Beyond Pooling: Matching for Robust Generalization Under Data Heterogeneity

Ayush Roy SUNY Buffalo
Rudrasis Chakraborty Lawrence Livermore National Lab
Lav R. Varshney SUNY Stony Brook
Vishnu Suresh Lokhande SUNY Buffalo

Abstract

Pooling heterogeneous datasets across domains is a common strategy in representation learning, but naive pooling can amplify distributional asymmetries and yield biased estimators, especially in settings where zero-shot generalization is required. We propose a matching framework that selects samples relative to an adaptive centroid and iteratively refines the representation distribution. The double robustness and the propensity score matching for the inclusion of data domains make matching more robust than naive pooling and uniform subsampling by filtering out the confounding domains (the main cause of heterogeneity). Theoretical and empirical analyses show that, unlike naive pooling or uniform subsampling, matching achieves better results under asymmetric meta-distributions, which are also extended to non-Gaussian and multimodal real-world settings. Most importantly, we show that these improvements translate to zero-shot medical anomaly detection, one of the extreme forms of data heterogeneity and asymmetry. The code is available on Github.

1 INTRODUCTION

Pooling datasets from multiple institutions presents an attractive approach for studies constrained by limited sample sizes, as is often the case in biomedical research involving human subjects (Thompson et al., 2014). The aggregation of data across sites yields larger sample sizes and enhanced statistical power, thereby facilitating more robust analyses. Frequently, biomedical investigations face restrictions in recruiting subjects, particularly when the population of interest is rare, the relevant measurements are challenging to acquire, or when logistical and financial barriers preclude large-scale data collection. Although single-site datasets may suffice to address primary research questions, there is often interest in secondary analyses aimed at uncovering subtle associations between specific predictors and response variables. Such endeavors benefit significantly from pooled datasets, which boost the statistical signal and enable exploration of scientific questions that would otherwise remain inaccessible with smaller cohorts from individual institutions (Zhou et al., 2018)(Lokhande et al., 2022).

The feasibility of pooling datasets across institutions has been demonstrated in practical settings, supported by a mature body of statistical literature that outlines best practices for such analyses, including covariate matching (Mahajan et al., 2021)(Lokhande et al., 2022) and meta-analysis (Kim and Steiner, 2016). However, careful attention is required when aggregating data from multiple sources, as increasing training data in these multi-domain scenarios can sometimes lead to diminished overall accuracy (Gulrajani and Lopez-Paz, 2020; Zhou et al., 2022), uncertainty in fairness outcomes, and reduced performance for the most challenging subgroups (Shen et al., 2024).

These concerns extend beyond biomedical research, appearing in machine learning and computer vision domains where pooling distinct datasets is common, especially in anomaly detection tasks (Huang et al., 2024). Unlike anomalies in natural scene data, which are typically pronounced and readily identifiable, medical anomalies are subtle and often reflect slight deviations within complex imaging modalities such as chest X-rays, MRIs, CT scans, and retinal OCTs. Although comprehensive frameworks exist for mitigating sample selection bias and accounting for differences in population characteristics, there remains a scarcity of adaptive methodologies capable of accommodating sequentially added domains to facilitate robust pooled analyses, an important consideration across a range of

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

scientific applications.

In this work, we present an in-depth theoretical study of pooled dataset construction via matching, focusing on scenarios where domains are sequentially incorporated. We systematically compare naive pooling, uniform subsampling, and matching as strategies for combining data across multiple sites/domains. The proposed matching framework adaptively selects samples relative to an evolving centroid, iteratively refining the representation. We provide rigorous guarantees demonstrating that while all pooling methods converge to the population mean, naive pooling and subsampling retain domain-level heterogeneity, whereas matching uniquely filters out this extra variance and aligns with the target distribution. In finite-domain settings, naive pooling and subsampling are susceptible to unbounded error and bias, while matching delivers explicit finite-sample robustness. Synthetic experiments further validate these theoretical results, showing that matching maintains stability under practical sample sizes, whereas subsampling exhibits amplified errors. Moreover, our findings extend robustly to zero-shot medical anomaly detection, supporting the generalizability of the approach.

2 PRELIMINARIES OF POOLING STRATEGIES

Let the target test distribution be $\mathscr{D}_{\text{test}} = \mathcal{N}(\boldsymbol{\mu}_*, \sigma^2\mathbf{I}_d)$, an isotropic Gaussian in $\mathbb{R}^d$. Let K Gaussian component distributions be denoted as $\{Q_k\}_{k=1}^K$, where $Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2\mathbf{I}_d)$, $\boldsymbol{\mu}_k \overset{i.i.d.}{\sim} \mathscr{D}_\mu$, with $\mathbb{E}[\boldsymbol{\mu}_k] = \mu_*$ and $\text{Cov}[\boldsymbol{\mu}_k] = \Sigma_\mu$. This setup can be understood hierarchically. Firstly, we sample a domain mean $\boldsymbol{\mu}_k$ from the meta-distribution $\mathscr{D}_\mu$. Then, conditioned on $\boldsymbol{\mu}_k$, we define the domain distribution $Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2\mathbf{I}_d)$, from which draw i.i.d. samples $\mathbf{x} \sim Q_k$. The complete data-generation process can be expressed as a Bayes-style factorization:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid Q_k)\, p(Q_k \mid \boldsymbol{\mu}_k)\, p(\boldsymbol{\mu}_k \mid \mathscr{D}_\mu). \tag{1}$$

That is $\boldsymbol{\mu}_k \sim \mathscr{D}_\mu, Q_k \mid \boldsymbol{\mu}_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2\mathbf{I}_d)$. This introduces two levels of randomness: (i) domain-level variation from $\mathscr{D}_\mu$ (inter-domain) (ii) sample-level variation within each Q_k (intra-domain) . The Q_ks are not identical, because each is centered at different $\boldsymbol{\mu}_k$s. Their randomness comes from drawing $\boldsymbol{\mu}_k \sim \mathscr{D}_\mu$. Note that although samples $\mathbf{x}_i \sim Q_i$ and $\mathbf{x}_j \sim Q_j$ are i.i.d. conditioned on the domain mean $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ respectively, $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *are not exchangeable.*

Intuition on Eq. 1: This factorization simply expresses a sample x is generated following these steps: i) first a domain mean μ_k is drawn from the meta-distribution $\mathscr{D}_\mu$ ii) this μ_k defines a domain Q_k (i.e., the distribution for that domain) iii) finally, data x is sampled from the distribution of that domain, $p(x \mid Q_k), p(Q_k \mid \mu_k), p(\mu_k \mid \mathscr{D}_\mu)$.

Thus, the meta-distribution $\mathscr{D}_\mu$ is the distribution over domain means μ_k (e.g., $\mathscr{D}_\mu$ is a medical imaging database of all hospitals, and μ_k is a table inside $\mathscr{D}_\mu$ corresponding to one hospital). The shape of $\mathscr{D}_\mu$ (e.g., symmetric vs. asymmetric, light-tailed vs. heavy-tailed) fully determines whether the collection of domains is balanced or skewed. Thus, symmetry or asymmetry arises entirely at the level of sampling μ_k (e.g., μ_1 being data from a hospital in the USA and μ_2 being data from a hospital in India will have different distributions). The variability across datasets comes from variability in μ_k (the domain heterogeneity), and each domain then generates its own data.

We formally define three pooling strategies, naive pooling (Def. 1), uniform subsampling (Def. 2), and causal matching (Def. 3). Each corresponds to a different way of aggregating samples across domains.

Definition 1 (**Naive Pooling**). *Naive pooling aggregates all available samples from every domain:*

$$\mathcal{U}^{(K)} = \bigcup_{k=1}^{K} \{\, \mathbf{x}_{k,i} : \mathbf{x}_{k,i} \sim Q_k,\ i = 1, \ldots, N_k \,\}. \tag{2}$$

Equivalently, the pooled empirical distribution is $\hat{P}^{(K)}_{\text{pool}}(\mathbf{x}) = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K \sum_{i=1}^{N_k} \delta(\mathbf{x} - \mathbf{x}_{k,i})$*, where* N_k *is the sample size of domain* k*.*

Definition 2 (**Uniform Subsampling**). *Uniform subsampling randomly selects domains without bias and then samples a small subset of points from them. Formally, let* $I \subset \{1, \ldots, K\}$ *be a set of indices chosen uniformly at random. For each selected domain* Q_k*, we draw* $n \ll N_k$ *samples (with replacement):*

$$\mathcal{U}^{(n)}_{\text{sub}} = \bigcup_{k \in I} \{\, \mathbf{x}_{k,j} : \mathbf{x}_{k,j} \sim Q_k,\ j = 1, \ldots, n \,\}. \tag{3}$$

The corresponding empirical distribution is $\hat{P}^{(n)}_{\text{sub}}(\mathbf{x}) = \frac{1}{|I| \cdot n} \sum_{k \in I} \sum_{j=1}^{n} \delta(\mathbf{x}) - \mathbf{x}_{k,j})$*. Thus subsampling introduces randomness both at the* domain level *(random choice of* Q_k*) and at the* sample level *(random choice of* n *samples from each selected domain).*

Definition 3 (**Matching**). *Matching selectively includes domains based on their proximity to a adaptive centroid. A domain* Q_k *is included if its mean* μ_k *lies within a* τ*-radius ball of the matching function* M*:* $Q_k \in \mathcal{S}^{(\tau)} \iff M(\boldsymbol{\mu}_k - \mathbf{c}_t) < \tau$.

If Q_k is included, then *all of its samples* are admitted:

$$\mathcal{U}^{(\tau)}_{\text{match}} = \bigcup_{k \in \mathcal{S}^{(\tau)}} \{\, \mathbf{x}_{k,i} : \mathbf{x}_{k,i} \sim Q_k,\ i = 1, \ldots, N_k \,\}. \tag{4}$$

The matched empirical distribution is $\hat{P}^{(\tau)}_{\text{match}}(\mathbf{x}) = \frac{1}{\sum_{k \in \mathcal{S}^{(\tau)}} N_k} \sum_{k \in \mathcal{S}^{(\tau)}} \sum_{i=1}^{N_k} \delta(\mathbf{x} - \mathbf{x}_{k,i})$. The centroid is updated iteratively as $\mathbf{c}_{t+1} = \frac{1}{\sum_{k \in \mathcal{S}^{(\tau)}} N_k} \sum_{k \in \mathcal{S}^{(\tau)}} \sum_{i=1}^{N_k} \mathbf{x}_{k,i}$. Thus, unlike subsampling, matching operates via a *biased domain inclusion step*, where entire domains are accepted or rejected.

Specifically, matching can be viewed through the lens of domain-level propensity scores 4.1. For each domain k, define $W_k = \mathbb{I}\big(M(\boldsymbol{\mu}_k - \mathbf{c}_t) < \tau\big)$, where $W_k = 1$ indicates that all samples from domain Q_k are included in the matched set. This parallels causal inference, where W_k represents treatment assignment given covariates (Motiian et al., 2017) (here, $\boldsymbol{\mu}_k$). The "treatment" is domain inclusion, and the covariate is the domain mean. Unbiased causal estimation relies on three well-known assumptions: *ignorability*, *positivity*, and *consistency*. These map naturally into our framework: *ignorability* requires that inclusion depends only on observable $\boldsymbol{\mu}_k$, *positivity* requires that each relevant domain has a nonzero chance of inclusion, and *consistency* requires that included domains genuinely contribute samples from their Q_k (which holds since intra-domain samples are i.i.d.).

These assumptions clarify why matching succeeds where naive pooling and subsampling can fail. Naive pooling and subsampling, while exchangeable in the probabilistic sense, they do not satisfy ignorability, i.e., do not enforce *conditional balance with the target* $\boldsymbol{\mu}_*$: they freely include domains whose $\boldsymbol{\mu}_k$ deviate systematically from the target, introducing target-irrelevant variation (akin to confounding in observational studies; see Section 4.1). Matching, by contrast, *leverages* exchangeability and explicitly conditions inclusion on proximity to the adaptive centroid, thereby ensuring *ignorability and overlap with respect to the target.* This mechanism yields a form of double robustness: consistency is achieved as long as either (i) the inclusion rule (propensity model W_k) or (ii) the outcome model (centroid update) is correctly specified . Consequently, mild misspecification of the threshold τ does not compromise limiting behavior, making matching a stable strategy under domain heterogeneity.

In the subsequent sections, we analyze the aforementioned points theoretically while validating them experimentally. Particularly, Section 3 shows convergence of all methods for infinte number of domains, Section 4 demonstrates the challenges under finite K. Section 7 and 6 shows the extension of the proposed method to multimodal data and non gaussian settings respectively. All the theoretical insights are validated by the experimental results (Section 8) with additional explanations and experimental evidence been provided in Appendix.

3 ASYMPTOTIC CONVERGENCE ANALYSIS ($K \to \infty$)

Here, we show that all the pooling strategies converge to $\boldsymbol{\mu}_*$ when we have infinite domains, i.e., $K \to \infty$.

Theorem 1 (**Asymptotic Behavior as** $K \to \infty$)**.** *Let* $\{\boldsymbol{\mu}_k\}_{k=1}^K \overset{i.i.d.}{\sim} \mathscr{D}_\mu$ *with* $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$ *and* $\mathrm{Cov}(\boldsymbol{\mu}_k) = \Sigma_\mu$*. Suppose each domain* Q_k *generates i.i.d. samples* $\mathbf{x}_{k,i} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 I_d)$*, where* $\sigma^2 I_d$ *is the within-domain covariance. Then, as* $K \to \infty$*: a)* ***Naive pooling:*** $\hat{P}^{(K)}_{\text{pool}} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\mu}_*,\, \sigma^2 I_d + \Sigma_\mu\big)$*. b)* ***Uniform subsampling (fixed** n**):*** $\hat{P}^{(n)}_{\text{sub}} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\mu}_*,\, \sigma^2 I_d + \Sigma_\mu\big)$*. c)* ***Matching (fixed** $\tau > 0$**):*** *If the iterative centroid satisfies* $\mathbf{c}_n \xrightarrow{p} \mu_*$*, then* $\hat{P}^{(\tau)}_{\text{match}} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\mu}_*,\, \sigma^2 I_d\big)$*.*

Takeaway: From Theorem 1 (see Supplementary Sec. A.2 for proof), we see that as $K \to \infty$, all pooling methods converge to distributions centered at $\boldsymbol{\mu}_*$, but with critical differences in their covariance structures. naive pooling and uniform subsampling converge to $\mathcal{N}(\boldsymbol{\mu}_*, \sigma^2 \mathbf{I}_d + \Sigma_\mu)$, retaining the inter-domain variance Σ_μ that reflects domain heterogeneity. In contrast, matching converges to $\mathcal{N}(\boldsymbol{\mu}_*, \sigma^2 \mathbf{I}_d)$, effectively filtering out the inter-domain variance and matching the target test distribution exactly. This shows matching's unique ability to eliminate domain-level heterogeneity while preserving within-domain variability.

Corollary 1 (Supplementary Sec. A.3) shows that all pooling strategies (naive pooling, subsampling, and matching) are exchangeable because they are symmetric functions of the i.i.d. domain draws $\{\mu_k\}$. This holds for finite K and as $K \to \infty$: permuting domain indices does not change the joint distribution of the pooled estimator. Exchangeability implies that, conditional on $\{\boldsymbol{\mu}_k\}$, domain inclusion is "as if randomized," which corresponds to the *ignorability* condition used by propensity-score methods (Rosenbaum and Rubin, 1983). The distinction lies in how exchangeability is *used*. Naive pooling and subsampling do not enforce balance with respect to the target μ_* and thus can suffer finite-sample bias under asymmetric $\mathscr{D}_\mu$. Matching explicitly exploits exchangeability within a propensity framework: by conditioning inclusion on distance to the centroid, it secures *ignorability and overlap relative to the target*, aligning domain inclusion with the causal estimand of interest. Thus, while all strategies are exchangeable, only matching operationalizes

Figure 1: Comparison of the proposed method with MVFA (Huang et al., 2024), AnomalyCLIP (Zhou et al., 2023), and BiLORA (Zhu et al., 2025). Performance is reported in terms of domain alignment (DA), anomaly classification (AC), and anomaly segmentation (AS). The x-axis represents the sequential domain addition, and the y-axis represents the AUC scores. Comparison of Domain Alignment (DA) scores across all datasets for the Base, Agnostic, and GeoDVar methods. MVFA (Huang et al., 2024) shows moderate alignment with DA scores clustered around 2 (e.g., 2.2365 for HIS, 2.1146 for Chest-XRay, 2.0501 for OCT17, 2.0466 for Brain MRI AC, 2.0335 for Liver CT AC). AnomalyCLIP (Zhou et al., 2023) yields inconsistent and often poorer alignment, with scores varying widely from 1.0102 to 3.2632 across different tasks. BiLORA (Zhu et al., 2025) achieved better DA scores than MVFA (Huang et al., 2024) and AnomalyCLIP (Zhou et al., 2023) (HIS AC-4.0, ChestXray AC-3.145, OCT17 AC-4.0, BrainMRI AC-4.0, BrainMRI AS-4.012, LiverCT AC-3.158, LiverCT AS-4.030, RESC AC-3.081, RESC AS-4.072). In contrast, our method achieves consistently superior domain alignment, with DA scores at or exceeding 4.0 for all datasets and tasks, including peaks of 4.0736 for Brain MRI detection (AC) and 4.0484 for Brain MRI segmentation (AS), demonstrating its robust performance in matching complex domain distributions.

this property to mitigate finite-sample bias. However, $K \to \infty$ is not a practical setting. In the subsequent sections we will analyze the unbiasness/biasness of the pooling techniques under finite K (for symmetric and assymetric of meta distribution $\mathscr{D}_\mu$ as seen in Def. 4).

4 CONVERGENCE ANALYSIS UNDER SYMMETRIC $\mathscr{D}_\mu$ for K

4.1 SYMMETRY OF META DISTRIBUTION $\mathscr{D}_\mu$

Definition 4 (Distributional Symmetry). *A meta-distribution $\mathscr{D}_\mu$ is* symmetric *around $\boldsymbol{\mu}_*$ if $\boldsymbol{\mu}_k - \boldsymbol{\mu}_* \overset{d}{=} \boldsymbol{\mu}_* - \boldsymbol{\mu}_k$.*

Intuitively, symmetry means domain means are equally likely to deviate above or below the population center, while on the contrary, asymmetry implies systematic bias in one direction. Furthermore, Assumption 2 highlights the relevance of $\mathscr{D}_{\boldsymbol{\mu}}$ for real world scenarios where sampling a domain (Q_k) that is far away from the mean $\boldsymbol{\mu}_*$ is highly likely and not a rare event. In example 1, the occasional $\boldsymbol{\mu}_* + 5$ domains are averaged out by the majority of $\boldsymbol{\mu}_*$ domains as K grows. Hence, in the infinite-sample limit, deviations vanish and the pooled mean coincides with $\boldsymbol{\mu}_*$.

