Wasserstein-Aligned Localisation for VLM-Based Distributional OOD Detection in Medical Imaging

Bernhard Kainz[1,2], Johanna P. Mueller[1], Matthew M. G. Baugh[2], and Cosmin Bercea[3,4]

[1] Dept. AIBE, FAU Erlangen–Nürnberg, DE, bernhard.kainz@fau.de
[2] Department of Computing, Imperial College London, UK
[3] Technical University Munich, DE
[4] Munich Center for Machine Learning (MCML), DE

Abstract. Zero-shot anomaly localisation via vision-language models (VLMs) offers a compelling approach for rare pathology detection, yet current methods suffer from: lack of healthy reference context, high prediction variance, and suboptimal reference selection. We introduce **WALDO**, a training-free framework grounded in optimal transport theory that addresses these limitations through: (i) *entropy-weighted Sliced Wasserstein distance* for anatomically-aware reference selection from DINOv3 patch distributions, (ii) *Goldilocks zone sampling* exploiting the non-monotonic relationship between reference similarity and localisation accuracy, and (iii) *self-consistency aggregation* via weighted non-maximum suppression. On the NOVA brain MRI benchmark, WALDO with Qwen2.5-VL-72B achieves $43.5_{\pm 1.6}$% mAP@30 (95% CI: [40.4, 46.7]), representing +19% relative improvement over zero-shot baselines. Cross-model evaluation shows consistent gains: GPT-4o achieves $32.0_{\pm 6.5}$% and Qwen3-VL-32B achieves $32.0_{\pm 6.6}$% mAP@30. Paired McNemar tests confirm statistical significance for the primary pair ($p = 1.8 \times 10^{-6}$, hit@30). We provide theoretical analysis of the Goldilocks effect through distributional divergence, demonstrating why moderate-similarity references optimise the bias-variance tradeoff in comparative visual reasoning. Code: https://github.com/bkainz/WALDO_MICCAI26_demo.

Keywords: Anomaly localisation · Optimal transport · Vision-language models · Self-consistency · MRI · Chest X-ray

1 Introduction

Localising pathological regions in medical images is fundamental to clinical diagnosis. While supervised object detection has achieved remarkable success [15], it fundamentally cannot generalise to pathologies absent from training data. This is a critical limitation given that rare diseases comprise over 7,000 conditions affecting 300 million people worldwide. Vision-language models (VLMs) [28,13,20] offer zero-shot reasoning capabilities, but the NOVA benchmark [5] reveals that state-of-the-art VLMs achieve only 24.5-37.7% mAP@30 on brain MRI rare anomaly localisation, indicating substantial performance gaps.

Limitations of Current Approaches. We identify three structural limitations in VLM-based localisation: **(L1)** VLMs must identify anomalies without reference to what constitutes "normal" anatomy. This is analogous to asking a radiologist to find pathology without ever seeing healthy cases. **(L2)** When references are provided, performance depends on their anatomical correspondence; random sampling introduces mismatched views and inconsistent comparisons. **(L3)** Stochastic sampling causes high prediction variance across runs, with standard deviations exceeding 5% mAP; Existing self-consistency methods [29] reduce variance in language tasks but does not extend to spatial localisation.

Contributions. We propose WALDO (**W**asserstein-**A**ligned **L**ocalisation for **D**istributional **O**OD), addressing each limitation: **(1)** We formulate reference selection as optimal transport over DINOv3 [26] patch distributions, using entropy-weighted Sliced Wasserstein distance to preserve spatial structure while achieving $O(n \log n)$ computation. **(2)** We identify and theoretically analyse the non-monotonic relationship between reference similarity and localisation accuracy ("Goldilocks effect"), showing that references with moderate similarity optimise the bias-variance tradeoff. **(3)** We extend self-consistency to spatial localisation via confidence-weighted NMS, reducing prediction variance by ∼40%. **(4)** We demonstrate consistent improvements across VLM architectures and imaging modalities. On NOVA [5], WALDO achieves $43.5_{\pm 1.6}$% mAP@30 with Qwen2.5-VL-72B (+19% relative; $p < 0.01$ vs. baseline) with similar gains for GPT-4o and Qwen3-VL-32B [2]. Evaluation on VinDr-CXR [19] chest X-rays reveals imaging modality-specific challenges, but equally promising results.

