and a Quality Control (predicted OOD) pool at R=0 and remains fixed thereafter. We benchmark the baselines against **Ours** (PromptGate), where the gate is refined by federated CoOp tuning in each round, following one of the configurations: Mixed (8G-8L; global+local), Global (16G; global only), and Local (16L; client-specific only).

Datasets. **FedISIC** [2, 20]: A public multi-center skin lesion benchmark with four sites (9,930; 3,163; 2,691; 1,807 samples) spanning eight classes. We simulate the OS-FAL scenario by injecting OOD samples (50% ratio relative to the local unlabeled ID pool) from DDI [3] and Fitzpatrick17k [6], capturing clinical shifts. Training data spans four clients with heterogeneous OOD prevalence (32.2%; 35.4%; 35.7%; 32.5%). The test set evaluates generalization across four centers with naturally varied OOD distributions: Center 0 (2,840 samples; 12.6% OOD), Center 1 (900; 12.1%), Center 2 (772; 13.0%), and Center 3 (522; 13.4%). **FedEMBED**: A breast density classification benchmark from the EMBED dataset [8] distributed across 222,700 training and 65,891 test samples [14]. The data spans four primary scanners—Selenia Dimensions (198,726 train / 58,925 test), Senograph 2000D ADS (10,043 / 2,810), Lorad Selenia (7,749 / 2,421), and Clearview CSm (6,182 / 1,735). To assess real-world OOD, we consider as OOD the clinical artifacts annotated by [16], which are organically present in the data. The total OOD prevalence varies heterogeneously across test domains, driven by distinct artifact profiles: Selenia Dimensions (13.12% total; predominantly 8.49% circles), Senograph 2000D ADS (11.03% total; predominantly 5.90% implants), Lorad Selenia (12.38% total; split between 5.95% circles and 4.10% implants), and Clearview CSm (4.78% total; 4.60% circles).

Evaluation Metrics. We report Balanced Multiclass Accuracy (BMA) on the test sets, Query Precision (QP; the fraction of true ID samples in each query round), Accumulated Query Recall (AQR; the cumulative fraction of the retrieved ID), and ID Purity (the fraction of true ID samples in the ID-candidate pool).

Implementation details. All closed-set AL (Random, Entropy [18], FAL (FEAL [1]) including OAL (PAL [21], LfOSA [12]) are introduced to OS-FAL using their default configurations. Local training follows standard FedAvg (batch 32, 15 epochs with lr 0.0005 on FedISIC and 50 epochs with lr 0.0003 on FedEMBED) across three seeds. We employ frozen BiomedCLIP [22] as the VLM backbone. Prompt tuning uses CSC CoOp with 8 global + 8 local learnable vectors per class (16 total) in the mixed configuration, trained via SGD (lr 0.002, momentum 0.9, weight decay 5e-4) for 15 epochs on FedISIC and 50 epochs on FedEMBED for a 128-shot subset per client; the gating temperature is fixed at $\tau_{\text{VLM}} = 0.0117$ following BiomedCLIP convention. The AL budget is 500 samples/round (R=5 for FedISIC, R=10 for FedEMBED). Downstream classifiers are EfficientNet-B0 for FedISIC [2] and a linear probe for FedEMBED [5]. For OpenPath* [23], we use its k-means strategy with a single static OOD prompt and no self-training, ensuring fair comparison. Base prompts follow dataset conventions: "A dermoscopy image of [CLASS]" / "Unknown or artifact" (FedISIC) and "A mammogram showing [CLASS] breast density" / "Unknown or artifact" (FedEMBED).

3.1 PromptGate vs. Baselines

Table 1 compares OS-FAL methods under different VLM filtering strategies across FedISIC and FedEMBED. A key structural difference is that Coldstart

Table 1: Comparison of VLM Filtering Strategies. Average scores of ID Purity and BMA over three seeds on the last AL round are reported. **Bold** indicates best purity, and Underline indicates best BMA within each dataset.