Example 1. *Consider a one-dimensional meta-distribution $\mathscr{D}_\mu$ defined as follows: $\boldsymbol{\mu}_k = \boldsymbol{\mu}_* + c$ with probability p, $\boldsymbol{\mu}_*$ with probability 1-p. This distribution satisfies Assumption 2: it has a finite mean and variance, sub-exponential tails, finite fourth moment, and importantly nontrivial directional mass at $\boldsymbol{\mu}_* + c$.*

Finite-K sampling: *If K domains are drawn, the probability that at least one Q_k is centered at $\boldsymbol{\mu}_* + c$ is $1 - (1-p)^K$ (p can be considered small based on the*

fact that as we go far from the mean μ_*, the probability decreases). Thus, with high probability we obtain at least one "outlier" domain whose mean is far from $\boldsymbol{\mu}_*$. This illustrates that asymmetric realizations are not rare but expected *under the assumption.* As $K \to \infty$: *The law of large numbers guarantees that* $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\mu}_k \longrightarrow \boldsymbol{\mu}_*$ *almost surely.*

4.2 UNBIASED ESTIMATION FOR FINITE K

Theorem 2 (**Finite-K Unbiasedness under Symmetry**). *Suppose the meta-distribution $\mathscr{D}_\mu$ is symmetric about $\boldsymbol{\mu}_*$ in the sense of Def. 4, so that $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$. For any finite number of domains $K \geq 1$:*

1) ***Naive Pooling:*** *Let $\bar{\mu}_K = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k} \mathbf{x}_{k,i}}{\sum_{k=1}^{K} N_k}$, $x_{k,i} \sim Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 I_d)$. Then $\mathbb{E}[\bar{\mu}_K] = \boldsymbol{\mu}*$. b)* ***Uniform Subsampling:*** *For a uniformly chosen subset $I \subset \{1, \ldots, K\}$ and n samples per selected domain, the subsample mean $\bar{\boldsymbol{\mu}}_I = {}^1\!/_{n|I|}\sum_{k\in I}\sum_{j=1}^{n} \mathbf{x}_{k,j}$ satisfies $\mathbb{E}[\bar{\mu}_I] = \boldsymbol{\mu}_*$. c)* ***Matching:*** *If the centroid is initialized at the target, $\mathbf{c}^{(0)} = \boldsymbol{\mu}_*$, then at every iteration the matched mean $\bar{\mu}_{\mathcal{S}} = \frac{1}{\sum_{k\in\mathcal{S}^{(\tau)}} N_k}\sum_{k\in\mathcal{S}^{(\tau)}}\sum_{i=1}^{N_k} \mathbf{x}_{k,i}$ remains unbiased, i.e. $\mathbb{E}[\bar{\mu}_{\mathcal{S}}] = \boldsymbol{\mu}_*$.*

Takeaway: Theorem 2 (see Supplementary Sec. A.4 for proof.) demonstrates that under symmetric $\mathscr{D}_{\boldsymbol{\mu}}$ (Def. 4), all three pooling strategies yield unbiased estimators of the target mean $\boldsymbol{\mu}_*$ for any finite K. This establishes a baseline where domain heterogeneity is balanced around the target, allowing all methods to perform well. Furtherore, for matching even if $\mathbf{c}^{(0)} \neq \boldsymbol{\mu}_*$, appropriate selection of τ ensures that relevant domains are selected and $\bar{\mu}_{\mathcal{S}}$ converges to $\boldsymbol{\mu}_*$.

This is often rare for real world datasets, thus in the next section we analyze the convergence rates of the pooling methods under asymmetric $\mathscr{D}_{\boldsymbol{\mu}}$ (Def. 4).

5 CONVERGENCE ANALYSIS UNDER ASYMETRIC $\mathscr{D}_\mu$ FOR FINITE K

While Theorem 1 establishes asymptotic convergence properties and Theorem 2 shows finite-K convergence under ideal symmetry, real-world applications involve finite K with distributional asymmetries that create significant challenges. This section analyzes the critical transition from K to $K+1$ domains under asymmetric conditions.

Theorem 3 (**Finite-Sample Robustness under Domain Addition**). *Let $\bar{\boldsymbol{\mu}}_K := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\mu}_k$ denote the empirical average of the first K domain means (with $\mu_k = \mathbb{E}_{x\sim Q_k}[f(x)]$), and let $\varepsilon_K := \|\bar{\boldsymbol{\mu}}_K - \boldsymbol{\mu}^*\|_2$. Fix $\delta \in (0,1)$. Under Assumption 2: 1.* ***Naive pooling.*** *With probability at least $1-\delta$, $\varepsilon_{K+1} \geq \frac{K}{K+1}\varepsilon_K - \frac{\beta\log(2/\delta)}{K+1} - O\left(\sqrt{\frac{\log(1/\delta)}{K}}\right)$. Moreover, there exist adversarial μ_{K+1} such that with constant probability, $\varepsilon_{K+1} \geq \varepsilon_K + \Theta(1)$ (unbounded per-step deterioration). 2.* ***Uniform subsampling.*** *If m domains are drawn uniformly at random and n samples per domain are averaged, then $\mathbb{E}[\varepsilon_{K+1}^2] \leq \varepsilon_K^2 + \frac{\lambda_{\max}(\Sigma_\mu)}{m} + \frac{d\sigma^2}{mn}$. Hence $\mathbb{E}[\varepsilon_{K+1}] \leq \varepsilon_K + O\big(\frac{1}{\sqrt{m}} \vee \frac{1}{\sqrt{mn}}\big)$, while worst-case deterioration per step is $\Omega(1/m)$. 3.* ***Matching.*** *Let S_K be the matched set of size $|S_K|$, with centroid $\hat{c}_K$. Suppose $|S_K| \geq C_0\log(1/\delta)$. If the domain $K+1$ is excluded, then $\varepsilon_{K+1} \leq \varepsilon_K + C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$. Furthermore, if it is included, then with probability $1-\delta$, $\varepsilon_{K+1} \leq \varepsilon_K + \frac{\tau}{m_M(|S_K|+1)} + C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$. If $M(\boldsymbol{\mu}_{K+1} - \boldsymbol{\mu}s^*) \leq \varepsilon_K + \tau/L_M$, then $\Pr(\varepsilon_{K+1} \leq \varepsilon_K) \geq 1 - \exp(-\Omega(|S_K|))$ (high-probability non-deterioration).*

Takeaway: From Theorem 3 (see Supplementary A.5 for proof) we get he comparison in Table 2 (see Supplementary A.5) highlights the fundamental difference between matching and the other strategies. For **naive pooling**, the error can jump by a $\Theta(1)$ amount even as K grows, meaning deterioration never vanishes. For **subsampling**, the worst-case increase is $\Omega(1/m)$, which shrinks only with the number of domains subsampled per step. **Matching** limits the worst-case increase to $O(1/|S_K|)$, where $|S_K|$ is the size of the matched set. Since $|S_K|$ typically grows with K, perturbation vanishes as more domains are included. Thus, matching converts constant or $1/m$-level risks into a self-shrinking $1/|S_K|$ bound (only strategy with explicit non-deterioration guarantees under heterogeneity).

Corollary 2 (see Supplementary Sec. A.6) provides a practical criterion for selecting the matching threshold τ to ensure safe domain addition. By setting τ appropriately relative to the current error ϵ_K and the distance of harmful domains under metric M, degradation is not observed under the introduciton of new domains. Table 2 summarizes these differences: only causal matching simultaneously provides bounded deterioration, explicit outlier robustness, and a non-deterioration safety guarantee. In the following section, we see that these theoretical guarantees of matching and the properties of the matched set S_K (see Supplementary B) translate to arbitrary distribution.

Synthetic experiments: We conducted controlled synthetic experiments to empirically validate the guarantees presented in Theorems 1, 2, and 3 under enhanced challenges. Each domain was generated as

$Q_k = \mathcal{N}(\mu_k, \sigma^2 I_d)$ with means $\mu_k \sim \mathscr{D}_\mu$, while the target was fixed as $\mathscr{D}_{\text{test}} = \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. All the analysis done so far assumed large per-domain sample size N, where $\hat{\mu}_k \to \mu_k$. In contrast, we used small finite N so that sample noise persists. This revealed that all three pooling strategies degrade compared to large-N theory, but to different extents (Nazarovs et al., 2021). Pooling and matching remain relatively stable, whereas subsampling suffers the largest errors due to $\sqrt{d/N}$ variance amplification. Experiments targeted the three regimes: (a) **Asymptotic behavior** (Theorem 1): convergence as $K \to \infty$ under asymmetric domains. (b) **Symmetric finite-K convergence** (Theorem 2): error bounds with symmetric domain means. (c) **Finite-sample robustness** (Theorem 3): stability when incrementally adding domains under asymmetry. All strategies were implemented as defined in Sec. 2, repeated over 10 seeds and averaged. Matching used iterative centroid refinement with L_2 metric, robust median initialization, and convergence criterion $|\mathbf{c}_{t+1} - \mathbf{c}_t|_2 < 10^{-4}$. Additional details can be seen in Supplementary C.1 due to space constrains.

6 USING NORMALIZING FLOW TO GENERALIZE FOR ARBITRARY DISTRIBUTIONS

Although analysis of matching for Gaussian distribution ensures theoretical validity, extended theoretical guarantee to arbitrary distributions would demonstrate the superiority of matching over naive pooling, in line with transport-regularized schemes in multi-marginal settings (Mehta et al., 2023). Theorem 4 (see Supplementary A.7 for proof) shows the preservation of the properties (see Supplementary B) of $\mathcal{S}$ under arbitrary distributional settings.

Theorem 4 (**Normalizing Flow Transport Theorem with Lipschitz Guarantees**). *Let p_{data} be a compactly supported data distribution and $p_Z = \mathcal{N}(0, I_d)$. Under Assumption 1, there exists a bijective normalizing flow $T : \mathbb{R}^d \to \mathbb{R}^d$ such that: 1) **Universal Approximation:** For any $\epsilon > 0$, $D_{KL}(T_{\#}p_Z \,\|\, p_{data}) < \epsilon$, where $T_{\#}p_Z$ denotes the* push-forward *of p_Z by T (i.e., the distribution of $T(z)$ for $z \sim p_Z$), and D_{KL} is the Kullback–Leibler divergence. 2) **Lipschitz Control:** T can be constructed with finite Lipschitz constant L_T controlled by architectural norms; in particular, if $T = T_L \circ \cdots \circ T_1$ and the Jacobian of layer l, $\|J_{T_\ell}\|_{\text{op}} \le C_\ell$ (C_l represents the maximum operator norm of the Jacobian for the l-th layer), then $L_T \le \prod_{\ell=1}^{L} C_\ell$. 3) **Variance Preservation:** For any matched set $\mathcal{S}_Z = \{z : \|z - \mathbf{c}_z\|_2 < \tau\}$, letting $\mathcal{S}_X = T(\mathcal{S}_Z)$, $\frac{1}{|\mathcal{S}_X|}\sum_{x\in\mathcal{S}_X} \|x - T(\mathbf{c}_z)\|_2^2 \le L_T^2\tau^2$. 4) **Concentration Guarantee:** With $\bar{x} = \frac{1}{|\mathcal{S}_X|}\sum_{x\in\mathcal{S}_X} x$, we have $\|\bar{x} - T(\mathbf{c}_z)\|_2 \le L_T\tau$.*

Takeaway: Theorem 4 (see Supplementary Sec. A.7 for proof) demonstrates that the favorable properties of matching extend from Gaussian to real-world non-Gaussian data through normalizing flows. The utilization of normalizing flow is motivated by the invariance of Bayes error under invertible transformations, as established by (Theisen et al., 2021). This allows us to employ normalizing flows as approximately invertible mappings, ensuring that the robustness bounds of Theorem 3 remain valid beyond Gaussian distributions. The key insight is that flows preserve the geometric structure of matched sets: tight clusters in the latent space ($\|z - \mathbf{c}_z\|_2 < \tau$) map to tight clusters in the data space (variance $\le L_T^2\tau^2$). This means the theoretical guarantees for Gaussian data (bounded error, variance reduction, and concentration) carry over to real-world distributions. The Lipschitz constant L_T acts as a distortion factor, showing that well-behaved flows (small L_T) maintain the matching benefits. This bridges our Gaussian analysis to practical applications where data exhibit complex, multimodal structure.

7 EXTENSION TO MULTIMODAL SETTINGS

The theoretical framework established thus far assumes a unimodal distribution where a single centroid $\mathbf{c}_n$ converges to the target mean $\boldsymbol{\mu}_*$. However, real-world applications, particularly in medical imaging, often involve multimodal distributions corresponding to multiple classes of the dataset. We define multimodal data distribution as in Def. 5.

Definition 5 (Multimodal Data Distribution)**.** *A data distribution P is* M-modal *if it can be expressed as a mixture of M unimodal components $P(x) = \sum_{m=1}^{M} \pi_m P_m(x)$, where: 1) $\pi_m \ge 0$ are mixing coefficients with $\sum_{m=1}^{M} \pi_m = 1$ 2) Each P_m is unimodal with mode $\boldsymbol{\mu}_m^*$ and covariance Σ_m 3) The component distributions may have different shapes and scales.*

The target distribution in multimodal settings becomes $\mathscr{D}_{\text{test}} = \sum_{m=1}^{M} \pi_m \mathcal{N}(\boldsymbol{\mu}_m^*, \sigma_m^2 \mathbf{I}_d)$.

We maintain centroids $\{\mathbf{c}_{n,m}\}_{m=1}^{M}$ and radii $\{\tau_m\}_{m=1}^{M}$, i.e., a sample $\mathbf{x}$ is assigned to mode m if $M(\mathbf{x} - \mathbf{c}_{n,m}) < \tau_m$ and $M(\mathbf{x} - \mathbf{c}_{n,j}) \ge \tau_j$ for all $j \ne m$. Centroids are updated per mode using only assigned samples. Lemma 1 (Supplementary Sec. A.8) ensures M-modal convergence as M unimodal convergence under sufficient separation of the M modes. Under the conditions of Lemma 1, matching guarantees that adding new domains never increases the error for any mode $\epsilon_{K+1,m} \le \epsilon_{K,m} \quad \forall m = 1, \ldots, M$ with high probability, where $\epsilon_{K,m} = \|\mathbf{c}_{n,m}^{(K)} - \boldsymbol{\mu}_m^*\|_2$. This follows directly from applying Theorem 3 to each mode independently. The separation condition ensures that harm-

ful domains affecting one mode do not interfere with other modes. Each mode's matching process evolves independently and enjoys the same non-deterioration guarantees as the unimodal case.

8 EXPERIMENTAL RESULTS FOR ZERO-SHOT ANOMALY DETECTION

The synthetic Gaussian experiments in Section C.1 confirmed our theoretical guarantees while exposing finite per-domain sample effects. Unlike the large-N assumptions in theory, limited N introduces sample-level noise, producing severe degradation under subsampling and milder but still visible declines for pooling and matching. This connects theory with practice: robust adaptation must handle finite-N asymmetries rather than relying solely on asymptotic guarantees.

Zero-shot medical anomaly detection provides the strongest stress test of this setting, extending beyond synthetic domains to one of the most challenging real-world scenarios for distributional generalization. Unlike natural scene anomalies (e.g., surface defects) that are structurally consistent and visually distinct, medical anomalies are subtle, heterogeneous, and often overlap with normal anatomical variation. This induces three extreme forms of heterogeneity: (a) **modality heterogeneity**, as imaging technologies (X-ray, MRI, CT, OCT) generate fundamentally distinct feature distributions; (b) **pathological similarity**, as abnormal and normal samples exhibit high feature overlap, making separation difficult; and (c) **institutional bias**, where scanner protocols and acquisition pipelines introduce asymmetric shifts even within the same modality. Together, this "triple heterogeneity" defines the exact worst-case conditions for naive pooling, breaking exchangeability assumptions and amplifying asymmetry. The zero-shot requirement—generalization to unseen modalities and institutions. further exacerbates these challenges.

Under such conditions, three failure modes emerge when $\mathcal{D}_{\text{test}}$ is unseen: (1) the meta-distribution $\mathscr{D}_\mu$ is inherently asymmetric (Def. 4); (2) finite-sample bias becomes unavoidable under pooling; and (3) asymmetry compounds as domains are sequentially added. This is precisely where causal matching is critical. Unlike domain-level pooling, our implementation operates at the *sample level*, using an adaptive threshold τ in the learned feature space. The feature space filters nuisance variability while preserving anomaly–normal semantics, making sub-exponential sample-level behavior more realistic. Consequently, causal matching provides robustness in the finite-N regime, yielding empirical non-deterioration exactly where pooling and subsampling fail, and extending theoretical guarantees (positivity, ignorability, consistency) beyond the asymptotic, domain-level analysis.

Setup: The baselines and related works are described in Supplementary Sec. C.2.1. We employ the BMAD benchmark (Bao et al., 2024) spanning five medical domains with six datasets: brain MRI (BrainMRI), liver CT (LiverCT), retinal OCT (OCT2017)), chest X-ray (ChestXray), and digital histopathology (HIS). This selection captures the extreme heterogeneity discussed in Section 8, with BrainMRI, LiverCT, and RESC supporting both anomaly classification (AC) and segmentation (AS), while OCT17, ChestXray, and HIS are used for AC only. Following the leave-one-out strategy (Huang et al., 2024), we train on datasets $D_1, D_2, ..., D_{i-1}$ and evaluate on unseen D_i. This setup directly instantiates the theoretical challenges of asymmetric domain adaptation: test domains may exhibit selective semantic alignment with specific training domains, creating the perfect conditions for the performance degradation predicted by Theorem 3. While selectively fine-tuning on aligned domains might seem appealing, medical data scarcity necessitates using all available data making robust pooling strategies essential.

Evaluation Metrics: We report standard Area Under the ROC Curve (AUC) metrics: image-level AUC for anomaly classification and pixel-level AUC for segmentation. However, these conventional metrics fail to capture a critical aspect of real-world deployment: the *monotonicity of improvement* when incorporating additional domains. Theorem 3 highlighted that under asymmetry, the critical quantity is the per-step deterioration $\varepsilon_{K+1} - \varepsilon_K$, which can be unbounded for naive pooling but is controlled under structured strategies. To empirically measure this effect, we introduce the **Data Addition Score (DA)**, which translates the theoretical notion of stepwise deterioration into an evaluation metric at the performance level. Formally, given performance sequence $\mathbf{y} = [y_1, y_2, y_3, y_4, y_5]$ where y_i denotes performance after incorporating i domains, and importance weights $\mathbf{s} = [0.1, 0.2, 0.3, 0.4]$ emphasizing later stages, the DA is defined as:

$$\text{DA}(\mathbf{y}) = \sum_{i=1}^{4} \mathbb{1}\{y_{i+1} \geq y_i\} \cdot \left(1 + \frac{y_{i+1} - y_i}{10} \cdot s_i\right). \quad (5)$$

Here, the indicator $\mathbb{1}\{y_{i+1} \geq y_i\}$ captures the non-deterioration requirement—if performance drops, the score for that step is zero—while the weighted improvement term rewards consistent gains. In this way, the DA directly operationalizes the deterioration analysis of Theorem 3, providing an empirical lens on a method's improvements as domains are added.