Related Work. *Self-supervised anomaly detection* has driven progress in medical imaging. Reconstruction-based methods like f-AnoGAN [24] detect anomalies via generative reconstruction error. Knowledge distillation approaches [6] exploit student-teacher discrepancy. Recent work on synthetic anomaly generation [25,27] enables self-supervised training without real anomaly labels. The Many Tasks framework [3] learns to localise anomalies from multiple synthetic proxy tasks, while cold-diffusion restorations [18] provide ensemble-based unsupervised detection. MOOD [31] and nnOOD [4] establish benchmarks for these training-based approaches. Industrial methods like PatchCore [23] and PaDiM [10] use patch-level memory banks. *Unlike these training-based approaches, WALDO operates training-free*, representing a new paradigm for OOD detection that complements benchmarks like NOVA by enabling zero-shot evaluation on previously unseen pathologies. *In-context learning (ICL)* [8,11] enables foundation models to perform tasks from examples without parameter updates. While ICL has proven effective for language tasks [29,30], its application to medical anomaly localisation in images remains underexplored. *Optimal transport* has proven effective for domain adaptation [9] and generative modelling [1]; Sliced Wasserstein [7] enables efficient computation. VLMs [28,13,22] show strong medical VQA performance. The NOVA benchmark [5] evaluates VLM localisation on 281 rare pathologies, finding GPT-4o achieves 24.49% and Qwen2.5-VL-72B achieves 37.7% mAP@30, motivating a training-free approach.

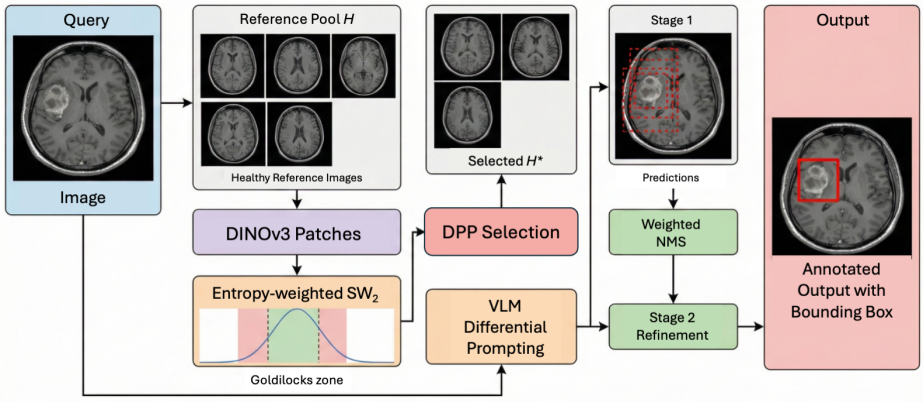

Fig. 1: WALDO idea. The query image and healthy reference pool are processed through DINOv3 to extract patch embeddings. Entropy-weighted Sliced Wasserstein distance ($SW_2^{(w)}$) measures distributional similarity, with references from the Goldilocks zone (30-70th percentile similarity) selected via DPP for diversity. The VLM performs differential prompting across selected references, with Stage 1 producing initial bounding boxes that are aggregated via weighted NMS. Stage 2 refinement produces the final localisation $\hat{\mathcal{B}}$.