Strategies	Random	Entropy	FEAL	PAL	LfOSA	OpenPath*	Avg
FedISIC (R = 5)							
Coldstart	61.7 (60.2)	59.1 (62.0)	45.9 (57.6)	61.1 (50.5)	52.1 (55.7)	84.1 (60.5)	60.7 (57.8)
Baseline	60.9 (59.5)	72.9 (61.5)	78.7 (58.0)	73.1 (49.7)	76.1 (54.1)	83.8 (61.0)	74.3 (57.3)
Ours (Mixed)	97.4 (64.4)	98.9 (66.1)	98.7 (57.2)	**98.7** (50.4)	98.6 (54.9)	86.5 (63.5)	96.5 (59.4)
Ours (Global)	98.1 (59.8)	99.0 (63.7)	**98.9** (58.5)	**98.7** (53.1)	99.0 (58.3)	**86.6** (64.4)	96.7 (59.6)
Ours (Local)	**98.2** (65.3)	**99.5** (64.2)	98.8 (60.1)	**98.7** (50.2)	**99.2** (60.0)	86.5 (62.2)	**96.8** (60.3)
UpperB.	100.0 (63.2)	100.0 (67.2)	100.0 (53.5)	100.0 (62.9)	100.0 (64.3)	100.0 (60.2)	100.0 (61.9)
FedEMBED (R = 10)							
Coldstart	89.6 (60.0)	80.9 (59.2)	82.6 (42.5)	87.5 (58.0)	91.4 (58.9)	95.1 (60.8)	87.9 (56.6)
Baseline	89.9 (59.9)	80.2 (59.2)	82.8 (42.5)	87.7 (57.7)	92.4 (58.7)	95.0 (60.5)	88.0 (56.4)
Ours (Mixed)	**93.1** (60.6)	85.7 (59.4)	87.0 (43.1)	88.8 (58.1)	92.6 (59.0)	**96.0** (60.6)	90.5 (56.8)
Ours (Global)	92.6 (60.5)	85.2 (59.3)	86.8 (42.5)	88.5 (57.6)	**92.7** (58.8)	95.9 (60.6)	90.3 (56.6)
Ours (Local)	**93.1** (60.2)	**85.8** (59.5)	**87.4** (42.9)	**89.2** (58.2)	92.7 (58.6)	95.8 (60.2)	**90.7** (56.6)
UpperB.	100.0 (60.8)	100.0 (59.9)	100.0 (43.4)	100.0 (57.5)	100.0 (58.6)	100.0 (60.2)	100.0 (56.7)

operates on the full, unfiltered, unlabeled pool, whereas VLM-based methods query from a pre-filtered ID-candidate pool; consequently, the two settings train on differently composed batches from R=1, which explains divergent early-round BMA and QP. Averaged across all strategies, every PromptGate variant outperforms the Baseline in Purity on both datasets, with average gains of +22% on FedISIC and +2–3% on FedEMBED, and up to 3% BMA improvements.

FedISIC. Without filtering, the Coldstart averages only 60.7% Purity, while the static Baseline caps all strategies at ≤79%, except OpenPath* (83.8%), which benefits from stronger pretrained features. PromptGate variants achieve >97% Purity uniformly (avg. 96.8% for *Local*), with *Mixed* delivering the best overall BMA. As shown in Fig. 2 (top row), VLM-based QP starts at ∼60% at R=1, reflecting zero-shot gating before prompt adaptation, then temporarily dips at R=2 as the first CoOp update on a small 128-shot sample can shift the decision boundary sub-optimally; from R=3 onward, accumulated labels stabilize the prompts, and QP rises above 95%. By contrast, Coldstart and Baseline QP steadily deteriorate as OOD samples accumulate. Under strong PromptGate filtering, closed-set strategies close the gap with open-set methods and in several cases surpass them, suggesting that a high-purity gate effectively converts the open-set problem into a closed-set one. Final BMA remains inherently dependent on the downstream acquisition strategy: information-dense methods like Entropy reach 66.1% BMA under *Mixed*, whereas Random achieves 64.4%.