Table 1: **Comparisons with state-of-the-art zero-shot anomaly detection methods.** The AUCs (in %) for AC and AS are reported. The best result is in bold, and the second-best result is underlined. * are the results we achieved after reimplementation.

Method	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
	AC	AC	AC	AC	AS	AC	AS	AC	AS
WinCLIP(Jeong et al., 2023)	69.85	<u>70.86</u>	46.64	66.49	85.99	64.20	96.20	42.51	80.56
APRIL-GAN(Chen et al., 2023)	72.36	57.49	92.61	76.43	**91.79**	70.57	97.05	75.67	85.23
AnomalyCLIP*(Deng and Li, 2022)	**80.27**	69.11	94.60	79.72	90.47	78.72	97.88	88.49	91.14
AdaCLIP(Cao et al., 2024)	-	-	68.8	73.1	-	-	-	-	-
Mao et al.(Mao et al., 2025)	-	-	91.2	73.7	-	-	-	-	-
MVFA-AD*(Huang et al., 2024)	<u>78.88</u>	69.11	96.62	75.05	<u>90.33</u>	**80.77**	98.06	88.53	91.27
BiLORA*(Zhu et al., 2025)	77.68	66.75	<u>97.05</u>	<u>87.65</u>	89.40	<u>80.64</u>	<u>98.53</u>	<u>89.15</u>	<u>93.46</u>
Ours	75.41	**74.94**	**97.21**	**87.67**	90.29	80.26	**98.75**	**89.42**	**95.09**

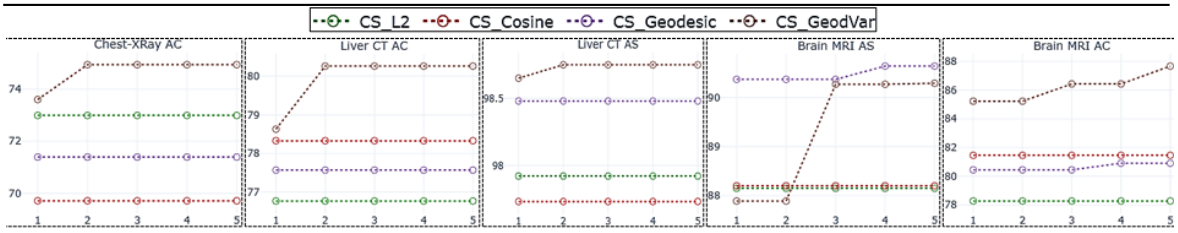

Figure 2: **Ablation comparing different matching metrics M:** Euclidean distance (CS_L2), cosine similarity (CS_Cosine), geodesic distance (CS_Geodesic), and geodesic distance with VACA (CS_GeodVar). Performance is reported in terms of domain alignment (DA), anomaly classification (AC), and anomaly segmentation (AS). X axis represents the sequential domain addition and the Y axis represents the AUC scores. CS_L2 and CS_Cosine achieves DA of 4 for BrainMRI, LiverCT, and ChestXRay. CS_Geo achieves DA of 4 for ChestXRay and LiverCT while achieving DA of 4.0138 for BrainMRI detection (AC) and 4.0081 for BrainMRI segmentation (AS). CS_GeodVar achieves DA of 4.0134 for ChestXRay, 4.0163 for LiverCT detection (AC), 4.0010 for LiverCT segmentation (AS), 4.0736 for BrainMRI detection (AC) and 4.0484 for BrainMRI segmentation (AS).

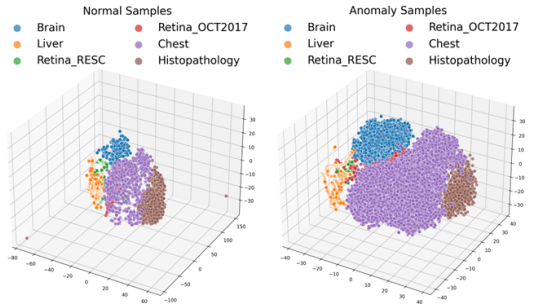

Figure 3: **Illustration of modality-induced clustering in CLIP feature space.** Embeddings of distinct modalities form disjoint angular regions on the hypersphere (see Table 3 in Supplementary Sec. C.2.2 for more details).

8.1 MATCHING FOR ANOMALY DETECTION

Geodesics for Matching: Zero-shot transfer is hindered by the well-documented *modality gap* (Levi and Gilboa, 2024; Liang et al., 2022), where CLIP embeddings from text and image subspaces occupy disjoint regions of the hypersphere. Our analysis extends this: even within the image space, modality-specific clustering (e.g., ChestXRay, BrainMRI, Histopathology) confines embeddings to distinct angular regions $\mathbb{S}^{d-1}$. This creates large inter-modality angular gaps (Figure 3), making zero-shot anomaly detection particularly challenging. If parameters lie closer to one modality's cluster, the semantic gap to others is amplified. Because CLIP embeddings are ℓ_2-normalized, they reside on $\mathbb{S}^{d-1}$, where the intrinsic metric is the geodesic distance $d_G(f_1, f_2) = \arccos(f_1^\top f_2)$. Unlike Euclidean distance (underestimates large gaps) or cosine similarity (lacks metric structure), d_G respects hyperspherical curvature, preserves neighborhoods, and faithfully represents modality separations. Under the bi-Lipschitzness assumption (Assumption 1), this ensures stability of matching and underpins theoretical guarantees.

Centroid Initialization and Update: We initialize normal and anomaly centroids at $\mathbf{c}^{(0)} = \mathbf{0}$, avoiding warm-start bias. For a feature f_j satisfying the geodesic matching criterion, centroids are updated by exponential moving average (EMA): $\tilde{\mathbf{c}}^{(t+1)} = \alpha \mathbf{c}^{(t)} + (1-\alpha) f_j$, $\mathbf{c}^{(t+1)} = \frac{\tilde{\mathbf{c}}^{(t+1)}}{\|\tilde{\mathbf{c}}^{(t+1)}\|_2}$, $\alpha = 0.5$. Projection ensures centroids remain on $\mathbb{S}^{d-1}$. Within the spherical cap where matching occurs, d_G behaves like Euclidean distance up to small scaling, so all guarantees—positivity, ignorability, consistency, and finite-K robustness (Theorem 3)—remain intact. The geometric advantage of d_G is that centroid neighborhoods expand consistently with the hypersphere, allowing a

tighter adaptive threshold τ.

Geodesic Loss for Intra-/Inter-Cluster Structure: We directly embed d_G in the learning objective:

Intra-cluster compactness: $\mathcal{L}_{\text{intra}} = \frac{1}{|\mathcal{S}^+|}\sum_{j\in\mathcal{S}^+} d_G(f_j, \mathbf{c}^+)^2 + \frac{1}{|\mathcal{S}^-|}\sum_{j\in\mathcal{S}^-} d_G(f_j, \mathbf{c}^-)^2$. *Inter-cluster separation:* $\mathcal{L}_{\text{inter}} = -d_G(\mathbf{c}^+, \mathbf{c}^-)$.

Maximizing $\mathcal{L}_{\text{inter}}$ prevents mode collapse and guarantees identifiability. The final geodesic loss is $\mathcal{L}_{\text{geo}} = \lambda_1\mathcal{L}_{\text{intra}} + \lambda_2\mathcal{L}_{\text{inter}}, \lambda_1, \lambda_2 > 0$, balancing compactness and separation. This directly addresses Lemma 1, which shows that minimal mode separation is necessary for consistent convergence.

Variance-Aware Channel Attention (VACA): Variance decomposition (Figure 8, Supplementary Sec. C.3) reveals that only a subset of embedding dimensions carries class-specific signal (Yu et al., 2024), while many encode modality-specific variance that induces hyperspherical clustering (Figure 3). We address this with *variance-aware channel attention* (VACA), which explicitly upweights class-discriminative channels and suppresses modality-driven ones. Let patch embeddings be $P \in \mathbb{R}^{B\times N\times D}$ (batch B, patches N, channels D), normalized along the channel dimension, and let class embeddings for *normal* and *anomaly* be $T \in \mathbb{R}^{D\times 2}$ (L2-normalized). Broadcasting and taking their Hadamard product yields per-channel contributions, aggregated into a discriminative score vector $\Delta \in \mathbb{R}^D$. Channels with high Δ_d align with the anomaly/normal axis, while low-Δ_d reflect modality noise. A small MLP with Softplus output produces nonnegative gains, scaled by γ to form channel weights $w = 1 + \gamma\, a$. Reweighted patches $\widetilde{P} = P \odot w$ emphasize class signal, while a variance surrogate $\mathsf{Var}_{\text{disc}}$, computed from weighted Δ, summarizes discriminative energy. This reweighting shrinks angular gaps between modalities and strengthens "local positivity" in hyperspherical neighborhoods, ensuring subsequent geodesic matching (Section 8.1) operates on richer, domain-invariant features.

Ablation experiments: Figure 2 and Supplementary Sec. C.3.1 highlight the effectiveness of geodesic distance in capturing faithful neighborhoods on the hypersphere, outperforming Euclidean and cosine metrics. When combined with VACA, DA scores improve consistently, as variance-aware reweighting suppresses modality-specific leakage and amplifies class-discriminative channels. This shifts modality clusters closer in the discriminative subspace, increasing successful cross-domain matching. Geometrically, this restores local domain overlap, directly supporting the *positivity assumption* by ensuring that normal and anomaly features across modalities share sufficient support for non-zero inclusion probability. Given the scarcity of medical data, this guarantees that relevant samples remain available for matching.[1] The overall framework is summarized in Supplementary Sec. C.3.3 (Algorithm 1).

Empirical comparison with the existing methods: Compared to the state of the art anomaly detection methods like MVFA (Huang et al., 2024), AnomalyCLIP (Zhou et al., 2023), and continual learning frameworks like BiLORA (Zhu et al., 2025) that utilize naive pooling for incremental learning setups where domains (finetuning as domains keeps on increasing), our proposed method performs better due to the inherent advantage of matching by filtering out the confounding data points (see Fig. 1). We see that the DA score, which measures the dips in performance of the model as newer domains are introduced for finetuning (in accordance to Theorem 3), are always above 4 (a DA score of 4 or above indicates no decrease or increase in the model's performance as we keep on finetuning on newer domains) for our proposed method unlike the existing methods (detailed analysis is provided in Supplementary Sec. C.3.2). Table 1 demonstartes that our method achieves comparable performance with respect to the state of the art zero-shot anomaly detection methods while also surpassing them in most cases. *Takeaway:* The proposed Geodesic-aware Matching method is the only method with non-detoriative performance behaviour with addition of more data compared to exisitng baselines along with superior downstream performance.

9 CONCLUSION

Pooling heterogeneous datasets is attractive for representation learning, but we showed both theoretically and empirically that naive pooling or uniform subsampling can amplify distributional asymmetries and harm generalization, especially under domain addition. Our matching framework addresses this by adaptively selecting samples around hypersphere centroids, ensuring double robustness and filtering confounding domains that violate positivity. Practically, we instantiated these ideas in zero-shot medical anomaly detection, an extreme heterogeneity case where CLIP embeddings cluster by modality. Using geodesic distance as the intrinsic matching metric and augmenting it with VACA, we emphasized class-discriminative while suppressing modality-driven dimensions.

[1] Unlike a fixed threshold, τ is defined adaptively as the minimum distance of a sample to the centroids (See Supplementary Sec. C.3.3). This ensures that matching decisions evolve naturally as centroids are updated via exponential moving averages and as feature space transforms while training.

Acknowledgments

Prof. Lokhande acknowledges support from University at Buffalo startup funds, an Adobe Research Gift, an NVIDIA Academic Grant, and the National Center for Advancing Translational Sciences of the NIH (award UM1TR005296 to the University at Buffalo). Dr. Chakraborty performed this work under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

J. Bao, H. Sun, H. Deng, Y. He, Z. Zhang, and X. Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053, 2024.

P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

X. Chen, Y. Han, and J. Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

H. Deng and X. Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022.

I. Diakonikolas and D. M. Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

E. Dunipace. Optimal transport weights for causal inference. *arXiv preprint arXiv:2109.01991*, 2021.

I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

F. Gunsilius and Y. Xu. Matching for causal effects via multimarginal unbalanced optimal transport. *arXiv preprint arXiv:2112.04398*, 2021.

C. Huang, A. Jiang, J. Feng, Y. Zhang, X. Wang, and Y. Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024.

C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International conference on machine learning*, pages 2078–2087. PMLR, 2018.

J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

A. Jiang, C. Huang, Q. Cao, S. Wu, Z. Zeng, K. Chen, Y. Zhang, and Y. Wang. Multi-scale cross-restoration framework for electrocardiogram anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2023.

Y. Kim and P. Steiner. Quasi-experimental designs for causal inference. *Educational psychologist*, 51(3-4): 395–405, 2016.

D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

H. Lee, C. Pabbaraju, A. P. Sevekari, and A. Risteski. Universal approximation using well-conditioned normalizing flows. *Advances in Neural Information Processing Systems*, 34:12700–12711, 2021.

M. Y. Levi and G. Gilboa. The double-ellipsoid geometry of clip. *arXiv preprint arXiv:2411.14517*, 2024.

H. Li, J. Wu, and V. Braverman. Memory-statistics tradeoff in continual learning with structural regularization. *arXiv preprint arXiv:2504.04039*, 2025.

V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

V. S. Lokhande, R. Chakraborty, S. N. Ravi, and V. Singh. Equivariance allows handling multiple nuisance variables when analyzing pooled neuroimaging datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10432–10441, 2022.

D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pages 7313–7324. PMLR, 2021.

K. Mao, P. Wei, Y. Lian, Y. Wang, and N. Zheng. Beyond single-modal boundary: Cross-modal anomaly detection through visual prototype and harmonization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9964–9973, 2025.

R. Mehta, J. Kline, V. S. Lokhande, G. Fung, and V. Singh. Efficient discrete multi-marginal optimal transport regularization. 2023.

S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

J. Nazarovs, R. R. Mehta, V. S. Lokhande, and V. Singh. Graph reparameterizations for enabling 1000+ monte carlo iterations in bayesian deep neural networks. In *Uncertainty in Artificial Intelligence*, pages 118–128. PMLR, 2021.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022.

M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3076–3085, 2017.

J. H. Shen, I. D. Raji, and I. Y. Chen. The data addition dilemma. *arXiv preprint arXiv:2408.04154*, 2024.

R. Theisen, H. Wang, L. R. Varshney, C. Xiong, and R. Socher. Evaluating state-of-the-art classification models against bayes optimality. *Advances in Neural Information Processing Systems*, 34:9367–9377, 2021.

P. M. Thompson, J. L. Stein, S. E. Medland, D. P. Hibar, A. A. Vasquez, M. E. Renteria, R. Toro, N. Jahanshad, G. Schumann, B. Franke, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior*, 8(2):153–182, 2014.

Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

X. Yu, S. Yoo, and Y. Lin. Clipceil: Domain generalization through clip via channel refinement and image-text alignment. *Advances in Neural Information Processing Systems*, 37:4267–4294, 2024.

H. H. Zhou, V. Singh, S. C. Johnson, G. Wahba, A. D. N. Initiative, Berkeley, and C. DeCArli. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proceedings of the National Academy of Sciences*, 115(7):1481–1486, 2018.

K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4396–4415, 2022.

Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

H. Zhu, Y. Zhang, J. Dong, and P. Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25613–25622, 2025.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A PROOFS OF THEORETICAL RESULTS

A.1 ASSUMPTIONS

Assumption 1 (Metric equivalence)**.** *Let $M : \mathbb{R}^d \to [0, \infty)$ be the matching metric used by the algorithm. There exist constants $0 < m_M \le L_M < \infty$ such that $m_M \|v\|_2 \le M(v) \le L_M \|v\|_2 \qquad$ for all $v \in \mathbb{R}^d$*

Assumption 1 states M is locally equivalent to the Euclidean norm on the data manifold. This means that M does not distort distances by more than a fixed multiplicative factor. If, for example, $M(v)$ were much larger or smaller than $\|v\|_2$, depending on v, matching could select domains with means very far from the centroid, breaking the tight concentration of the matched set. In practice, this means distances measured under M cannot collapse to zero or blow up arbitrarily relative to ℓ_2 distances: the constants (m_M, L_M) control this distortion. Such equivalence guarantees that concentration arguments and robustness bounds established in Euclidean space remain valid when extended to M, since neighborhoods defined by M correspond to neighborhoods on the manifold up to constant factors. Moreover, because compact manifolds admit finite bi-Lipschitz charts, this equivalence allows our theoretical results to generalize from Gaussian settings in $\mathbb{R}^d$ to arbitrary compact Riemannian manifolds, ensuring that matching remains stable and generalizable across domains.

Assumption 2 (Moment and Tail Conditions)**.** *We assume all domain means μ_k are drawn i.i.d. from a meta-distribution D with: a) Finite mean and covariance: $\mathbb{E}[\mu_k] = \mu_*$ and $\mathrm{Cov}(\mu_k) = \Sigma_{\boldsymbol{\mu}} < \infty$ b) Sub-exponential tails: $\exists C, \beta > 0$ such that $\mathbb{P}(\|\mu_k - \mu_*\| > t) \le C \exp(-\beta t)$ c) Finite fourth moment: $\mathbb{E}[\|\mu_k - \mu_*\|_2^4] < \infty$ d) There exist a unit vector $v \in \mathbb{R}^d$, constants $p_0 \in (0, 1]$ and $\delta > 0$, such that the half-space $S_{v,\delta} = \{\mu : (\mu - \mu_*) \cdot v \ge \delta\}$ satisfies $\mathbb{P}(\mu_k \in S_{v,\delta}) \ge p_0$.*

Throughout our proofs, we use (i) the Strong Law of Large Numbers (SLLN), (ii) Lévy's Continuity Theorem, which states that almost sure convergence of characteristic functions implies convergence in distribution. All applications assume the moment and tail conditions in Assumption 2.