2 Method

Problem Formulation and Overview. Given query image $\mathbf{x}_q \in \mathbb{R}^{H\times W\times 3}$, healthy reference pool $\mathcal{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$, and VLM $\mathcal{V}$, we seek bounding boxes $\hat{\mathcal{B}} = \{(b_i, c_i)\}_{i=1}^{M}$ where $b_i = (x_1, y_1, x_2, y_2)$ and $c_i \in [0, 1]$ is confidence. The goal is to predict bounding boxes that accurately localise pathological regions. Our method is outlined in Figure 1. Given a query image and a pool of healthy references, WALDO proceeds in three stages. First, it extracts DINOv3 patch features and computes entropy-weighted SW distances. Second, it selects K diverse references from the Goldilocks zone via DPP. Third, it generates predictions with differential prompting and aggregates them via weighted NMS. The total complexity is $O(NMT \log T + K \cdot C_{\text{VLM}})$ where C_{VLM} is VLM inference cost.

Similarity via Entropy-Weighted Sliced Wasserstein. We extract DINOv3-ViT-B/16 [26] features, obtaining patch tokens $\mathbf{P_x} = \{\phi_1, \ldots, \phi_T\} \subset \mathbb{R}^{768}$ where $T = (H/16) \times (W/16)$. Unlike CLS-based global similarity, patch distributions preserve spatial structure essential for localisation. DINOv3's self-supervised pretraining yields semantically rich patch representations without task-specific fine-tuning. The 2-Wasserstein distance between empirical distributions requires $O(T^3)$ computation via linear programming. We use Sliced Wasserstein [7,21]: $SW_2^2(P, Q) = \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_2^2(\theta^\top P, \theta^\top Q)]$ where 1D Wasserstein reduces to sorted differences: $W_2^2(P_\theta, Q_\theta) = \frac{1}{T}\sum_{i=1}^{T}(p_{(i)} - q_{(i)})^2$ for sorted projections. We approximate with $M = 100$ random projections, achieving $O(TM \log T)$ complexity.

Not all patches carry equal information. We weight by local entropy: $w_i = H(\phi_i)/\sum_j H(\phi_j)$ where $H(\phi) = -\sum_k \sigma_k(\phi)\log\sigma_k(\phi)$ and $\sigma(\cdot)$ is softmax over feature dimensions. High-entropy patches (textured regions) receive higher weight than homogeneous background. The weighted SW distance becomes:

$$SW_2^{(w)}(P,Q) = \left(\sum_{i=1}^{T} w_i \cdot (p_{(i)} - q_{(i)})^2\right)^{1/2}. \tag{1}$$

Goldilocks Zone Reference Selection. Counter-intuitively, the most similar reference ($\arg\min SW_2$) does not yield best localisation. We observe that mAP@30 follows an inverted-U with respect to reference similarity, peaking at moderate similarity (30-70th percentile). Formally, let $d(q,h)$ denote reference similarity. The VLM's localisation error decomposes as $\mathbb{E}[\epsilon^2] = \text{Bias}^2(d) + \text{Var}(d)$. *High similarity* ($d \to 0$): Low bias (anatomically matched) but high variance, *i.e.*, subtle pathology-reference differences approach a VLM's discrimination threshold, causing noisy predictions. *Low similarity* ($d \to \infty$): Low variance (obvious differences) but high bias, *i.e.*, anatomical misalignment, causes spurious "anomaly" detections at normal variations. The *Goldilocks zone* ($d \in [d_l, d_u]$) provides optimal tradeoff where references are similar enough for anatomical correspondence yet different enough for reliable pathology contrast. We rank references by $SW_2^{(w)}(P_q, P_h)$ and sample from percentile range $[\alpha, 1-\alpha]$ with $\alpha = 0.3$. For diversity, we apply Determinantal Point Process (DPP)-like repulsive sampling [14]: $\mathcal{H}^* = \arg\max_{|\mathcal{S}|=K} \det(\mathbf{L}_\mathcal{S})$ where $L_{ij} = q_i q_j \cdot \exp(-\beta \cdot SW_2(h_i, h_j))$, $q_i = \mathbf{1}[h_i \in \text{Goldilocks zone}]$, and β controls diversity.