FedEMBED. Because organic OOD prevalence is low (4–13% across scanners), the Coldstart already reaches 87.9% average Purity and ∼90% QP at R=1—most random queries are naturally ID. At R=2, QP drops to ∼75% as uncertainty-driven strategies begin targeting ambiguous boundary samples that include subtle artifacts (e.g., circular markers); QP then recovers from R=3 on-

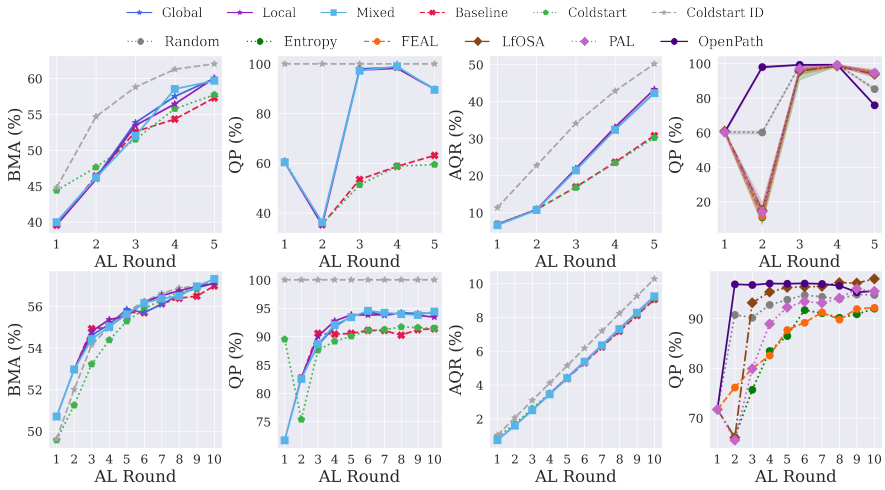

Fig. 2: Average BMA, QP, and AQR across all OS-FAL methods for different VLM filter variants. Top row: FedISIC. Bottom row: FedEMBED.

ward as the easy ID pool is consumed and query composition stabilizes. VLM-gated methods initially show slightly lower BMA (∼49% vs. ∼51% for Coldstart) because the filtered pool is smaller and somewhat less diverse; however, as prompt learning refines the boundary, this gap closes, and PromptGate yields a net advantage. At the final round, *Mixed* combined with OpenPath* reaches the highest Purity (96.0%) and BMA (60.6%), versus the Baseline's 95.0% / 60.5%. Scanner-specific artifacts make fully local token adaptation the strongest per-client filter (avg. 90.7% Purity).

Adapter Strategy. Ablating our approach (Table 1) reveals that *Local* strictly outperforms *Global* and *Mixed* aggregation. On FedISIC, the *Local* adapter achieves the highest average Purity (96.8%) and BMA (60.3%). On FedEMBED, *Local* maintains peak Purity (90.7%) with a negligible (-0.2%) BMA trade-off against *Mixed*. This confirms that tailoring semantic alignment to client-specific artifacts yields a superior gatekeeper compared to generalized global prompts. However, no single variant dominates universally (Table 1): *Global* leads in OpenPath* (86.6% Purity, 64.4 BMA) and in FEAL purity, and *Mixed* achieves the best peak BMA with a strong AL strategy.

4 Discussion and Conclusion

PromptGate Efficacy. When OOD is rare (4.8–13.1%; FedEMBED), the unfiltered pool is already mostly ID by chance, so gating adds only a modest ∼ 3% query purity. The decisive regime is the high-OOD FedISIC benchmark (50% OOD), where gating is essential; raising query purity from ∼ 60 to ∼ 98%. Qualitative examples of selected and discarded samples are shown in Fig. 4.