A.2 PROOF OF THEOREM 1

Theorem 1 (**Asymptotic Behavior as $K \to \infty$**)**.** *Let $\{\boldsymbol{\mu}_k\}_{k=1}^K \overset{i.i.d.}{\sim} \mathscr{D}_\mu$ with $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$ and $\mathrm{Cov}(\boldsymbol{\mu}_k) = \Sigma_\mu$. Suppose each domain Q_k generates i.i.d. samples $\mathbf{x}_{k,i} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 I_d)$, where $\sigma^2 I_d$ is the within-domain covariance. Then, as $K \to \infty$: a) **Naive pooling:** $\hat{P}_{\text{pool}}^{(K)} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\mu}_*, \sigma^2 I_d + \Sigma_\mu\big)$. b) **Uniform subsampling (fixed n):** $\hat{P}_{\text{sub}}^{(n)} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\mu}_*, \sigma^2 I_d + \Sigma_\mu\big)$. c) **Matching (fixed $\tau > 0$):** If the iterative centroid satisfies $\mathbf{c}_n \xrightarrow{p} \mu_*$, then $\hat{P}_{\text{match}}^{(\tau)} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\mu}_*, \sigma^2 I_d\big)$.*

Proof. **Naive Pooling Convergence:** Let $\{\boldsymbol{\mu}_k\}_{k=1}^K$ be i.i.d. draws from $\mathscr{D}_{\boldsymbol{\mu}}$ with $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$ and $\mathrm{Cov}[\boldsymbol{\mu}_k] = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}$. The characteristic function of $\hat{P}_{\text{pool}}^{(K)}$ is given by $\phi_P^{(K)}(\mathbf{t}) = \mathbb{E}\left[\exp(i\mathbf{t}^\top \mathbf{x})\right] \quad$ where $\quad \mathbf{x} \sim \hat{P}_{\text{pool}}^{(K)}$

Since $\hat{P}_{\text{pool}}^{(K)} = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K \sum_{i=1}^{N_k} \delta(\mathbf{x} - \mathbf{x}_{k,i})$ and each $\mathbf{x}_{k,i} \sim Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}_d)$, we have $\phi_P^{(K)}(\mathbf{t}) = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K N_k \cdot \mathbb{E}_{\mathbf{x} \sim Q_k}\left[\exp(i\mathbf{t}^\top \mathbf{x})\right] = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K N_k \cdot \exp\left(i\mathbf{t}^\top \boldsymbol{\mu}_k - \frac{1}{2}\sigma^2 \|\mathbf{t}\|^2\right)$

Assuming balanced domains where $N_k = N$ for all k (or that $\frac{N_k}{\sum_{j=1}^K N_j} \to \frac{1}{K}$ as $K \to \infty$), we obtain $\phi_P^{(K)}(\mathbf{t}) = \frac{1}{K} \sum_{k=1}^K \exp\left(i\mathbf{t}^\top \boldsymbol{\mu}_k - \frac{1}{2}\sigma^2 \|\mathbf{t}\|^2\right) + o(1)$

By the Strong Law of Large Numbers for i.i.d. random vectors: $\frac{1}{K} \sum_{k=1}^K \exp(i\mathbf{t}^\top \boldsymbol{\mu}_k) \xrightarrow{a.s.} \mathbb{E}_{\boldsymbol{\mu} \sim \mathscr{D}_{\boldsymbol{\mu}}}\left[\exp(i\mathbf{t}^\top \boldsymbol{\mu})\right] =: \phi_{\boldsymbol{\mu}}(\mathbf{t})$

Under the moment conditions ($\mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}_*$, $\mathrm{Cov}[\boldsymbol{\mu}] = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}$), the characteristic function admits the expansion: $\phi_{\boldsymbol{\mu}}(\mathbf{t}) = \exp\left(i\mathbf{t}^\top \boldsymbol{\mu}_* - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{t} + o(\|\mathbf{t}\|^2)\right)$

Thus, for any fixed $\mathbf{t} \in \mathbb{R}^d$: $\phi_P^{(K)}(\mathbf{t}) \xrightarrow{a.s.} \exp\left(-\frac{1}{2}\sigma^2\|\mathbf{t}\|^2\right) \cdot \exp\left(i\mathbf{t}^\top \boldsymbol{\mu}_* - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{t}\right) = \phi_\infty(\mathbf{t})$, where $\phi_\infty(\mathbf{t})$ is the characteristic function of $\mathcal{N}(\boldsymbol{\mu}_*, \sigma^2\mathbf{I}_d + \boldsymbol{\Sigma}_{\boldsymbol{\mu}})$. By Lévy's Continuity Theorem, almost sure convergence of characteristic functions implies convergence in distribution $\hat{P}_{\text{pool}}^{(K)} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}_*, \sigma^2\mathbf{I}_d + \boldsymbol{\Sigma}_{\boldsymbol{\mu}})$

Uniform Subsampling Convergence: The uniform subsampling procedure selects domains uniformly at random and draws $n \ll N_k$ samples from each selected domain. The key insight is that this process is equivalent to $\hat{P}_{\text{sub}}^{(n)}(\mathbf{x}) = \frac{1}{|I|\cdot n}\sum_{k\in I}\sum_{j=1}^{n}\delta(\mathbf{x} - \mathbf{x}_{k,j})$ where $I \subset \{1, \ldots, K\}$ is a random subset of domains chosen uniformly.

This can be reinterpreted as: each subsampled point $\mathbf{x}_{k,j}$ is obtained by 1) Sampling a domain index $k \sim \text{Uniform}\{1, \ldots, K\}$ (through the random selection of I) 2) Sampling $\mathbf{x} \sim Q_k$ within that domain.

As $K \to \infty$, the domain selection process converges to drawing $\boldsymbol{\mu} \sim \mathscr{D}_{\boldsymbol{\mu}}$ then $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$. The characteristic function of this limiting process is $\phi_{\text{sub}}^{(\infty)}(\mathbf{t}) = \mathbb{E}_{\boldsymbol{\mu}\sim\mathscr{D}_{\boldsymbol{\mu}}}\left[\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\mathbf{I}_d)}\left[\exp(i\mathbf{t}^\top\mathbf{x})\right]\right] = \mathbb{E}_{\boldsymbol{\mu}\sim\mathscr{D}_{\boldsymbol{\mu}}}\left[\exp(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\sigma^2\|\mathbf{t}\|^2)\right]$. This matches $\phi_\infty(\mathbf{t})$ from the naive pooling case. The finite sample size n affects the rate of convergence but not the asymptotic limit, as the empirical characteristic function converges to its expectation by the Law of Large Numbers.

Matching Convergence: We analyze the matching distribution defined as $\hat{P}_{\text{match}}^{(\tau)}(\mathbf{x}) = \frac{1}{\sum_{k\in\mathcal{S}^{(\tau)}} N_k}\sum_{k\in\mathcal{S}^{(\tau)}}\sum_{i=1}^{N_k}\delta(\mathbf{x} - \mathbf{x}_{k,i})$, where $\mathcal{S}^{(\tau)} = \{k : M(\boldsymbol{\mu}_k - \mathbf{c}_n) < \tau\}$ and $\mathbf{c}_n$ is the running centroid.

The characteristic function of the matched distribution is $\phi_{\text{match}}^{(\tau)}(\mathbf{t}) = \mathbb{E}\left[\exp(i\mathbf{t}^\top\mathbf{x})\right] \quad \text{where} \quad \mathbf{x} \sim P_{\text{match}}^{(\tau)}$

Since $\hat{P}_{\text{match}}^{(\tau)}$ is an empirical mixture over selected domains, we have $\phi_{\text{match}}^{(\tau)}(\mathbf{t}) = \frac{\sum_{k\in\mathcal{S}^{(\tau)}} N_k \cdot \mathbb{E}_{\mathbf{x}\sim\hat{Q}_k}[\exp(i\mathbf{t}^\top\mathbf{x})]}{\sum_{k\in\mathcal{S}^{(\tau)}} N_k}$ where $\hat{Q}_k$ is the empirical distribution of domain k. As $K \to \infty$ and for each domain k, as $N_k \to \infty$ (or under the assumption that domain sample sizes are sufficiently large), the empirical characteristic function converges: $\mathbb{E}_{\mathbf{x}\sim\hat{Q}_k}[\exp(i\mathbf{t}^\top\mathbf{x})] \to \mathbb{E}_{\mathbf{x}\sim Q_k}[\exp(i\mathbf{t}^\top\mathbf{x})] = \exp(i\mathbf{t}^\top\boldsymbol{\mu}_k - \frac{1}{2}\sigma^2\|\mathbf{t}\|^2)$

The matching criterion selects domains where $M(\boldsymbol{\mu}_k - \mathbf{c}_n) < \tau$. Given that $\mathbf{c}_n \xrightarrow{p} \boldsymbol{\mu}_*$ (which follows from the iterative centroid update converging to the target mean), we have $W_k = \mathbb{I}\{M(\boldsymbol{\mu}_k - \mathbf{c}_n) < \tau\} \xrightarrow{p} \mathbb{I}\{M(\boldsymbol{\mu}_k - \boldsymbol{\mu}_*) < \tau\}$

By the law of large numbers for the weighted average over selected domains: $\phi_{\text{match}}^{(\tau)}(\mathbf{t}) \xrightarrow{p} \frac{\mathbb{E}_{\boldsymbol{\mu}\sim\mathscr{D}_{\boldsymbol{\mu}}}\left[\mathbb{I}\{M(\boldsymbol{\mu}-\boldsymbol{\mu}_*)<\tau\}\cdot\exp(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\sigma^2\|\mathbf{t}\|^2)\right]}{\mathbb{E}_{\boldsymbol{\mu}\sim\mathscr{D}_{\boldsymbol{\mu}}}[\mathbb{I}\{M(\boldsymbol{\mu}-\boldsymbol{\mu}_*)<\tau\}]}$. As the matching process becomes increasingly selective (in the sense that τ can be chosen to decrease appropriately with K), we consider the limit $\lim_{\tau\to 0}\phi_{\text{match}}^{(\tau)}(\mathbf{t}) = \frac{\mathbb{E}_{\boldsymbol{\mu}\sim\mathscr{D}_{\boldsymbol{\mu}}}\left[\delta(\boldsymbol{\mu}-\boldsymbol{\mu}_*)\cdot\exp(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\sigma^2\|\mathbf{t}\|^2)\right]}{\mathbb{E}_{\boldsymbol{\mu}\sim\mathscr{D}_{\boldsymbol{\mu}}}[\delta(\boldsymbol{\mu}-\boldsymbol{\mu}_*)]}$, where $\delta(\cdot)$ represents the limiting density concentration at $\boldsymbol{\mu}_*$.

Under the condition that $\mathscr{D}_{\boldsymbol{\mu}}$ has positive density at $\boldsymbol{\mu}_*$ (which holds for continuous meta-distributions), we obtain $\phi_{\text{match}}^{(\tau)}(\mathbf{t}) \to \exp(i\mathbf{t}^\top\boldsymbol{\mu}_* - \frac{1}{2}\sigma^2\|\mathbf{t}\|^2)$. This is the characteristic function of $\mathcal{N}(\boldsymbol{\mu}_*, \sigma^2\mathbf{I}_d)$. By Lévy's Continuity Theorem: $\hat{P}_{\text{match}}^{(\tau)} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}_*, \sigma^2\mathbf{I}_d)$. The double robustness property ensures that this convergence holds even with mild misspecification of the matching radius τ, as long as the centroid estimation is consistent. □

A.3 PROOF OF COROLLARY 1

Corollary 1 (**Exchangeability of Pooling Strategies**). *Each pooling strategy is invariant to permutations of domain indices under the following conditions: a)* ***Naive Pooling:*** *The pooled distribution* $\hat{P}_{\text{pool}}^{(K)}(\mathbf{x}) = \frac{1}{\sum_{k=1}^{K} N_k}\sum_{k=1}^{K}\sum_{i=1}^{N_k}\delta(\mathbf{x} - \mathbf{x}_{k,i})$ *is permutation-invariant by commutativity of addition. Relabeling domain indices leaves the sum unchanged. b)* ***Uniform Subsampling:*** *The subsampled distribution* $\hat{P}_{\text{sub}}^{(n)}(\mathbf{x}) = \frac{1}{|I|\cdot n}\sum_{k\in I}\sum_{j=1}^{n}\delta(\mathbf{x} - \mathbf{x}_{k,j})$, *where* $I \subset \{1, \ldots, K\}$ *is chosen uniformly at random, is exchangeable. Uniform sampling of domains preserves the joint law under permutations of indices. c)* ***Matching:*** *The matched distribution* $\hat{P}_{\text{match}}^{(\tau)}(\mathbf{x}) = \frac{1}{\sum_{k\in\mathcal{S}^{(\tau)}} N_k}\sum_{k\in\mathcal{S}^{(\tau)}}\sum_{i=1}^{N_k}\delta(\mathbf{x} - \mathbf{x}_{k,i})$, *where* $\mathcal{S}^{(\tau)} = \{k : M(\boldsymbol{\mu}_k - \mathbf{c}_n) < \tau\}$, *is*

exchangeable. The centroid $\mathbf{c}_n$ is permutation-invariant, and inclusion depends only on distances to this centroid, which are unaffected by relabeling.

Thus, all three pooling strategies are exchangeable, both in case of finite K and as $K \to \infty$. □

Proof. We fix any finite K and any permutation $\pi : \{1, \dots, K\} \to \{1, \dots, K\}$. We show that permuting domain labels leaves each pooled empirical distribution unchanged in law; in fact, for (a) it is unchanged pointwise.

a) Naive pooling. Recall $\hat{P}^{(K)}_{\text{pool}}(\mathbf{x}) = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K \sum_{i=1}^{N_k} \delta(\mathbf{x} - \mathbf{x}_{k,i})$. Under the relabeling $k \mapsto \pi(k)$ we obtain

$$\frac{1}{\sum_{k=1}^K N_{\pi(k)}} \sum_{k=1}^K \sum_{i=1}^{N_{\pi(k)}} \delta(\mathbf{x} - \mathbf{x}_{\pi(k),i}) = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K \sum_{i=1}^{N_k} \delta(\mathbf{x} - \mathbf{x}_{k,i}) = \hat{P}^{(K)}_{\text{pool}}(\mathbf{x}),$$

since $\{N_{\pi(k)}\}_{k=1}^K$ is just a reordering of $\{N_k\}_{k=1}^K$ and addition is commutative. Thus naive pooling is permutation-invariant (exchangeable) for every realization and hence in distribution.

(b) Uniform subsampling. Let $I \subset \{1, \dots, K\}$ be chosen uniformly at random and

$$\hat{P}^{(n)}_{\text{sub}}(\mathbf{x}) = \frac{1}{|I|\, n} \sum_{k \in I} \sum_{j=1}^{n} \delta(\mathbf{x} - \mathbf{x}_{k,j}).$$

The law of I is invariant under relabeling: for any permutation π, $\pi(I)$ is uniform iff I is uniform. Therefore, conditional on the data $\{\mathbf{x}_{k,j}\}$, the distribution of $\frac{1}{|\pi(I)|\, n} \sum_{k \in \pi(I)} \sum_{j=1}^n \delta(\mathbf{x} - \mathbf{x}_{k,j})$ coincides with that of $\hat{P}^{(n)}_{\text{sub}}(\mathbf{x})$. Hence $\hat{P}^{(n)}_{\text{sub}}$ is exchangeable.

(c) Matching. Let $\mathcal{S}^{(\tau)} = \{\, k : M(\mu_k - \mathbf{c}_n) < \tau \,\}$, $\hat{P}^{(\tau)}_{\text{match}}(\mathbf{x}) = \frac{1}{\sum_{k \in \mathcal{S}^{(\tau)}} N_k} \sum_{k \in \mathcal{S}^{(\tau)}} \sum_{i=1}^{N_k} \delta(\mathbf{x} - \mathbf{x}_{k,i})$. The centroid $\mathbf{c}_n$ is a permutation-invariant statistic of the multiset of observed samples (and, recursively, of prior matched sets); in particular, it depends on domains only through symmetric sums/averages. Consequently, the inclusion set $\mathcal{S}^{(\tau)}$ depends on $\{\mu_k\}$ only via distances to $\mathbf{c}_n$ and is unchanged by relabeling domain indices. Therefore, $\hat{P}^{(\tau)}_{\text{match}}$ is invariant under permutations and hence exchangeable.