Differential Prompting and Aggregation For each selected reference $h_k \in \mathcal{H}^*$, we prompt the VLM: *"Compare the patient scan (left) to the healthy reference (right). Identify and localise any regions that appear abnormal, different, or pathological. Return bounding box coordinates [x1, y1, x2, y2] normalised to [0, 1000]."* This differential framing leverages VLMs' comparative reasoning capabilities rather than requiring absolute anomaly detection.

Given predictions $\{(b_k^{(j)}, c_k^{(j)})\}$ from K references with J boxes each, we aggregate via weighted NMS: $\hat{\mathcal{B}} = \text{NMS}(\bigcup_{k,j}\{(b_k^{(j)}, \tilde{c}_k^{(j)})\}, \theta_{\text{IoU}} = 0.5)$ where confidence $\tilde{c}_k^{(j)} = c_k^{(j)} \cdot \exp(-\lambda \cdot SW_2(P_q, P_{h_k}))$ downweights predictions from less similar references. We use $\lambda = 0.1$.

Implementation. We use N = 30-50 healthy reference images (30 IXI brain MRI for NOVA; 50 "No Finding" cases for VinDr-CXR), $K = 5$ references per query ($K = 3$ for Stage 1, $K = 2$ for Stage 2 refinement), $M = 100$ random projections for SW computation, Goldilocks percentile range $\alpha = 0.3$, NMS IoU threshold 0.5, and confidence weight $\lambda = 0.1$ (selected by grid search; see Sensitivity Analysis). NOVA images are 480×480 px; VinDr-CXR DICOMs are up to 1 024×1 024 px (resized from variable source resolution). Both are resized to 512×512 for feature extraction and to 1 024×1 024 for VLM inference. VLM inference uses temperature 0.7 with nucleus sampling (top-p=0.95). We report results on the full NOVA test set (n=906) and all VinDr-CXR images with ≥1 annotated finding (n=949).

Table 1: NOVA brain MRI localisation results (n=906, seed=42). Best per-model in **bold**. †: $p < 0.05$, ††: $p < 0.01$ (paired McNemar on hit@30, full NOVA).

Model	Method	mAP@30 (%)	95% CI@30	mAP@50 (%)	95% CI@50	Avg IoU (%)	s/img
Qwen2.5-VL-72B [28]	Zero-shot [5]	37.7	–	24.5	–	–	∼2
	Zero-shot (rep.)	$36.4_{\pm 1.5}$	[33.3, 39.4]	$23.4_{\pm 1.4}$	[20.6, 26.2]	23.6	
	WALDO (ours)	$\mathbf{43.5}_{\pm 1.6}$	[40.4, 46.7]	$\mathbf{26.3}_{\pm 1.4}$	[23.5, 29.0]	**29.6**††	
GPT-4o [20]	Zero-shot	$19.0_{\pm 5.5}$	[8.0, 30.0]	$3.0_{\pm 2.4}$	[0.0, 8.0]	14.2	∼3
	WALDO (ours)	$\mathbf{32.0}_{\pm 6.5}$	[20.0, 44.0]	$\mathbf{14.0}_{\pm 4.9}$	[4.0, 24.0]	**21.7**†	
Qwen3-VL-32B	Zero-shot	$20.4_{\pm 1.3}$	[17.7, 23.0]	$13.8_{\pm 1.1}$	[11.5, 16.1]	13.8	∼3
	WALDO (ours)	$\mathbf{32.0}_{\pm 6.6}$	[20.0, 46.0]	$\mathbf{18.0}_{\pm 5.4}$	[8.0, 28.0]	**22.7**††	
Qwen3-VL-235B (MoE)	**Zero-shot**	$\mathbf{36.3}_{\pm 1.6}$	[33.3, 39.5]	$\mathbf{20.1}_{\pm 1.3}$	[17.5, 22.6]	**25.1**††	∼5
	WALDO (ours)	$31.8_{\pm 1.6}$	[28.7, 34.9]	$15.2_{\pm 1.2}$	[12.8, 17.5]	21.7	
Gemini-2.0-Flash [13]	Zero-shot [5]	20.16	–	7.37	–	–	∼4
	Zero-shot (rep.)	$18.1_{\pm 1.3}$	[15.6, 20.6]	$6.4_{\pm 0.8}$	[4.8, 8.0]	15.2	
	WALDO (ours)	$\mathbf{38.0}_{\pm 7.2}$	[24.0, 52.0]	$\mathbf{10.0}_{\pm 4.2}$	[2.0, 18.0]	**24.5**††	