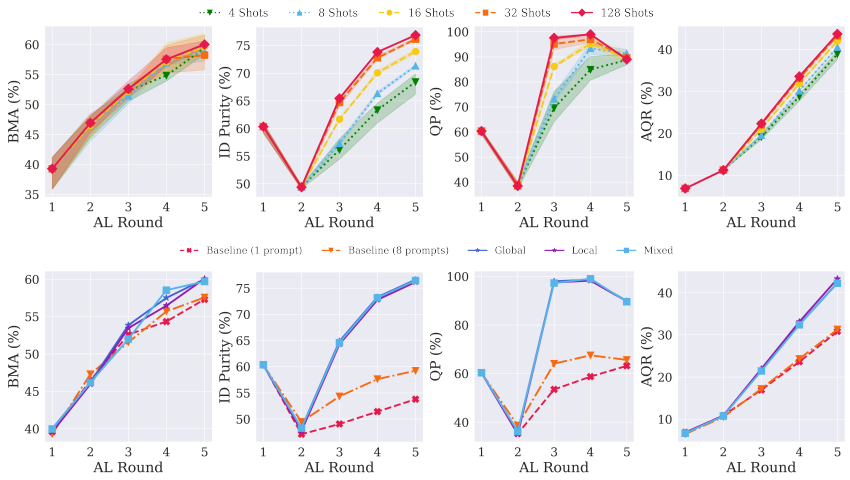

Fig. 3: **Influence of the number of shots and OOD prompts.** Average BMA, ID purity, QP, and AQR across all AL methods for both Baseline and PromptGate variants (Mixed, Local, and Global) for FedISIC benchmark.

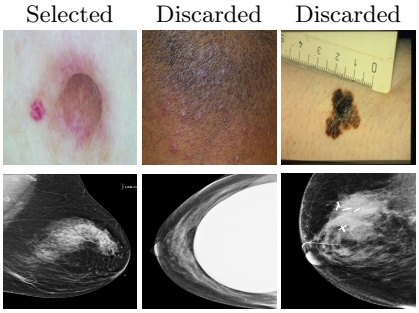

Fig. 4: Qualitative examples of PromptGate's gating. Top: FedISIC. Bottom: FedEMBED. The adapter accepts ID samples and discards OOD/artifact-corrupted samples, effectively performing implicit dataset cleaning before passing them to the AL Acquisition rule for Query selection.

Table 2: VLM test-set performance on FedISIC using random acquisition at AL round 5. We compare Global, Local, and Mixed CSC variants against the Static baseline across clients C0–C3.

Metric (%)	C0	C1	C2	C3	Avg
Baseline					
ID BMA	30.6	27.6	31.2	26.6	27.5
Acc.	83.2	87.9	85.8	82.4	84.4
OOD Recall	0.6	0.9	0.0	0.0	0.5
Ours (Global)					
ID BMA	33.9	62.9	**44.9**	36.7	41.8
Acc.	**99.6**	99.6	**99.7**	**99.9**	**99.6**
OOD Recall	97.9	99.7	**99.3**	99.1	98.5
Ours (Local)					
ID BMA	32.5	55.4	44.7	**38.9**	40.1
Acc.	**99.6**	99.4	**99.7**	99.8	**99.6**
OOD Recall	**98.3**	**100.0**	99.0	**99.1**	**98.8**
Ours (Mixed)					
ID BMA	**35.7**	**64.4**	43.2	35.3	**42.0**
Acc.	99.5	99.6	**99.7**	99.7	**99.6**
OOD Recall	97.6	99.7	**99.3**	97.6	98.2

Interestingly, our PromptGate discarded Breast Mammography samples, which were initially labeled as ID, while they had some artifacts, e.g., surgical clips, confirming the efficacy of PromptGate. Besides, we noticed that the preferred PromptGate variant depends on the OOD heterogeneity across clients, e.g.,

global prompts form a stronger shared prior when artifacts are similar across sites, whereas local prompts adapt better to client-specific OOD. We therefore recommend that the reader select the variant that best fits their setup.

The paradox of expanding static OOD prompts. Expanding static VLM prompts with additional OOD descriptions (e.g., eight instead of one) seems a natural way to enhance OS detection. However, Fig. 3 (bottom) shows a clear limitation: more static prompts yield only marginal gains, as one cannot anticipate the full diversity of clinical artifacts across clients. In contrast, dynamically adapting text embeddings to each local distribution provides a much more reliable gatekeeper. As shown in Fig. 3 (top), even a small few-shot adaptation budget (4–128 shots) improves anomaly query recall by nearly 20% while keeping downstream BMA stable (within $\pm 1.5\%$). Thus, dedicating a small fraction of the annotation budget to prompt initialization is more effective than the static prompt, yielding robust and autonomous filtering in subsequent AL rounds.