In all three cases, permutation of domain labels leaves the empirical distribution unchanged (pointwise for naive pooling; in law for subsampling and matching). Thus each pooling strategy is exchangeable for any finite K, and with even stronger guarentees as $K \to \infty$. □

A.4 PROOF OF THEOREM 2

Theorem 2 (**Finite-K Unbiasedness under Symmetry**). *Suppose the meta-distribution $\mathscr{D}_\mu$ is symmetric about $\boldsymbol{\mu}_*$ in the sense of Def. 4, so that $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$. For any finite number of domains $K \geq 1$:*

*1) **Naive Pooling:** Let $\bar{\mu}_K = \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} \mathbf{x}_{k,i}}{\sum_{k=1}^K N_k}$, $x_{k,i} \sim Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 I_d)$. Then $\mathbb{E}[\bar{\mu}_K] = \boldsymbol{\mu}*$. b) **Uniform Subsampling:** For a uniformly chosen subset $I \subset \{1, \dots, K\}$ and n samples per selected domain, the subsample mean $\bar{\boldsymbol{\mu}}_I = {}^1\!/_{n|I|} \sum_{k \in I} \sum_{j=1}^n \mathbf{x}_{k,j}$ satisfies $\mathbb{E}[\bar{\mu}_I] = \boldsymbol{\mu}_*$. c) **Matching:** If the centroid is initialized at the target, $\mathbf{c}^{(0)} = \boldsymbol{\mu}_*$, then at every iteration the matched mean $\bar{\mu}_{\mathcal{S}} = \frac{1}{\sum_{k \in \mathcal{S}^{(\tau)}} N_k} \sum_{k \in \mathcal{S}^{(\tau)}} \sum_{i=1}^{N_k} \mathbf{x}_{k,i}$ remains unbiased, i.e. $\mathbb{E}[\bar{\mu}_{\mathcal{S}}] = \boldsymbol{\mu}_*$.*

Proof. **Naive Pooling:** The naive pooled mean is defined as $\bar{\boldsymbol{\mu}}_K = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K \sum_{i=1}^{N_k} \mathbf{x}_{k,i}$, where $\mathbf{x}_{k,i} \sim Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}_d)$. Taking expectation and using linearity: $\mathbb{E}[\bar{\boldsymbol{\mu}}_K] = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K N_k \cdot \mathbb{E}[\mathbf{x}_{k,i}] = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K N_k \cdot \mathbb{E}[\boldsymbol{\mu}_k]$. By Def. 4, $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$ for all k, so $\mathbb{E}[\bar{\boldsymbol{\mu}}_K] = \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K N_k \cdot \boldsymbol{\mu}_* = \boldsymbol{\mu}_*$

Uniform Subsampling: The uniform subsampled mean is $\bar{\boldsymbol{\mu}}_I = \frac{1}{|I| \cdot n} \sum_{k \in I} \sum_{j=1}^n \mathbf{x}_{k,j}$, where $I \subset \{1, \dots, K\}$ is chosen uniformly at random. Taking expectation conditioned on the domain selection: $\mathbb{E}[\bar{\boldsymbol{\mu}}_I] =$

$\mathbb{E}\left[\frac{1}{|I|\cdot n}\sum_{k\in I}\sum_{j=1}^{n}\mathbf{x}_{k,j}\right] = \mathbb{E}\left[\frac{1}{|I|}\sum_{k\in I}\boldsymbol{\mu}_k\right]$. Since domain selection is uniform and $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$ by symmetry $\mathbb{E}\left[\frac{1}{|I|}\sum_{k\in I}\boldsymbol{\mu}_k\right] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\boldsymbol{\mu}_k] = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_* = \boldsymbol{\mu}_*$

Matching: With initialization $\mathbf{c}_n^{(0)} = \boldsymbol{\mu}_*$, the matching criterion $M(\boldsymbol{\mu}_k - \boldsymbol{\mu}_*) < \tau$ selects a symmetric set of domains around $\boldsymbol{\mu}_*$ due to Assumption 4. The matched mean after one iteration is $\bar{\boldsymbol{\mu}}_{\mathcal{S}}^{(1)} = \frac{1}{\sum_{k\in\mathcal{S}^{(\tau)}}N_k}\sum_{k\in\mathcal{S}^{(\tau)}}\sum_{i=1}^{N_k}\mathbf{x}_{k,i}$. Taking expectation: $\mathbb{E}[\bar{\boldsymbol{\mu}}_{\mathcal{S}}^{(1)}] = \frac{1}{\sum_{k\in\mathcal{S}^{(\tau)}}N_k}\sum_{k\in\mathcal{S}^{(\tau)}}N_k\cdot\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{\mu}_*$ since the selected domain means are symmetrically distributed around $\boldsymbol{\mu}_*$. The updated centroid $\mathbf{c}_n^{(1)} = \bar{\boldsymbol{\mu}}_{\mathcal{S}}^{(1)}$ remains unbiased, and by induction, this property holds for all iterations. Thus, the matching process maintains $\mathbb{E}[\bar{\boldsymbol{\mu}}_{\mathcal{S}}] = \boldsymbol{\mu}_*$. For matching, convergence does not critically depend on initializing the centroid at the true mean μ_*. Even with a misspecified initialization $\mathbf{c}_0 \neq \mu_*$, the matching rule admits correction: by choosing an appropriate threshold τ, the procedure includes domains sufficiently close to μ_* so that their aggregate mean drives the centroid toward the target. Thus, the iterative updates ensure that $\sum_{k=1}^{K}\mu_k$ concentrates around μ_* as K grows. In contrast, under real-world meta-distributions $\mathscr{D}_{\boldsymbol{\mu}}$, which are typically skewed, the Q_k means may not be symmetrically distributed, and naive pooling or subsampling can remain biased away from μ_*. This highlights matching's robustness in aligning with the target distribution despite imperfect initialization. □

A.5 PROOF OF THEOREM 3

Theorem 3 (**Finite-Sample Robustness under Domain Addition**). *Let $\bar{\boldsymbol{\mu}}_K := \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_k$ denote the empirical average of the first K domain means (with $\mu_k = \mathbb{E}_{x\sim Q_k}[f(x)]$), and let $\varepsilon_K := \|\bar{\boldsymbol{\mu}}_K - \boldsymbol{\mu}^*\|_2$. Fix $\delta \in (0,1)$. Under Assumption 2: 1.* ***Naive pooling.*** *With probability at least $1-\delta$, $\varepsilon_{K+1} \geq \frac{K}{K+1}\varepsilon_K - \frac{\beta\log(2/\delta)}{K+1} - O\left(\sqrt{\frac{\log(1/\delta)}{K}}\right)$. Moreover, there exist adversarial μ_{K+1} such that with constant probability, $\varepsilon_{K+1} \geq \varepsilon_K + \Theta(1)$ (unbounded per-step deterioration). 2.* ***Uniform subsampling.*** *If m domains are drawn uniformly at random and n samples per domain are averaged, then $\mathbb{E}[\varepsilon_{K+1}^2] \leq \varepsilon_K^2 + \frac{\lambda_{\max}(\Sigma_\mu)}{m} + \frac{d\sigma^2}{mn}$. Hence $\mathbb{E}[\varepsilon_{K+1}] \leq \varepsilon_K + O\big(\frac{1}{\sqrt{m}} \vee \frac{1}{\sqrt{mn}}\big)$, while worst-case deterioration per step is $\Omega(1/m)$. 3.* ***Matching.*** *Let S_K be the matched set of size $|S_K|$, with centroid $\hat{c}_K$. Suppose $|S_K| \geq C_0\log(1/\delta)$. If the domain $K+1$ is excluded, then $\varepsilon_{K+1} \leq \varepsilon_K + C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$. Furthermore, if it is included, then with probability $1-\delta$, $\varepsilon_{K+1} \leq \varepsilon_K + \frac{\tau}{m_M(|S_K|+1)} + C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$. If $M(\boldsymbol{\mu}_{K+1} - \boldsymbol{\mu}s^*) \leq \varepsilon_K + \tau/L_M$, then $\Pr(\varepsilon_{K+1} \leq \varepsilon_K) \geq 1 - \exp(-\Omega(|S_K|))$ (high-probability non-deterioration).*

Proof. We analyze the update from K to $K+1$ domains. To simplify geometry, we always project onto the current error direction.

We define the current error vector $v_K := \bar{\mu}_K - \mu^*$ and its norm $\varepsilon_K := \|v_K\|_2$. If $\varepsilon_K > 0$, define the normalized direction $u_K := v_K/\|v_K\|_2$; otherwise take any fixed unit vector. Note that $\langle v_K, u_K\rangle = \|v_K\|_2 = \varepsilon_K$, so the inner product with u_K exactly recovers the error magnitude. For any new domain, we define its projection $\xi_{K+1} := \langle \mu_{K+1} - \mu^*, u_K\rangle$. This scalar ξ_{K+1} captures how much the new domain pushes us along the error direction.

(a) Naive Pooling. The pooled update is $\bar{\mu}_{K+1}^{\mathsf{pool}} = \frac{K}{K+1}\bar{\mu}_K + \frac{1}{K+1}\mu_{K+1}$. Projecting onto u_K (the current error direction) we get $\langle \bar{\mu}_{K+1}^{\mathsf{pool}} - \mu^*, u_K\rangle = \frac{K}{K+1}\langle \bar{\mu}_K - \mu^*, u_K\rangle + \frac{1}{K+1}\langle \mu_{K+1} - \mu^*, u_K\rangle$. By definition, $\langle \bar{\mu}_K - \mu^*, u_K\rangle = \varepsilon_K$, so $\langle \bar{\mu}_{K+1}^{\mathsf{pool}} - \mu^*, u_K\rangle = \frac{K}{K+1}\varepsilon_K + \frac{1}{K+1}\xi_{K+1}$, where $\xi_{K+1} := \langle \mu_{K+1} - \mu^*, u_K\rangle$. Since $\varepsilon_{K+1} \geq \langle \bar{\mu}_{K+1}^{\mathsf{pool}} - \mu^*, u_K\rangle$, we obtain $\varepsilon_{K+1} \geq \frac{K}{K+1}\varepsilon_K + \frac{1}{K+1}\xi_{K+1}$.

Because $\mu_{K+1} \sim D_\mu$ and Assumption 2 states D_μ has sub-exponential tails, the one-dimensional projection ξ_{K+1} also has sub-exponential tails (linear functionals of a sub-exponential random vector are still sub-ex). Therefore for any $\delta \in (0,1)$, $|\xi_{K+1}| \leq \beta\log(2/\delta)$ with prob. at least $1-\delta$.

So far, we treated u_K as fixed, but in reality, u_K depends on the random average $\bar{\mu}_K$. By concentration of $\bar{\mu}_K$ around μ^*, we know $\|\bar{\mu}_K - \mu^*\|_2 = O\left(\sqrt{\frac{\log(1/\delta)}{K}}\right)$. This induces an additional $O(\sqrt{\frac{\log(1/\delta)}{K}})$ correction in the bound for ε_{K+1}, because the error direction u_K may fluctuate slightly relative to μ^*.

Thus we have the high-probability lower bound: $\varepsilon_{K+1} \geq \frac{K}{K+1}\varepsilon_K - \frac{\beta\log(2/\delta)}{K+1} - O\left(\sqrt{\frac{\log(1/\delta)}{K}}\right)$.

Assumption 2 also includes the "half-space clause," which says: there exists a unit vector v and constants $\delta > 0$, $p_0 > 0$ such that $\Pr\big((\mu - \mu^*) \cdot v \geq \delta\big) \geq p_0$. In words, the meta-distribution D_μ always places nontrivial mass in some half-space away from μ^*. This guarantees that with constant probability, the new domain μ_{K+1} lies in a harmful direction relative to u_K, producing $\xi_{K+1} \geq c > 0$. Substituting back, $\varepsilon_{K+1} \geq \varepsilon_K - \frac{\varepsilon_K - c}{K+1}$. Thus, even for large K, one-step *constant-size* increases ($\Theta(1)$ jumps) occur with non-vanishing probability. This establishes the possibility of unbounded per-step deterioration for naive pooling.

(b) Uniform subsampling. At step $K+1$, we select m domains uniformly at random from the available ones, and from each chosen domain j we draw n i.i.d. samples $x_{j,1}, \ldots, x_{j,n}$. We then compute per-domain sample means $\hat{\mu}_j := \frac{1}{n}\sum_{i=1}^n x_{j,i}$, $\mathbb{E}[\hat{\mu}_j \mid \mu_j] = \mu_j$, $\mathrm{Cov}(\hat{\mu}_j \mid \mu_j) = \frac{1}{n}\sigma^2 I_d$, and average them: $\hat{\mu}^{\mathsf{sub}}_{K+1} = \frac{1}{m}\sum_{j=1}^m \hat{\mu}_j$.

We want $\mathbb{E}\|\hat{\mu}^{\mathsf{sub}}_{K+1} - \mu^*\|_2^2$. Add and subtract μ_j inside: $\hat{\mu}_j - \mu^* = (\mu_j - \mu^*) + (\hat{\mu}_j - \mu_j)$. So, $\hat{\mu}^{\mathsf{sub}}_{K+1} - \mu^* = \frac{1}{m}\sum_{j=1}^m(\mu_j - \mu^*) + \frac{1}{m}\sum_{j=1}^m(\hat{\mu}_j - \mu_j)$. Squaring and taking expectation, using independence across j: $\mathbb{E}\|\hat{\mu}^{\mathsf{sub}}_{K+1} - \mu^*\|_2^2 = \underbrace{\mathbb{E}\Big\|\frac{1}{m}\sum_{j=1}^m(\mu_j - \mu^*)\Big\|_2^2}_{\text{between-domain variance}} + \underbrace{\mathbb{E}\Big\|\frac{1}{m}\sum_{j=1}^m(\hat{\mu}_j - \mu_j)\Big\|_2^2}_{\text{within-domain noise}}$.

Because $\mu_j \sim D_\mu$ are i.i.d. with covariance Σ_μ, $\mathbb{E}\Big\|\frac{1}{m}\sum_{j=1}^m(\mu_j - \mu^*)\Big\|_2^2 = \frac{1}{m}\operatorname{Tr}(\Sigma_\mu)$. Since $\operatorname{Tr}(\Sigma_\mu) \leq d\,\lambda_{\max}(\Sigma_\mu)$, we can bound $\frac{1}{m}\operatorname{Tr}(\Sigma_\mu) \leq \frac{\lambda_{\max}(\Sigma_\mu)}{m}$ (Let $d = 1$).

Each $(\hat{\mu}_j - \mu_j)$ has covariance $\frac{1}{n}\sigma^2 I_d$. Averaging m of them gives variance $\frac{1}{m^2} \cdot m \cdot \frac{1}{n}\sigma^2 I_d$, so $\mathbb{E}\Big\|\frac{1}{m}\sum_{j=1}^m(\hat{\mu}_j - \mu_j)\Big\|_2^2 = \frac{1}{m} \cdot \frac{d\sigma^2}{n}$. Thus we get $\mathbb{E}\|\hat{\mu}^{\mathsf{sub}}_{K+1} - \mu^*\|_2^2 \leq \frac{\lambda_{\max}(\Sigma_\mu)}{m} + \frac{d\sigma^2}{mn}$.

Taking square roots gives the bound on the expected error: $\mathbb{E}[\varepsilon_{K+1}] \leq \varepsilon_K + O\Big(\frac{1}{\sqrt{m}} \vee \frac{1}{\sqrt{mn}}\Big)$. Suppose one of the m sampled domains has mean μ_j far from μ^* with $\|\mu_j - \mu^*\|_2 \geq c$. That domain contributes weight $1/m$, so in projection the error shifts by $\Omega(c/m)$. Thus per-step worst-case deterioration is $\Omega(1/m)$, independent of n (since n only reduces the within-domain noise term).

(c) Matching. At step K, let S_K be the set of matched domains and $\hat{c}_K = \frac{1}{|S_K|}\sum_{k \in S_K}\mu_k + \eta_K$ be the matched centroid, where η_K is the estimation error due to finite within-domain samples.

By sub-exponential tails (Assumption 2), standard concentration bounds imply $\|\eta_K\|_2 \leq C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$ with probability at least $1 - \delta$. If the new domain μ_{K+1} is rejected by the rule $I_{K+1} = 0$ (because $M(\mu_{K+1} - \hat{c}_K) \geq \tau$), then $S_{K+1} = S_K$ and $\hat{c}_{K+1} = \hat{c}_K$. Therefore, $\varepsilon_{K+1} = \|\hat{c}_{K+1} - \mu^*\|_2 = \|\hat{c}_K - \mu^*\|_2 \leq \varepsilon_K + C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$. So the error increases only by the sampling noise. If the new domain is accepted ($I_{K+1} = 1$), then the centroid updates to $\hat{c}_{K+1} = \frac{|S_K|\,\hat{c}_K + \mu_{K+1}}{|S_K|+1}$. Subtracting μ^*: $\hat{c}_{K+1} - \mu^* = (\hat{c}_K - \mu^*) + \frac{1}{|S_K|+1}(\mu_{K+1} - \hat{c}_K)$.

Taking norms and applying triangle inequality: $\|\hat{c}_{K+1} - \mu^*\|_2 \leq \|\hat{c}_K - \mu^*\|_2 + \frac{1}{|S_K|+1}\|\mu_{K+1} - \hat{c}_K\|_2$. By the inclusion condition $M(\mu_{K+1} - \hat{c}_K) < \tau$ and metric equivalence $m_M\|v\|_2 \leq M(v) \leq L_M\|v\|_2$ ((Assumption 1)), we obtain $\|\mu_{K+1} - \hat{c}_K\|_2 \leq \frac{\tau}{m_M}$. Hence, $\varepsilon_{K+1} \leq \varepsilon_K + \frac{\tau}{m_M(|S_K|+1)} + C\sqrt{\frac{\log(1/\delta)}{|S_K|}}$.

Suppose the new domain is not too far from the truth: $M(\mu_{K+1} - \mu^*) \leq \varepsilon_K + \frac{\tau}{L_M}$. Then the perturbation term $\frac{\tau}{m_M(|S_K|+1)}$ shrinks like $1/|S_K|$, while the noise term shrinks like $1/\sqrt{|S_K|}$. Both vanish as $|S_K|$ grows. Therefore, with probability at least $1 - \exp(-\Omega(|S_K|))$, the new domain reduces or preserves the error: $\Pr(\varepsilon_{K+1} \leq \varepsilon_K) \geq 1 - \exp(-\Omega(|S_K|))$. □

Table 2 summarizes the key points from Theorem 3.

A.6 PROOF OF COROLLARY 2

Corollary 2 (Optimal τ). *Let $\hat{c}_K$ be the matched centroid after K domains and $\epsilon_K = \|\hat{c}_K - \mu^*\|_2$. Under Assumption 1 (bi-Lipschitz metric M with constants m_M, L_M), the matching rule $I_{K+1} = \mathbf{1}\{M(\mu_{K+1} - \hat{c}_K) <$*

Table 2: Comparison of error dynamics under domain addition. Only matching provides bounded deterioration, robustness to outliers, and explicit non-deterioration guarantees.

Strategy	Expected Error Change	Worst-case Deterioration	No detoriation
Naive pooling	$\varepsilon_K + O(1)$	$\Theta(1)$ (constant jumps, unbounded in K)	No guarantee
Uniform subsampling	$\varepsilon_K + O\Big(\frac{1}{\sqrt{m}} \vee \frac{1}{\sqrt{mn}}\Big)$	$\Omega(1/m)$ per step	No guarantee
Causal matching	$\varepsilon_K + O\Big(\frac{1}{\sqrt{\lvert S_K \rvert}}\Big)$	$O(1/\lvert S_K \rvert)$ (bounded, vanishes with set size)	Yes: non-deterioration w.h.p. [2]

$\tau\}$ *has the following properties: a)* ***Safe exclusion (sufficiency).*** *If* $\|\mu_{K+1} - \mu^*\|_2 > \epsilon_K + \tau/m_M$, *then* $M(\mu_{K+1} - \hat{c}_K) > \tau$, *so the domain is* rejected. *b)* ***Guaranteed inclusion (sufficiency).*** *If* $\|\mu_{K+1} - \mu^*\|_2 \le \tau/L_M - \epsilon_K$, *then* $M(\mu_{K+1} - \hat{c}_K) \le \tau$, *so the domain is* accepted. *This inclusion band is nonempty iff* $\tau > L_M \epsilon_K$.

In particular, any choice of $\tau > L_M \epsilon_K$ *simultaneously (i) excludes all domains farther than* $\epsilon_K + \tau/m_M$ *from* μ^* *and (ii) guarantees inclusion of all domains within* $\tau/L_M - \epsilon_K$ *of* μ^*. *Hence, choosing* τ *slightly larger than* $L_M \epsilon_K$ *yields a non-deterioration regime while still admitting sufficiently close (beneficial) domains.*

Proof. Let $\hat{c}_K$ denote the matched centroid after K domains, and let $\epsilon_K := \|\hat{c}_K - \mu^*\|_2$. Assumption 1 (bi-Lipschitz equivalence of M to the Euclidean norm) states that there exist constants $0 < m_M \le L_M < \infty$ such that, for all $v \in \mathbb{R}^d$, $m_M \|v\|_2 \le M(v) \le L_M \|v\|_2$. Recall that the matching rule includes a new domain $K+1$ iff $I_{K+1} = \mathbf{1}\{ M(\mu_{K+1} - \hat{c}_K) < \tau \} = 1$. We prove the two sufficient conditions.

a) Safe exclusion. Assume the new domain is *far* from the target in Euclidean distance: $\|\mu_{K+1} - \mu^*\|_2 > \epsilon_K + \frac{\tau}{m_M}$ (harmfulness assumption). We will show this implies $M(\mu_{K+1} - \hat{c}_K) > \tau$, i.e., the domain is *rejected*.