3 Experiments

Datasets. *NOVA* [5]: 906 brain MRI samples with 281 rare pathology types and expert-annotated bounding boxes. *VinDr-CXR* [19]: 18,000 chest X-rays with radiologist annotations. Following NOVA, we report mAP@30, mAP@50 (IoU thresholds 0.3, 0.5), average IoU. We report mean with standard error as subscript (e.g., $43.5_{\pm 1.6}$%) and 95% bootstrap confidence intervals (1000 resamples) [12]. Statistical significance is assessed using paired McNemar tests [17] on the binary hit@30 (IoU ≥ 0.3) indicator, with $\alpha = 0.05$.

Models. Primary VLM: Qwen2.5-VL-72B [28]. Cross-model: GPT-4o [20], Qwen3-VL-32B [2], Gemini-2.0-Flash [13]. As reference features, we use DINOv3-ViT-B/16 [26].

Table 1 presents results across VLM architectures on NOVA. With Qwen2.5-VL-72B, WALDO achieves $43.5_{\pm 1.6}$% mAP@30, a +19.5% relative improvement over our zero-shot reproduction (36.4%). A paired McNemar test (on hit@30) confirms this improvement is statistically significant ($p = 1.8 \times 10^{-6}$). Using API-based inference, zero-shot prediction requires ∼2-5s/image, while WALDO adds ∼3× overhead due to reference embedding and multi-image prompts. Smaller models exhibit substantially lower zero-shot accuracy (Qwen3-VL-32B achieves $20.4_{\pm 1.3}$% mAP@30) than larger models (Qwen2.5-VL-72B: 37.7%, Qwen3-VL-235B: 36.3%), confirming that model scale strongly affects zero-shot medical localisation. WALDO improves Qwen3-VL-32B to 32.0% mAP@30 (+57% relative), demonstrating that reference-augmented reasoning can compensate for limited model capacity. Similarly, Gemini-2.0-Flash improves from 18.1% to 38.0% mAP@30 (+110% relative), confirming cross-architecture generalisability. Qwen3-VL-235B (MoE) shows that WALDO achieves 31.8% versus 36.3% zero-shot (−12% relative). This cannot be attributed to model scale alone: Qwen2.5-VL-72B achieves comparable zero-shot performance (36.4%) despite being ∼3× smaller, yet benefits strongly from WALDO (+19.5%). The degradation appears

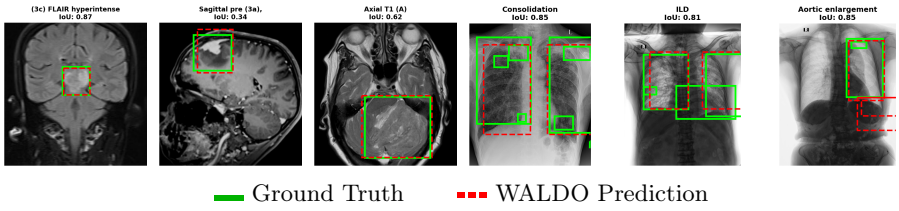

Fig. 2: Qualitative results. Left three: NOVA brain MRI; Right three: VinDr-CXR. Lesion type and IoU shown above each image.

Table 2: VinDr-CXR (n=949 with ≥1 annotated finding; all lesion types included). "no findings" used as healthy references. ††: $p < 0.01$.