Limitations and future work. The PromptGate's fine-grained ID classification remains limited, e.g., 42% BMA on FedISIC (Table 2), compared to the downstream classifier $\sim$64.4% BMA in Table 1, confirming that the VLM should remain an OOD gatekeeper rather than a downstream classifier. Early-round prompt adaptation can also temporarily reduce query precision (Section 3), and the system inherits the VLM backbone's domain biases. Future work includes enforcing cross-modal agreement between the VLM and task model to stabilize early AL rounds, and selectively querying from the Quality Control pool. PromptGate itself introduces negligible overhead: only 16 prompt vectors ($\sim$12K parameters) per client with a frozen backbone. Its privacy-preserving, plug-and-play design makes PromptGate readily deployable across heterogeneous hospital networks without centralized data curation or site-specific engineering.

Acknowledgments. AN and DDG acknowledge the use of large language model–based assistants (ChatGPT, OpenAI) for editorial help in improving the grammar and readability of the manuscript, and (Anthropic, Claude) for code assistance.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Chen, J., Ma, B., Cui, H., Xia, Y.: Federated evidential active learning for medical image analysis. In: Proceedings of MICCAI 2024 Workshops. pp. 45–53 (2024). https://doi.org/10.1007/978-3-031-72390-2_7
2. Chen, J., Ma, B., Cui, H., Xia, Y.: Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11439–11449 (2024)
3. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science advances **8**(31), eabq6147 (2022)

4. Fang, C., Huang, W., Wan, G., Yang, Y., Ye, M.: FedPHA: Federated prompt learning for heterogeneous client adaptation. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 15960–15975. PMLR (13–19 Jul 2025)
5. Germani, E., Selin-Türk, I., Zeineddine, F., Mourad, C., Albarqouni, S.: Bias and generalizability of foundation models across datasets in breast mammography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2025)
6. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1820–1828 (2021)
7. Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16312–16322. IEEE (2023)
8. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., et al.: The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence **5**(1), e220047 (2023)
9. Li, H., Huang, W., Wang, J., Shi, Y.: Global and local prompts cooperation via optimal transport for federated learning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12151–12161 (2024). https://doi.org/10.1109/CVPR52733.2024.01155
10. Liao, X., Liu, W., Qian, J., Zhou, P., Xu, J., Wang, W., Chen, C., Zheng, X., Chua, T.S.: FOCoOp: Enhancing out-of-distribution robustness in federated prompt learning for vision-language models. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 37528–37554. PMLR (13–19 Jul 2025)
11. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. Pmlr (2017)
12. Ning, K.P., Zhao, X., Li, Y., Huang, S.J.: Active learning for open-set annotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41–49 (2022)
13. Qu, L., Ma, Y., Yang, Z., Wang, M., Song, Z.: Openal: An efficient deep active learning framework for open-set pathology image classification. In: International conference on medical image computing and computer-assisted intervention. pp. 3–13. Springer (2023)
14. Roschewitz, M., De Sousa Ribeiro, F., Xia, T., Khara, G., Glocker, B.: Robust image representations with counterfactual contrastive learning. Medical Image Analysis **105**, 103668 (2025)
15. Schmidt, S., Schenk, L., Schwinn, L., Günnemann, S.: Joint out-of-distribution filtering and data discovery active learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25677–25687 (2025)
16. Schueppert, A., Glocker, B., Roschewitz, M.: Radio-opaque artefacts in digital mammography: automatic detection and analysis of downstream effects. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)

17. Settles, B.: Active learning literature survey (2009)
18. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
19. Tang, J., Yue, Y., Wan, P., Wang, M., Zhang, D., Shao, W.: Osal-nd: Open-set active learning for nucleus detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2024)
20. Ogier du Terrail, J., Ayed, S.S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al.: Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. Advances in Neural Information Processing Systems **35**, 5315–5334 (2022)
21. Yang, Y., Zhang, Y., Song, X., Xu, Y.: Not all out-of-distribution data are harmful to open-set active learning. Advances in Neural Information Processing Systems **36**, 13802–13818 (2023)
22. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
23. Zhong, L., Liao, X., Zhang, S., Zhang, S., Wang, G.: Openpath: Open-set active learning for pathology image classification via pre-trained vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 481–490. Springer (2025)
24. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)