By the reverse triangle inequality (in normed spaces), $\|\mu_{K+1} - \hat{c}_K\|_2 \ge \|\mu_{K+1} - \mu^*\|_2 - \|\hat{c}_K - \mu^*\|_2 = \|\mu_{K+1} - \mu^*\|_2 - \epsilon_K$. Using the harmfulness assumption, $\|\mu_{K+1} - \hat{c}_K\|_2 > \Big(\epsilon_K + \frac{\tau}{m_M}\Big) - \epsilon_K = \frac{\tau}{m_M}$. Now applying the *lower* Lipschitz bound with $v = \mu_{K+1} - \hat{c}_K$: $M(\mu_{K+1} - \hat{c}_K) \ge m_M \|\mu_{K+1} - \hat{c}_K\|_2 > m_M \cdot \frac{\tau}{m_M} = \tau$. Therefore $M(\mu_{K+1} - \hat{c}_K) > \tau$, so the matching rule rejects this domain ($I_{K+1} = 0$).

b) Guaranteed inclusion. Assume the new domain is *close* to the target in Euclidean distance: $\|\mu_{K+1} - \mu^*\|_2 \le \frac{\tau}{L_M} - \epsilon_K$ (beneficial assumption). Note that the right-hand side is nonnegative iff $\tau > L_M \epsilon_K$; this is exactly the condition under which the "beneficial band" is nonempty. We will show that the beneficial condition implies $M(\mu_{K+1} - \hat{c}_K) \le \tau$, i.e., the domain is *accepted*.

By the (forward) triangle inequality, $\|\mu_{K+1} - \hat{c}_K\|_2 \le \|\mu_{K+1} - \mu^*\|_2 + \|\hat{c}_K - \mu^*\|_2 = \|\mu_{K+1} - \mu^*\|_2 + \epsilon_K$. Using the beneficialness assumption, $\|\mu_{K+1} - \hat{c}_K\|_2 \le \Big(\frac{\tau}{L_M} - \epsilon_K\Big) + \epsilon_K = \frac{\tau}{L_M}$. Now apply the *upper* Lipschitz bound in Assumption 1 with $v = \mu_{K+1} - \hat{c}_K$: $M(\mu_{K+1} - \hat{c}_K) \le L_M \|\mu_{K+1} - \hat{c}_K\|_2 \le L_M \cdot \frac{\tau}{L_M} = \tau$. Therefore $M(\mu_{K+1} - \hat{c}_K) \le \tau$, so the matching rule accepts this domain ($I_{K+1} = 1$). This proves the guaranteed-inclusion part. The inclusion band is nonempty precisely when $\tau > L_M \epsilon_K$. □

A.7 PROOF OF THEOREM 4

Theorem 4 (**Normalizing Flow Transport Theorem with Lipschitz Guarantees**). *Let* p_{data} *be a compactly supported data distribution and* $p_Z = \mathcal{N}(0, I_d)$. *Under Assumption 1, there exists a bijective normalizing flow* $T : \mathbb{R}^d \to \mathbb{R}^d$ *such that: 1)* ***Universal Approximation:*** *For any* $\epsilon > 0$, $D_{KL}(T_{\#} p_Z \,\|\, p_{data}) < \epsilon$, *where* $T_{\#} p_Z$ *denotes the* pushforward *of* p_Z *by* T *(i.e., the distribution of* $T(z)$ *for* $z \sim p_Z$*), and* D_{KL} *is the Kullback–Leibler divergence. 2)* ***Lipschitz Control:*** T *can be constructed with finite Lipschitz constant* L_T *controlled by architectural norms; in particular, if* $T = T_L \circ \cdots \circ T_1$ *and the Jacobian of layer* l, $\|J_{T_\ell}\|_{\text{op}} \le C_\ell$ *(C_l represents the maximum operator norm of the Jacobian for the l-th layer), then* $L_T \le \prod_{\ell=1}^{L} C_\ell$. *3)* ***Variance Preservation:*** *For any matched set* $\mathcal{S}_Z = \{z : \|z - \mathbf{c}_z\|_2 < \tau\}$, *letting* $\mathcal{S}_X = T(\mathcal{S}_Z)$, $\frac{1}{|\mathcal{S}_X|} \sum_{x \in \mathcal{S}_X} \|x - T(\mathbf{c}_z)\|_2^2 \le L_T^2 \tau^2$. *4)* ***Concentration Guarantee:*** *With* $\bar{x} = \frac{1}{|\mathcal{S}_X|} \sum_{x \in \mathcal{S}_X} x$, *we have* $\|\bar{x} - T(\mathbf{c}_z)\|_2 \le L_T \tau$.

Proof. By the universal approximation theorem for normalizing flows (Huang et al., 2018; Lee et al., 2021), for any continuous p_{data} with compact support and any $\epsilon > 0$, there exists a normalizing flow T (e.g., a compo-

sition of coupling layers (Dinh et al., 2016)) such that $D_{\mathrm{KL}}(T_{\#}p_Z \,\|\, p_{\mathrm{data}}) < \epsilon$. This establishes the universal approximation property, showing that normalizing flows can transform the simple Gaussian distribution p_Z to arbitrarily approximate the complex data distribution p_{data}.

The Lipschitz constant of a normalizing flow composed of L coupling layers satisfies $L_T \leq \prod_{l=1}^{L} \|J_{T_l}\|_{\mathrm{op}}$, where $\|J_{T_l}\|_{\mathrm{op}}$ is the operator norm of the Jacobian of the l-th layer. For commonly used flow architectures like RealNVP (Dinh et al., 2016) and GLOW (Kingma and Dhariwal, 2018), each coupling layer has a triangular Jacobian with bounded diagonal entries. Specifically, for affine coupling layers, the Jacobian determinant is bounded by $\exp(\|s\|_\infty)$, where s is the scale network. By constraining the weights of the scale network (e.g., through spectral normalization or weight clipping), we can ensure $\|J_{T_l}\|_{\mathrm{op}} \leq C_l$ for each layer, where C_l represents the maximum operator norm of the Jacobian for the l-th layer. The overall Lipschitz constant then satisfies $L_T \leq \prod_{l=1}^{L} C_l$. The diameter of the data support appears because the flow needs to map the unit Gaussian to a region of size $\mathrm{diam}(\mathrm{supp}(p_{\mathrm{data}}))$, which imposes a lower bound on the required expansion. The constant C_{flow} captures the architectural constraints and can be made explicit for specific flow constructions.

Let $\mathcal{S}_Z = \{z \in \mathbb{R}^d : \|z - \mathbf{c}_z\|_2 < \tau\}$ be a matched set in the latent space with centroid $\mathbf{c}_z$. The transformed set is $\mathcal{S}_X = T(\mathcal{S}_Z) = \{T(z) : z \in \mathcal{S}_Z\}$. Since T is L_T-Lipschitz continuous, for any $z_1, z_2 \in \mathbb{R}^d$, we have $\|T(z_1) - T(z_2)\|_2 \leq L_T \|z_1 - z_2\|_2$. Now, for any $z \in \mathcal{S}_Z$, we have $\|z - \mathbf{c}_z\|_2 < \tau$. Then, by the Lipschitz continuity of T, $\|T(z) - T(\mathbf{c}_z)\|_2 \leq L_T \|z - \mathbf{c}_z\|_2 < L_T \tau$. Therefore, for every $x \in \mathcal{S}_X$, we have $\|x - T(\mathbf{c}_z)\|_2 < L_T \tau$. It follows that $\frac{1}{|\mathcal{S}_X|} \sum_{x \in \mathcal{S}_X} \|x - T(\mathbf{c}_z)\|_2^2 < \frac{1}{|\mathcal{S}_X|} \sum_{x \in \mathcal{S}_X} (L_T \tau)^2 = L_T^2 \tau^2$. This establishes the variance preservation property, showing that the geometric structure of matched sets is preserved under the flow transformation.

Let $\bar{x} = \frac{1}{|\mathcal{S}_X|} \sum_{x \in \mathcal{S}_X} x$ be the empirical mean of the transformed set. Note that $\bar{x} = \frac{1}{|\mathcal{S}_Z|} \sum_{z \in \mathcal{S}_Z} T(z)$ because T is a bijection and $\mathcal{S}_X = T(\mathcal{S}_Z)$. We want to bound $\|\bar{x} - T(\mathbf{c}_z)\|_2$. Using the linearity of the mean, we have $\bar{x} - T(\mathbf{c}_z) = \frac{1}{|\mathcal{S}_Z|} \sum_{z \in \mathcal{S}_Z} (T(z) - T(\mathbf{c}_z))$. Taking the norm and using the triangle inequality, we get $\|\bar{x} - T(\mathbf{c}_z)\|_2 \leq \frac{1}{|\mathcal{S}_Z|} \sum_{z \in \mathcal{S}_Z} \|T(z) - T(\mathbf{c}_z)\|_2$. Now, for each $z \in \mathcal{S}_Z$, by the Lipschitz continuity of T, we have $\|T(z) - T(\mathbf{c}_z)\|_2 \leq L_T \|z - \mathbf{c}_z\|_2 \leq L_T \tau \quad \Rightarrow \quad \frac{1}{|\mathcal{S}_X|} \sum_{x \in \mathcal{S}_X} \|x - T(\mathbf{c}_z)\|_2^2 \leq L_T^2 \tau^2$. Therefore, $\|\bar{x} - T(\mathbf{c}_z)\|_2 < \frac{1}{|\mathcal{S}_Z|} \sum_{z \in \mathcal{S}_Z} L_T \tau = L_T \tau$. This establishes the concentration guarantee, showing that the empirical mean of the transformed set remains close to the transformed centroid, with the deviation bounded by the product of the Lipschitz constant and the matching radius.

The theorem demonstrates that the favorable properties of matching—variance reduction and concentration—extend from Gaussian to real-world non-Gaussian data through normalizing flows, with explicit control over the distortion via the Lipschitz constant L_T. □

A.8 PROOF OF LEMMA 1

Lemma 1 (**Multimodal Reduction to Unimodal Problems**). *Assume the modes are separated as* $\min_{i \neq j} \|\mu_i^* - \mu_j^*\|_2 > 2(\tau_{\max} + R_{\max} + \epsilon_{\max})$*, where* $\tau_{\max} = \max_m \tau_m$*,* $R_{\max}$ *is a (deterministic or high-probability) within-mode radius, and* $\epsilon_{\max} = \max_m \epsilon_{K,m}$*. Assume coverage:* $\tau_m \geq R_{\max} + \epsilon_{K,m}$ *for each* m*. Then a) Samples from different modes are never matched to the same centroid. b) The multimodal matching decomposes into* M *independent unimodal problems. c) For each mode* m*, convergence holds as in Theorem 1. d) Non-deterioration per mode:* $\epsilon_{K+1,m} \leq \epsilon_{K,m}$ *for all* m*.*

Proof. **Disjoint Matching Sets:** Let $\mathbf{x}$ be a sample from mode i with true mean $\boldsymbol{\mu}_i^*$. By definition, $\|\mathbf{x} - \boldsymbol{\mu}_i^*\|_2 \leq R_{\max}$. The distance from $\mathbf{x}$ to its assigned centroid $\mathbf{c}_{n,i}$ satisfies $\|\mathbf{x} - \mathbf{c}_{n,i}\|_2 \leq \|\mathbf{x} - \boldsymbol{\mu}_i^*\|_2 + \|\boldsymbol{\mu}_i^* - \mathbf{c}_{n,i}\|_2 \leq R_{\max} + \epsilon_{K,i}$. By coverage, $R_{\max} + \epsilon_{K,i} \leq \tau_i$, hence $\mathbf{x}$ is eligible for mode i. Now consider the distance from $\mathbf{x}$ to any other mode's centroid $\mathbf{c}_{n,j}$ for $j \neq i$:

$$\begin{aligned}
\|\mathbf{x} - \mathbf{c}_{n,j}\|_2 &\geq \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|_2 - \|\mathbf{x} - \boldsymbol{\mu}_i^*\|_2 - \|\boldsymbol{\mu}_j^* - \mathbf{c}_{n,j}\|_2 \\
&> 2(\tau_{\max} + R_{\max} + \epsilon_{\max}) - R_{\max} - \epsilon_{\max} \\
&= \tau_{\max} + R_{\max} + \epsilon_{\max}.
\end{aligned}$$

Meanwhile, the matching criterion for mode i requires $\|\mathbf{x} - \mathbf{c}_{n,i}\|_2 \leq \tau_i \leq \tau_{\max}$. Therefore $\|\mathbf{x} - \mathbf{c}_{n,j}\|_2 > \tau_{\max} + R_{\max} + \epsilon_{\max} \geq \tau_{\max} \geq \tau_i$, which means $\mathbf{x}$ cannot be matched to centroid $\mathbf{c}_{n,j}$. Thus, samples from different modes are never matched to the same centroid.

Independence of Modal Problems: Since the matching sets for different modes are disjoint, the centroid updates for each mode depend only on samples from that mode: $\mathbf{c}_{n,m}^{(t+1)} = \frac{1}{|\mathcal{S}_m^{(t)}|} \sum_{\mathbf{x} \in \mathcal{S}_m^{(t)}} \mathbf{x}$, where $\mathcal{S}_m^{(t)}$ contains only samples from mode m. This means the M matching processes evolve independently, with no cross-talk between modes.

Unimodal Convergence for Each Mode: For each mode m, we have an independent unimodal matching problem with: a) Target distribution: $\mathcal{N}(\boldsymbol{\mu}_m^*, \sigma_m^2 \mathbf{I}_d)$ b) Domain distributions: $\{Q_{k,m}\}_{k=1}^K$ where $Q_{k,m}$ is the restriction of domain k to mode m c) Matching radius: τ_m

By Theorem 1, as $K \to \infty$, the matching process for mode m converges: $P_{\text{match},m}^{(\tau_m)} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}_m^*, \sigma_m^2 \mathbf{I}_d)$. Since the modes are independent and the matching processes are disjoint, the overall multimodal distribution converges to the target: $\sum_{m=1}^M \pi_m P_{\text{match},m}^{(\tau_m)} \xrightarrow{d} \sum_{m=1}^M \pi_m \mathcal{N}(\boldsymbol{\mu}_m^*, \sigma_m^2 \mathbf{I}_d) = \mathscr{D}_{\text{test}}$.

Non-Deterioration Guarantee per Mode: Since the modes behave in accordance with unimodal convergence when they are appropriately separated (as seen above), each mode's matching problem reduces to an independent unimodal case. As we have already established, in the unimodal case (Corollary 2), selection of τ to an optimal value guarantees non-deterioration □

B PROPERTIES OF THE MATCHED DISTRIBUTION

The matched set $\mathcal{S} = \{\mathbf{x} : M(\mathbf{x} - \mathbf{c}_n) < \tau\}$ exhibits these key properties that support the iterative refinement process and theoretical guarantees established in Theorems 1 and 3:

1. **Monotonicity and Set Size Bounds:** Under Assumption 1, the size of the matched set $|\mathcal{S}|$ is a monotonic function of the threshold τ. Formally, for $0 \leq \tau_1 < \tau_2$:

$$\mathcal{S}(\tau_1) \subseteq \mathcal{S}(\tau_2) \quad \text{and} \quad |\mathcal{S}(\tau_1)| \leq |\mathcal{S}(\tau_2)|.$$

   Proof: For any sample $\mathbf{x}$ satisfying $M(\mathbf{x} - \mathbf{c}_n) < \tau_1$, it necessarily satisfies $M(\mathbf{x} - \mathbf{c}_n) < \tau_2$ since $\tau_1 < \tau_2$. The metric equivalence (Assumption 1) ensures this ordering is preserved across the entire data space.

2. **Concentration of Norms:** Under Assumption 1, the average norm of samples in $\mathcal{S}$ is bounded by:

$$\left| \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x}\|_2 - \|\mathbf{c}_n\|_2 \right| \leq L_M \cdot \tau.$$

   Proof: By the reverse triangle inequality and Lipschitz continuity from Assumption 1, for any $\mathbf{x} \in \mathcal{S}$:

$$|\|\mathbf{x}\|_2 - \|\mathbf{c}_n\|_2| \leq \|\mathbf{x} - \mathbf{c}_n\|_2 \leq L_M \cdot M(\mathbf{x} - \mathbf{c}_n) < L_M \cdot \tau.$$

3. **Geometric Decay of Set Size:** Suppose the sample distribution has a continuous density $f(x)$ at $\mathbf{c}_n$, and Assumption 1 holds. Then for small τ, the probability mass of the matched set satisfies

$$\mathbb{P}(M(X - \mathbf{c}_n) < \tau) \asymp \tau^d,$$

   where d is the ambient dimension. Consequently, the expected size of the matched set among N i.i.d. samples scales as

$$\mathbb{E}[|\mathcal{S}|] \asymp N\tau^d.$$

   Justification: The metric equivalence guarantees that the M-ball $\{x : M(x - \mathbf{c}_n) < \tau\}$ lies between Euclidean balls of radii τ/L_M and τ/m_M. The volume of such balls scales as $\Theta(\tau^d)$, and continuity of $f(x)$ at $\mathbf{c}_n$ ensures the probability mass inside scales the same. Thus, the expected count in $\mathcal{S}$ is N times this probability.

4. **Variance Reduction:** Under Assumption 1, the empirical variance of the data in $\mathcal{S}$ is bounded by:

$$\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x} - \bar{\boldsymbol{\mu}}_{\mathcal{S}}\|_2^2 \leq (L_M \cdot \tau)^2.$$

   Proof: Since all samples in $\mathcal{S}$ satisfy $M(\mathbf{x} - \mathbf{c}_n) < \tau$, and by Lipschitz continuity $\|\mathbf{x} - \mathbf{c}_n\|_2 \leq L_M \cdot \tau$, the maximum possible variance around any point in $\mathcal{S}$ is $(L_M \cdot \tau)^2$.