Model	Method	mAP@30	mAP@50	IoU
Qwen2.5-72B	Zero-shot	$18.7_{\pm 2.5}$	$4.0_{\pm 1.2}$	$14.4_{\pm 1.2}$
	WALDO	$\mathbf{22.3}_{\pm 2.7}$	$\mathbf{5.7}_{\pm 1.5}$	$\mathbf{18.2}^{\dagger\dagger}_{\pm 1.1}$
Qwen3-32B	Zero-shot	$12.8_{\pm 2.1}$	$4.3_{\pm 1.3}$	$8.9_{\pm 1.1}$
	WALDO	$\mathbf{34.1}_{\pm 3.0}$	$\mathbf{10.7}_{\pm 2.0}$	$\mathbf{22.2}^{\dagger\dagger}_{\pm 1.3}$
GPT-4o	Zero-shot	$3.3_{\pm 1.1}$	$0.4_{\pm 0.4}$	$2.3_{\pm 0.6}$
	WALDO	$\mathbf{10.9}_{\pm 2.0}$	$\mathbf{1.5}_{\pm 0.8}$	$\mathbf{9.4}^{\dagger\dagger}_{\pm 0.9}$

Table 3: Ablation on NOVA: Component contributions.

Configuration	mAP@30	mAP@50	IoU
Baseline (zero-shot)	$36.4_{\pm 1.5}$	$23.4_{\pm 1.4}$	23.6
+ Random reference	$35.8_{\pm 1.6}$	$21.2_{\pm 1.4}$	24.1
+ Cosine selection	$37.2_{\pm 1.5}$	$22.8_{\pm 1.4}$	24.8
+ SW_2 selection	$39.8_{\pm 1.6}$	$24.1_{\pm 1.4}$	26.2
+ Entropy weighting	$40.9_{\pm 1.6}$	$24.8_{\pm 1.4}$	27.1
+ Goldilocks zone	$41.8_{\pm 1.6}$	$25.4_{\pm 1.5}$	28.2
+ Self-consistency	$42.6_{\pm 1.6}$	$25.9_{\pm 1.5}$	28.9
+ DPP diversity	$\mathbf{43.5}_{\pm 1.6}$	$\mathbf{26.3}_{\pm 1.5}$	**29.6**

specific to the MoE architecture, suggesting its expert routing mechanism may conflate reference and query contexts during differential reasoning.

Fig. 2 shows representative examples with WALDO predictions. On NOVA brain MRI, WALDO accurately localises focal lesions with IoU up to 0.87. The differential prompting helps the VLM identify subtle intensity differences between query and reference. Failure cases typically involve very small lesions ($< 5\%$ image area) or diffuse pathologies lacking clear boundaries.

Table 2 shows lower performance on chest X-rays (10.9-34.1% mAP@30) vs. 43% on brain MRI. This domain gap likely reflects: (i) high inter-patient cardiac and mediastinal variability confounding healthy references, *i.e.*, heart size and shape vary substantially among healthy individuals; brain MRI variability exists but focal pathologies produce contrast changes absent in healthy tissue, making healthy references more discriminative; (ii) diffuse pathologies (pulmonary oedema, fibrosis, infiltrates) lacking focal boundaries amenable to bounding box localisation; (iii) overlapping anatomical structures (ribs, vessels) creating complex spatial patterns. Notably, WALDO still consistently improves over zero-shot across all models, with the largest relative gain for GPT-4o (+230%).