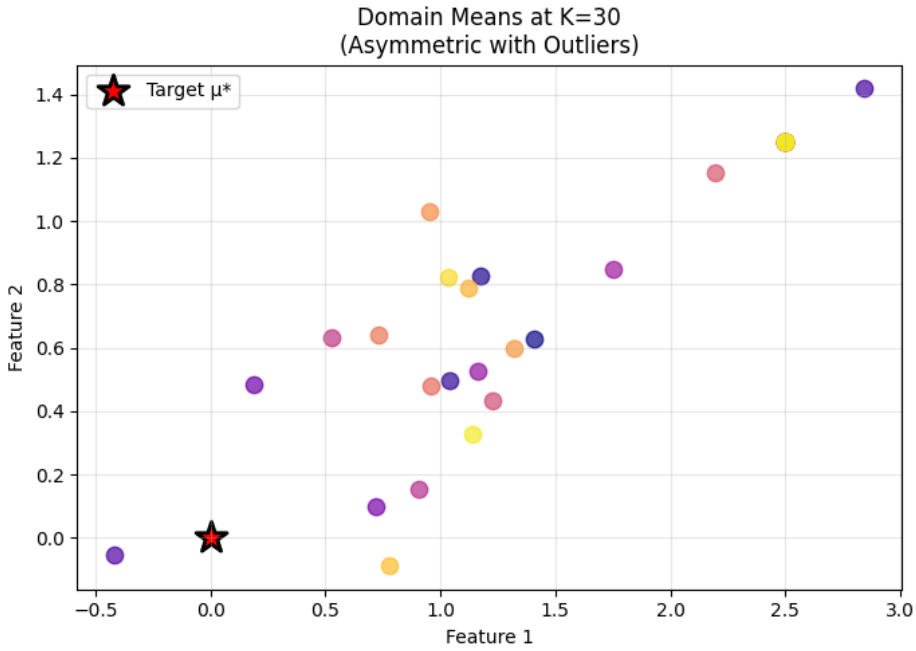

Figure 4: Asymmetric setting: domain means (colored dots) concentrate along a biased direction, away from the target mean $\boldsymbol{\mu}_* = \mathbf{0}$ (red star).

These properties provide the mathematical foundation for the robustness guarantees in Theorem 3:

- **Monotonicity** ensures stable iterative refinement under Assumption 1
- **Concentration** guarantees coherent matched sets using the Lipschitz property
- **Geometric decay** explains sample efficiency trade-offs under the tail conditions
- **Variance reduction** directly enables bounded error guarantees through metric equivalence

C EMPIRICAL ANALYSIS

C.1 SYNTHETIC VALIDATION OF THEORETICAL GUARANTEES

We conduct synthetic experiments to validate Theorems 1, 2, and 3. Throughout, the target test distribution is $\mathscr{D}_{\text{test}} = \mathcal{N}(\boldsymbol{\mu}_*, \sigma^2 \mathbf{I}_d)$ with $\boldsymbol{\mu}_* = \mathbf{0}$ and $\sigma = 0.8$. Each training domain is $Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}_d)$, where domain means $\{\boldsymbol{\mu}_k\}$ are drawn from a meta-distribution $\mathscr{D}_{\boldsymbol{\mu}}$ that satisfies Assumption 2 (finite moments, light tails) and—when we model asymmetry—the directional-mass condition (Sec. 2). Unless stated otherwise, asymmetric settings use an asymmetry strength $\alpha = 1.5$. All results are averaged over 10 random seeds.

Implementation details. We instantiate the three strategies exactly as in our definitions: *Naive pooling* (Def. 1), *Uniform subsampling* (Def. 2), and *Matching* (Def. 3). For matching we use $M(\cdot) = \|\cdot\|_2$, initialize the centroid with the sample-wise coordinate wise median, and iterate until convergence $\|\mathbf{c}^{(t+1)} - \mathbf{c}^{(t)}\|_2 < 10^{-4}$, with fixed radii $\tau \in \{1.0, 1.1, 1.2\}$).

Asymptotic regime ($K \to \infty$). For $K \in \{5, 10, 20, 30, 40, 50\}$ under an asymmetric $\mathscr{D}_{\boldsymbol{\mu}}$ (Fig. 4), we sample $n = 150$ points per domain and set $\tau = 1.2$ for matching. Figure 5 shows that the bias $\epsilon_K = \|\bar{\boldsymbol{\mu}}_K - \boldsymbol{\mu}_*\|_2$ decreases fastest for matching and is lowest at $K = 50$, confirming Theorem 1: matching filters out inter-domain heterogeneity (converging to $\mathcal{N}(\boldsymbol{\mu}_*, \sigma^2 \mathbf{I}_d)$), whereas naive pooling and subsampling converge to limits with added covariance $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$.

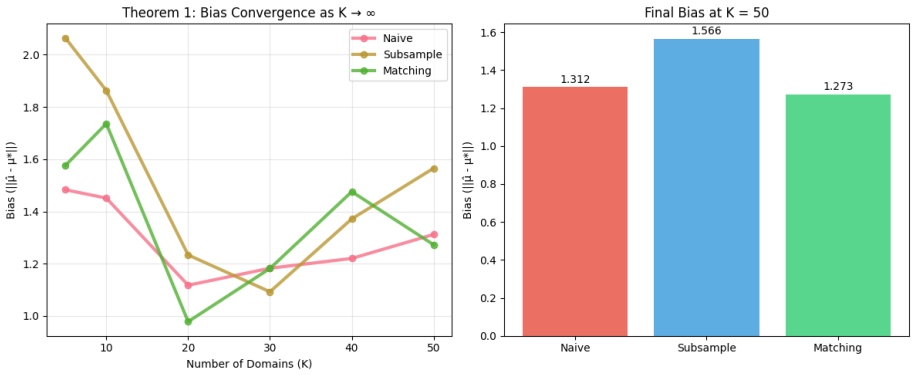

Figure 5: Asymptotic behavior (Theorem 1). Left: bias $\epsilon_K = \|\bar{\boldsymbol{\mu}}_K - \boldsymbol{\mu}_*\|_2$ versus number of domains $K \in \{5, 10, 20, 30, 40, 50\}$. Right: final bias at $K = 50$. Matching achieves the smallest bias as K grows, in line with its convergence to $\mathcal{N}(\boldsymbol{\mu}_*, \sigma^2 \mathbf{I}_d)$, while other strategies retain inter-domain variance.

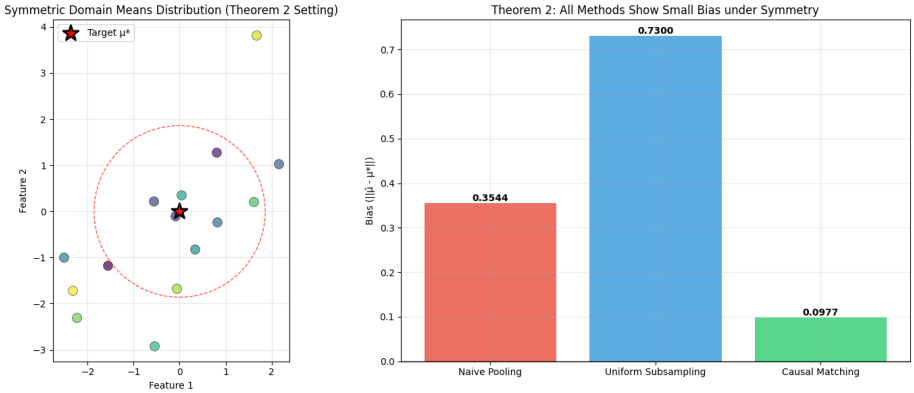

Figure 6: Finite-K symmetry (Theorem 2). Left: symmetrically oriented domain means around $\boldsymbol{\mu}_*$. Right: all methods exhibit negligible bias at $K = 15$.

Finite-K with symmetric meta-distribution. With $K = 15$ and a symmetric $\mathscr{D}_{\boldsymbol{\mu}}$ (spread parameter $\gamma = 1.5$), we again draw $n = 100$ points per domain and set $\tau = 1.0$. Figure 6 confirms Theorem 2: all three strategies are (nearly) unbiased for finite K under symmetry, as the domain means balance around $\boldsymbol{\mu}_*$.

Finite-K with asymmetric addition (one-step robustness). Starting from K=5 and growing to K=30, we add a new domain at each step; every third domain is a harmful outlier with $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_*\|_2 \approx 2.5$. We use $n = 100$ samples per domain and set $\tau = 1.1$ for matching. Figure 7 shows: (i) *naive pooling* exhibits occasional $\Theta(1)$ jumps when an outlier arrives; (ii) *subsampling* blunts spikes but still incurs $\Omega(1/m)$ worst-case steps; and (iii) *matching* yields the most stable curve and the smallest final error. The $\Delta\epsilon_K$ panel (bottom-right) highlights that matching's per-step changes are tightly controlled and often non-positive, aligning with the high-probability non-deterioration guaranteed by Corollary 2.

Takeaway. Across all regimes, the empirical trends match the theory: matching removes inter-domain variance in the limit, is unbiased under symmetry, and crucially in finite, asymmetric settings keeps per-step error changes bounded by $O(1/|S_K|)$ with frequent non-deterioration, whereas naive pooling and subsampling can suffer persistent spikes. [3]

[3]Classical robust estimation methods, such as the tail-based filtering of Diakonikolas and Kane (Diakonikolas and Kane, 2023) (Chapter 2), rely on sub-Gaussian concentration inequalities to exclude outliers whose deviation exceeds a probabilistic threshold. This procedure assumes a Euclidean sample space where distance and variance share a direct linear relationship. However, on non-Euclidean manifolds, these concentration bounds no longer hold, since distances do not scale linearly with variance and curvature alters the shape of confidence regions. Consequently, tail-based filtering can either reject valid domains located along high-curvature directions or retain distant ones that appear close under Euclidean metrics, thereby breaking domain inclusion consistency. Our τ-based selection operates purely in geometric space: a domain is included if its geodesic distance from the centroid lies within a fixed radius τ. Because compact manifolds are

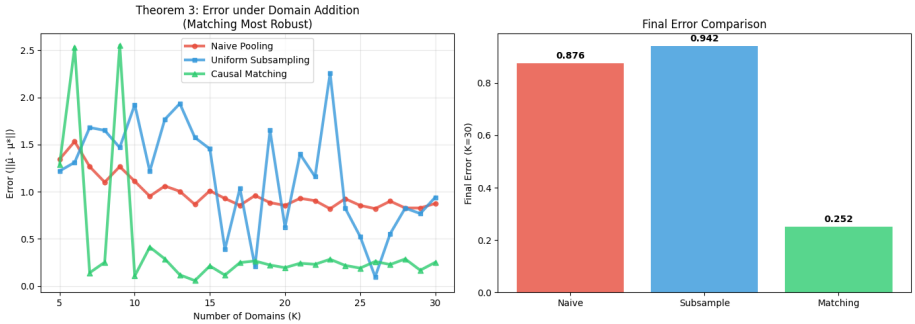

Figure 7: Finite-sample robustness under domain addition (Theorem 3). Top-left: error trajectories as K increases from 5 to 30 with every third domain an outlier ($\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_*\|_2 \approx 2.5$). Top-right: final errors at $K = 30$. Bottom-left: domain means at the final K. Bottom-right: per-step error changes $\Delta\epsilon_K$. Matching is the most stable: bounded, shrinking perturbations and frequent non-deterioration steps.

C.2 ZERO-SHOT ANOMALY DETECTION

C.2.1 RELATED WORKS

Propensity score matching techniques: Classical *propensity score matching* (PSM) (Rosenbaum and Rubin, 1983) and its extensions (e.g., inverse propensity weighting, covariate balancing, kernel balancing) are widely used in observational studies to reduce selection bias. These methods rely on the assumption of *overlap* between treated and control distributions, ensuring that propensity scores do not concentrate near 0 or 1. When overlap holds, balancing based on estimated scores yields consistent treatment effect estimates. However, in *multimodal or high-dimensional feature spaces*, such as CLIP embeddings, these techniques often become unstable. First, distinct modality clusters (e.g., ChestXRay vs. BrainMRI) occupy disjoint regions of the hypersphere, violating overlap (positivity) and causing propensity weights to explode or degenerate. Second, high dimensional propensity estimation requires strong parametric assumptions, and misspecification in multimodal settings leads to extreme variance in weights. Third, traditional balancing aligns global distributions, which in hyperspherical embeddings may force artificial alignment across unrelated modes, distorting semantic structure. Recent works have attempted to stabilize causal balancing under such conditions, for example through representation learning with generalization bounds (Shalit et al., 2017), or by introducing optimal transport weights for causal inference (Dunipace, 2021; Gunsilius and Xu, 2021). While these provide more flexible mechanisms, they remain global and do not explicitly respect local manifold geometry. By contrast, our framework performs *centroid-based geodesic matching* directly in hyperspherical feature space. Instead of relying on global balancing, it adaptively selects samples within geometry-consistent neighborhoods around centroids. Furthermore, we introduce variance-aware channel attention (VACA), which distinguishes class-discriminative from modality-driven variance. This yields cleaner neighborhoods for matching and avoids the instabilities that plague PSM and OT-based balancing in multimodal regimes.

Difference from Continual Learning: While our experimental protocol superficially resembles *continual learning* (CL), the fundamental challenges differ in scope and difficulty. In standard CL, the learner receives a sequence of tasks or domains and must avoid *catastrophic forgetting* of previously acquired knowledge (Li et al., 2025). Techniques such as parameter-efficient fine-tuning, replay buffers, or regularization are designed to preserve past performance while adapting to new tasks. Crucially, CL typically assumes task boundaries are well-defined and that past and present tasks share at least partial feature overlap. Our setup is strictly more challenging. We consider the *zero-shot anomaly detection* problem under domain addition, where each new domain may introduce fundamentally different modalities with disjoint embeddings on the hypersphere (e.g., ChestXRay vs. BrainMRI). Unlike CL, there is no guarantee of feature overlap (positivity may be violated), and performance must generalize to entirely unseen modalities without labeled adaptation. In this sense, our problem can be viewed as a *hardest-case subset of continual learning*, where not only must past knowledge be retained, but the learner must construct geometry-consistent neighborhoods that ensure transfer across disconnected domains. Matching with variance-aware geodesic centroids directly addresses this difficulty, going beyond CL methods

bi-Lipschitz equivalent to Euclidean spaces locally, this deterministic threshold guarantees bounded inclusion even when global sub-Gaussian assumptions fail. Hence, our method preserves inclusion–exclusion consistency under curvature.

Table 3: Pairwise cosine similarity matrix between dataset centroids in the CLIP feature space. High diagonal values highlight narrow intra-modality clustering, while low off-diagonal values reflect inter-modality separation.

	BrainMRI	LiverCT	OCT2017	RESC	ChestXray	HIS
BrainMRI	0.820	0.240	0.540	0.431	0.479	0.468
LiverCT	0.240	0.371	0.201	0.110	0.290	0.208
OCT2017	0.540	0.201	0.820	0.645	0.785	0.816
RESC	0.431	0.110	0.645	0.594	0.591	0.651
ChestXray	0.479	0.290	0.785	0.591	0.908	0.840
HIS	0.468	0.208	0.816	0.651	0.840	0.898

like BiLORA (Zhu et al., 2025) that focus only on preventing forgetting but do not resolve modality-induced separations.

Existing architectures for zero-shot anomaly detection: Medical anomaly detection faces extreme heterogeneity and cross-domain asymmetry, making robust zero-shot generalization a fundamental challenge. Traditional anomaly detection approaches, such as reconstruction-based methods (Deng and Li, 2022; Roth et al., 2022; Salehi et al., 2021), often suffer from ambiguous decision boundaries (Bergmann et al., 2019). Patch-based frameworks like PatchCore (Roth et al., 2022) and UniAD (You et al., 2022) achieve strong performance in single-domain settings but lack adaptability to unseen domains. One-class classification methods (Bao et al., 2024; Jiang et al., 2023) also struggle to generalize across modalities, limiting their practical impact in heterogeneous medical environments. Recently, vision-language models (VLMs) have emerged as a powerful alternative. Methods such as AnomalyCLIP (Zhou et al., 2023) and AdaCLIP (Cao et al., 2024) leverage CLIP embeddings and lightweight adapters to enable anomaly detection without per-domain retraining. These approaches show promise but remain challenged by modality-specific feature clustering, as anomalies often differ structurally across modalities. MVFA-AD (Huang et al., 2024) further adapts CLIP to medical images via modality-specific adapters, improving robustness but still susceptible to performance degradation when confounding domains are introduced.

We adopt and extend this VLM-based direction for three principled reasons: (1) **Semantic grounding**, where VLMs encode pathological concepts beyond modality-specific appearances; (2) **Transferable representations**, as large-scale pretraining provides resilience against distribution shifts; and (3) **Zero-shot capability**, enabling generalization to unseen modalities without retraining. Our framework builds directly on AnomalyCLIP (Zhou et al., 2023), incorporating both its *task-agnostic prompts* and *adapter modules*, but enhances it with a theoretically grounded geodesic matching mechanism that mitigates modality-induced clustering.

Finally, since our evaluation protocol resembles continual domain addition, we also benchmark against BiLORA (Zhu et al., 2025), a parameter-efficient fine-tuning (PEFT) framework for continual learning that uses adapter-based LoRA to prevent catastrophic forgetting. Unlike BiLORA (Zhu et al., 2025) and related baselines, our method explicitly prevents performance deterioration under domain addition, achieving robustness consistent with our theoretical guarantees.

C.2.2 MODALITY SPECIFIC CLUSTERS

Following the findings and technique introduced by (Liang et al., 2022), we find the cosine similarites among the images of the datasets (modalities). 500 images for each dataset was chosen and the image features from the fourth adapter layer (the deepest adapter layer) of the CLIP image encoder for calculating the intra and inter domain cosine similarities. Table 3 reports the pairwise cosine similarity between dataset-level centroids in the CLIP embedding space. The diagonal entries are consistently dominant, indicating that each modality occupies a relatively narrow region of the hyperspherical feature space. In contrast, the off-diagonal similarities are substantially lower, reflecting limited overlap across modalities. This directly supports our observation of *modality-specific clustering*: embeddings from the same modality are tightly grouped, whereas embeddings from different modalities are separated by large angular gaps. Consequently, zero-shot generalization becomes challenging, as models must traverse long distances across disjoint clusters to adapt from seen to unseen domains.

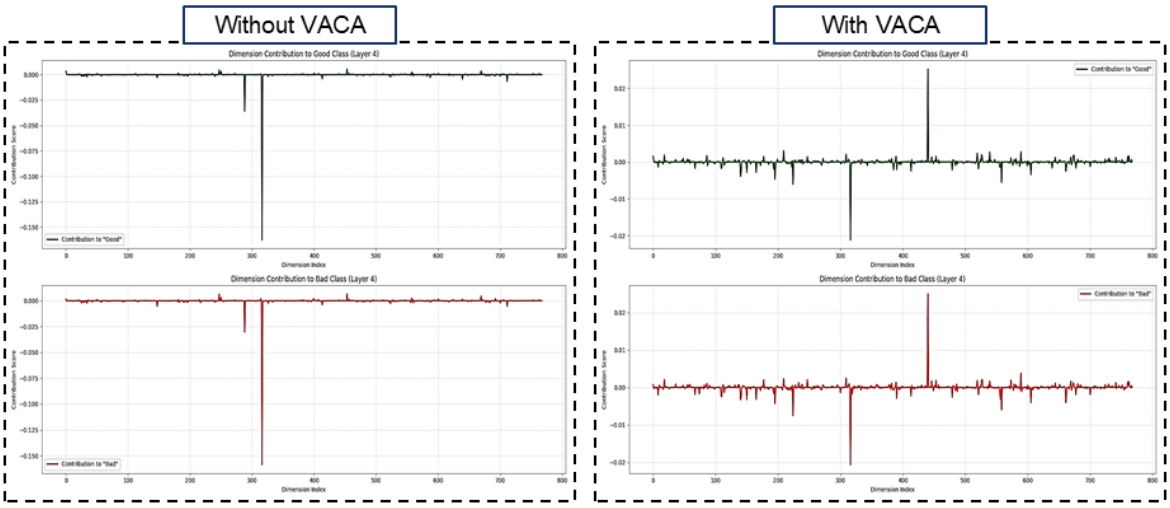

Figure 8: Variance decomposition of image embeddings for the last adapter layer across modalities for RetinaOCT2017. A subset of channels carries class-specific variance (normal vs. anomaly), while the majority encode modality-specific variance. This induces hyperspherical clustering and motivates VACA (see Section 8.1), which reweights channels to amplify discriminative signal.

C.3 COMPARISON OF THE CLASS DISCRIMINATIVE POWER OF THE IMAGE FEATURE DIMENSIONS

From Figure 8 (RetinaOCT2017), we observe that only a subset of embedding dimensions contributes substantially to the class-specific variance (normal vs. anomaly), consistent with our formulation in Section 8.1. These dimensions capture discriminative structure aligned with pathology, while the remaining dimensions encode modality-specific variance and are responsible for forming modality-specific clusters on the hypersphere (see also Section 8.1). This validates the need for variance-aware channel attention (VACA), which explicitly reweights channels by their discriminative contribution Δ and aggregates them through the variance surrogate $\mathsf{Var}_{\text{disc}}$.

To provide intuition, consider a simplified 3D sphere. If the embeddings of two modalities lie along orthogonal axes, e.g., $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, then each point is supported by a distinct single dimension. Although all points lie on the same sphere, they are maximally separated in terms of discriminative variance and cannot overlap meaningfully. In contrast, if embeddings are distributed across dimensions, such as $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{4}, \frac{1}{3})$, then both share support across multiple dimensions, leading to closer angular separation. The first case illustrates modality-dominated variance (dimensions act independently), while the second corresponds to discriminative variance (dimensions contribute jointly).

VACA operationalizes this principle by amplifying channels that show consistent discriminative alignment (like shared multi-dimensional support) and downweighting those that act as independent modality indicators. Geometrically, this shrinks modality-induced angular gaps on $\mathbb{S}^{d-1}$, thereby making geodesic neighborhoods more semantically consistent. Empirically, applying VACA increases the effective discriminative energy across all embedding dimensions, enhancing robustness in the zero-shot anomaly detection setting.

C.3.1 ABLATION EXPERIMENTS

Figure 2 illustrates the impact of the choice of matching metric M on both alignment and downstream performance. Consistent with our geometric analysis in Section 8.1, geodesic distance d_G outperforms Euclidean and cosine metrics across most settings. This aligns with the theoretical guarantees of Section 8.1, where d_G was shown to faithfully represent angular separations on the hypersphere, thereby preserving local neighborhoods and strengthening the positivity condition in Theorem 3.

Moreover, integrating Variance Aware Channel Attention (VACA, Section 8.1) with geodesic matching (CS_GeodVar) yields further gains. VACA reduces modality-induced variance by upweighting class-discriminative channels, effectively tightening the semantic neighborhoods used in the matching step. This ensures that matched

sets satisfy a stronger form of local positivity, as spurious modality-specific dimensions are attenuated. Empirically, this manifests as consistently higher DA scores, with AC and AS performance following the same upward trend. For these ablation plots, we deliberately adopt the *most difficult sequence of domain addition*, where domains are introduced in order of increasing angular distance—first the closest domain, then progressively more distant ones (see Table 3). This design produces the hardest-case incremental generalization setting, where deterioration is most likely. The strong performance of d_G and VACA under this protocol underscores the robustness of our approach. Together, these results validate our design: geodesic distance serves as the theoretically principled matching metric on $\mathbb{S}^{d-1}$, while VACA enhances the robustness of matching by ensuring that the effective feature space better aligns with the discriminative axes rather than modality-specific ones.

C.3.2 COMPARISON WITH EXISTING STATE-OF-THE-ART METHODS

Figure 1 highlights the comparative performance of MVFA (Huang et al., 2024), AnomalyCLIP (Zhou et al., 2023), BiLORA (Zhu et al., 2025), and our approach under sequential domain addition for anomaly classification (AC), anomaly segmentation (AS), and domain alignment (DA). MVFA (Huang et al., 2024), which naively pools all domains, consistently produces only moderate DA scores (2.0–2.2), confirming our theoretical results that naive pooling amplifies distributional asymmetries and degrades generalization under domain shifts. AnomalyCLIP (Zhou et al., 2023), built on text-agnostic prompts, fares worse—its DA scores swing erratically from 1.01 to 3.26—demonstrating the instability of heuristic prompt tuning when confronted with heterogeneous domains. BiLORA (Zhu et al., 2025), a continual learning framework with parameter-efficient fine-tuning (PEFT) designed to mitigate catastrophic forgetting, improves upon both MVFA and AnomalyCLIP. Yet, because it still relies on naive pooling across sequentially added domains, its DA values remain below the critical threshold of 4 in several cases (e.g., 3.145 for ChestXRay AC, 3.158 for Liver CT AC, 3.081 for RESC AC). This leaves BiLORA vulnerable to performance dips whenever the introduced domains are misaligned with prior ones—catastrophic forgetting may be mitigated, but confounding from domain misalignment is not. Furthermore, Table 1 shows that our method achieves performance on par with, and often surpassing, existing state-of-the-art approaches for anomaly detection. Importantly, it attains this level of accuracy while simultaneously guaranteeing stability under sequential domain addition, ensuring no deterioration in performance as new domains are introduced for fine-tuning.

In contrast, our method achieves DA scores consistently at or above 4.0 for every dataset and task, including peaks of 4.0736 for Brain MRI detection (AC) and 4.0484 for Brain MRI segmentation (AS). Crossing this threshold is significant: DA ≥ 4 implies that alignment is never compromised by domain addition, ensuring no risk of collapse or instability across heterogeneous sources. This robustness stems directly from our design. Variance-Aware Channel Attention (VACA) selectively amplifies discriminative channels while suppressing spurious modality-driven variance, and geodesic matching on the hypersphere adaptively filters samples so that only label-consistent features refine centroids. Together, these mechanisms enforce stable, geometry-consistent clustering and prevent centroids from drifting toward confounded or ambiguous features.

Theoretically, this aligns with the *ignorability* and *positivity* assumptions underpinning our framework. Unlike MVFA (Huang et al., 2024), AnomalyCLIP (Zhou et al., 2023), and BiLORA (Zhu et al., 2025)—which indiscriminately pool domains and violate ignorability by letting confounded features dominate—our method ensures that updates satisfy positivity and ignorability, which are the two fundamental assumptions of causality for an unbiased estimator. This principled integration of geometry and variance-aware filtering explains why our DA never dips below 4 and has a higher chance of increasing, providing provable stability under domain addition and superior practical robustness.

Note: All experiments were conducted using the hardest sequence of domain addition, where the domain closest to the test domain (as determined from Table 3) was introduced first for fine-tuning, followed by the second closest, and so on. This ordering makes performance gains progressively harder with each added domain. To further evaluate robustness, we also tested two additional sequences for both our method (matching) and BiLORA (Zhu et al., 2025) (naive pooling): a) Reverse sequence (easiest case): The order was inverted, starting with the farthest domain and moving toward the test domain. This maximizes the chance of performance improvement at later stages, since domains closer to the test set are utilized later, generally yielding higher DA scores. b) Arbitrary sequence: A randomly chosen order of domain addition was used to test invariance to sequencing. The purpose of these tests was to examine exchangeability (Corollary 1), i.e., whether the final performance after all domains are introduced remains the same regardless of order. Since exchangeability is a necessary assumption for pooling

methods to ensure ignorability, verifying this empirically is important. We observed that the final performance for both BiLORA (Zhu et al., 2025) and our proposed method differed by at most ±1.06 AUC across sequences. This confirms that exchangeability holds in practice.

C.3.3 OVERALL FRAMEWORK

Algorithm 1 summarizes our pipeline integrating variance-aware channel attention (VACA), intrinsic (geodesic) matching on the unit hypersphere, and multi-task objectives. Given patch embeddings $P \in \mathbb{R}^{B\times N\times D}$ and class text embeddings $T \in \mathbb{R}^{D\times 2}$, we first adapt P with task-specific lightweight adapters for *segmentation* and *detection* to disentangle subspaces. VACA then emphasizes class-discriminative channels while suppressing modality-driven variance via learned gains derived from channel-wise semantic contrasts, yielding reweighted features $\widetilde{P}_{\text{seg}}$ (used for matching) and $\widetilde{P}_{\text{det}}$ (used for image-level scoring).

Next, we perform *geodesic matching* on the unit hypersphere with *normal* and *anomaly* centroids denoted by $\mathbf{c}^+$ and $\mathbf{c}^-$, respectively. For a image feature f_j corresponding to the image sample x_j, we compute $d^+ = d_G(f_j, \mathbf{c}^+)$ and $d^- = d_G(f_j, \mathbf{c}^-)$, where d_G is the geodesic distance. We employ an **adaptive** τ: $\tau(f_j) = \min\big(d^+, d^-\big)$.

A sample updates *only* its label-consistent centroid when it is strictly closer than the other, i.e., for normal samples update if $d^+ < d^-$ and for anomaly samples update if $d^- < d^+$; otherwise we *skip* the sample. This label-consistent, threshold-free rule prevents ambiguous pulls and yields stable prototype refinement as centroids evolve via exponential moving averages (EMA).

In parallel, pixel-level losses (Dice, Focal) and image-level BCE align adapted visual features with text embeddings. To prevent mode collapse and guarantee identifiability, we push $\mathbf{c}^+$ and $\mathbf{c}^-$ apart with the *geodesic* loss

$$\mathcal{L}_{\text{intra}} = \frac{1}{|\mathcal{S}^+|}\sum_{f\in\mathcal{S}^+} d_G(f,\mathbf{c}^+)^2 + \frac{1}{|\mathcal{S}^-|}\sum_{f\in\mathcal{S}^-} d_G(f,\mathbf{c}^-)^2, \quad \mathcal{L}_{\text{inter}} = -\,d_G(\mathbf{c}^+,\mathbf{c}^-),$$

and

$$\mathcal{L}_{\text{geo}} = \lambda_1\,\mathcal{L}_{\text{intra}} + \lambda_2\,\mathcal{L}_{\text{inter}},$$

with $\lambda_1, \lambda_2 > 0$ balancing compactness and separation (consistent with Sec. C.3.1 discussion). The full objective is

$$\mathcal{L}_{\text{total}} = \underbrace{\beta_1\,\mathcal{L}_{\text{Dice}} + \beta_2\,\mathcal{L}_{\text{Focal}}}_{\mathcal{L}_{\text{seg}}} + \underbrace{\beta_3\,\mathcal{L}_{\text{BCE}}}_{\mathcal{L}_{\text{det}}} + \underbrace{\mathcal{L}_{\text{geo}}}_{\text{centroids}} + \underbrace{\mathcal{L}_{\text{var}}}_{\text{VACA reg.}},$$

where $\beta_1, \beta_2, \beta_3 > 0$ weight the task losses, and $\mathcal{L}_{\text{var}}$ is the VACA variance regularizer. $\mathcal{L}_{\text{var}}$ is only used to update the parameters of MLP_θ. All hyperparameters ($\lambda_1, \lambda_2, \beta_{1..3}, \lambda_{\text{var}}$) are selected via extensive experimentation per modality.

Algorithm 1 Variance-Aware Geodesic Matching for Zero-Shot Medical Anomaly Detection

Require: Patch embeddings $P \in \mathbb{R}^{B\times N\times D}$; class embeddings $T \in \mathbb{R}^{D\times 2}$
Require: Initial centroids $\mathbf{c}^+_{(0)}, \mathbf{c}^-_{(0)} \in \mathbb{R}^D$ (unit-norm or zero-init then normalized)
Require: EMA factor $\alpha \in (0,1)$; channel amplification $\gamma > 0$
Require: Loss weights $\beta_1, \beta_2, \beta_3 > 0$; $\lambda_1, \lambda_2 > 0$; variance weight $\lambda_{\text{var}} > 0$

Initialize: $\mathbf{c}^+ \leftarrow \mathbf{c}^+_{(0)}$, $\mathbf{c}^- \leftarrow \mathbf{c}^-_{(0)}$, $\mathcal{S}^+ \leftarrow \emptyset$, $\mathcal{S}^- \leftarrow \emptyset$
for each batch $(P, T, \mathcal{Y}_{\text{seg}}, \mathcal{Y}_{\text{det}})$ **do**
  (1) Task-Specific Adapters and Normalization
  $P_{\text{seg}} \leftarrow \text{Adapter}_{\text{seg}}(P)$; $P_{\text{det}} \leftarrow \text{Adapter}_{\text{det}}(P)$
  $P_{\text{norm}} \leftarrow \ell_2\text{-normalize}(P_{\text{seg}})$; $T_{\text{norm}} \leftarrow \ell_2\text{-normalize}(T)$
  (2) Variance-Aware Channel Attention (VACA)
  $\tilde{P} \leftarrow \text{broadcast}(P_{\text{norm}}) \in \mathbb{R}^{B\times N\times 1\times D}$; $\tilde{T} \leftarrow \text{broadcast}(T_{\text{norm}}) \in \mathbb{R}^{1\times 1\times D\times 2}$
  $\mathcal{C} \leftarrow \tilde{P} \odot \tilde{T}$; $\overline{\mathcal{C}} \leftarrow \frac{1}{BN}\sum_{b,n} \mathcal{C}_{b,n,:,:} \in \mathbb{R}^{D\times 2}$
  $\Delta_d \leftarrow \left|\overline{\mathcal{C}}_{d,\text{normal}} - \overline{\mathcal{C}}_{d,\text{anomaly}}\right|$ ▷ channel discriminability
  $a \leftarrow \text{Softplus}(\text{MLP}_\theta(\Delta))$; $w \leftarrow 1 + \gamma a$
  $\widetilde{P}_{\text{seg}} \leftarrow P_{\text{norm}} \odot w$; $\widetilde{P}_{\text{det}} \leftarrow P_{\text{det}}$
  VACA variance regularizer: $\mathsf{Var}_{\text{disc}} \leftarrow \text{Var}(\Delta \odot w)$; $\mathcal{L}_{\text{var}} \leftarrow \lambda_{\text{var}} \cdot (\mathsf{Var}_{\text{seg}} + \mathsf{Var}_{\text{det}})$
  (3) Geodesic Matching with Adaptive Gate
  for each feature f_j in $\widetilde{P}_{\text{seg}}$ with label $y \in \{\text{normal}, \text{anomaly}\}$ **do**
    $d^+ \leftarrow d_G(f_j, \mathbf{c}^+) = \arccos\big(\langle f_j, \mathbf{c}^+\rangle\big)$; $d^- \leftarrow d_G(f_j, \mathbf{c}^-) = \arccos\big(\langle f_j, \mathbf{c}^-\rangle\big)$
    $\tau(f_j) \leftarrow \min(d^+, d^-)$ ▷ adaptive, sample-specific
    if $y = \text{normal}$ **and** $d^+ < d^-$ **then**
      $\mathcal{S}^+ \leftarrow \mathcal{S}^+ \cup \{f_j\}$; $\tilde{\mathbf{c}}^+ \leftarrow \alpha\,\mathbf{c}^+ + (1-\alpha) f_j$; $\mathbf{c}^+ \leftarrow \tilde{\mathbf{c}}^+ / \|\tilde{\mathbf{c}}^+\|_2$
    else if $y = \text{anomaly}$ **and** $d^- < d^+$ **then**
      $\mathcal{S}^- \leftarrow \mathcal{S}^- \cup \{f_j\}$; $\tilde{\mathbf{c}}^- \leftarrow \alpha\,\mathbf{c}^- + (1-\alpha) f_j$; $\mathbf{c}^- \leftarrow \tilde{\mathbf{c}}^- / \|\tilde{\mathbf{c}}^-\|_2$
    else
      skip ▷ ambiguous sample; no centroid update
    end if
  end for
  (4) Losses
  Segmentation: $\hat{\mathcal{Y}}_{\text{seg}} \leftarrow \text{softmax}(\widetilde{P}_{\text{seg}} T^\top_{\text{norm}})$; $\mathcal{L}_{\text{Dice}} \leftarrow 1 - \text{Dice}(\hat{\mathcal{Y}}_{\text{seg}}, \mathcal{Y}_{\text{seg}})$; $\mathcal{L}_{\text{Focal}} \leftarrow \text{FocalLoss}(\hat{\mathcal{Y}}_{\text{seg}}, \mathcal{Y}_{\text{seg}})$; $\mathcal{L}_{\text{seg}} \leftarrow \beta_1 \mathcal{L}_{\text{Dice}} + \beta_2 \mathcal{L}_{\text{Focal}}$
  Detection: $\mathbf{s}_{\text{det}} \leftarrow \max_{\text{patch}}(\widetilde{P}_{\text{det}} T^\top_{\text{norm}})$; $\mathcal{L}_{\text{det}} \leftarrow \beta_3\,\text{BCE}(\sigma(\mathbf{s}_{\text{det}}), \mathcal{Y}_{\text{det}})$
  Geodesic (centroids): $\mathcal{L}_{\text{intra}} \leftarrow \frac{1}{|\mathcal{S}^+|}\sum_{f\in\mathcal{S}^+} d_G(f, \mathbf{c}^+)^2 + \frac{1}{|\mathcal{S}^-|}\sum_{f\in\mathcal{S}^-} d_G(f, \mathbf{c}^-)^2$; $\mathcal{L}_{\text{inter}} \leftarrow -\,d_G(\mathbf{c}^+, \mathbf{c}^-)$
  $\mathcal{L}_{\text{geo}} \leftarrow \lambda_1 \mathcal{L}_{\text{intra}} + \lambda_2 \mathcal{L}_{\text{inter}}$
  Total: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{var}}$
end for