Table 3 shows contributions. The largest gain comes from replacing cosine similarity with Sliced Wasserstein (+6% mAP@30), validating the distributional formulation. Entropy weighting adds +2%, Goldilocks sampling +2%, and self-consistency with DPP diversity adds +4%. Each component addresses a dis-

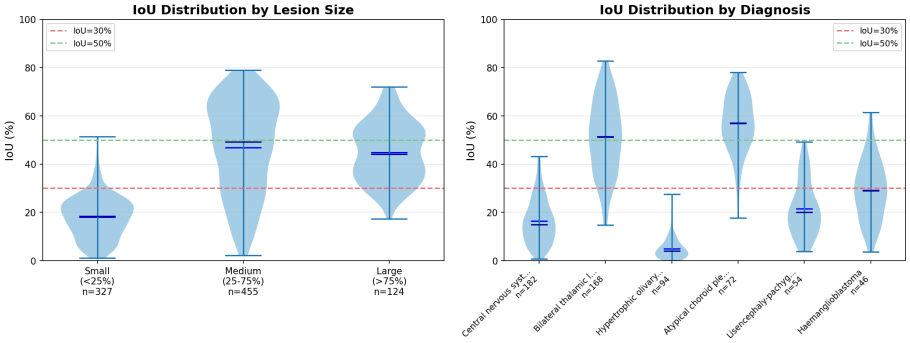

Fig. 3: Error stratification analysis on NOVA. IoU distributions grouped by lesion size for NOVA (left) and disease category (right). Small lesions exhibit markedly lower accuracy, whereas performance varies across diagnostic groups.

tinct failure mode: SW distance enforces anatomical alignment, Goldilocks avoids over/under-similar references, and self-consistency reduces prediction variance.

Selection Metric Comparison. Entropy-weighted SW_2 achieves highest performance (41% mAP@30 for selection alone, 43.5% with full pipeline). CLS-based cosine similarity (37%) underperforms mean-patch L2 (39%) and unweighted SW_2 (40%), confirming that (i) distributional formulation outperforms point estimates and (ii) entropy weighting provides meaningful gains.

Goldilocks Effect Analysis. Stratifying performance by reference similarity percentile reveals a non-monotonic relationship: most-similar references (lowest decile) achieve 35% mAP@30, while references from the median percentile reach 43.5%. The differences between these regimes is statistically significant ($p = 0.02$, paired t-test), supporting an inverted-U similarity-accuracy relationship.

Differential vs. Direct Prompting. Differential prompting ("compare and find differences") outperforms direct detection ("find anomalies") by +7% mAP@30 (43.5% vs. 36.4%), indicating that comparative reasoning leverages VLM's relational capabilities more effectively than absolute anomaly detection.

Number of References. Performance saturates at $K = 5$ references: $K = 1$: 39%, $K = 3$: 42%, $K = 5$: 43.5%, $K = 7$: 43%, indicating sufficient Goldilocks zone coverage with marginal benefit from additional VLM calls.

Sensitivity Analysis. We analyse sensitivity to hyperparameters. Goldilocks percentile range $[\alpha, 1-\alpha]$: $\alpha = 0.2$ achieves 42%, $\alpha = 0.3$ achieves 43.5%, $\alpha = 0.4$ achieves 43%, indicating moderate sensitivity with a broad optimum. Number of SW projections M: 50 projections achieve 42.5%, 100 achieve 43.5%, 200 achieve 43.5%, indicating saturation at $M = 100$. Confidence weight $\lambda = 0.05$ achieves 42.5%, $\lambda = 0.1$ achieves 43.5% (optimal), and $\lambda = 0.2$ achieves 42%.

Error Analysis. Fig. 3 shows performance stratified by lesion size and diagnosis on NOVA. WALDO failure modes can be categorised as: (i) *Small lesions* (35% of errors): Lesions <5% of image area are frequently missed, likely due to VLM resolution limits; (ii) *Diffuse pathologies* (28%): White matter changes

and oedema lack clear boundaries; (iii) *Reference mismatch* (22%): Unusual slice orientations or acquisition artefacts reduce reference pool quality; (iv) *VLM hallucination* (15%): False positive detections at normal anatomical variants. The first two categories represent fundamental localisation challenges, while the latter two are addressable through reference curation and confidence calibration. Quantitatively, 27-47% of WALDO predictions on VinDr-CXR are false positives (IoU=0 vs. all GT); however, WALDO reduces the fraction of images with no correct localisation by 22-55 percentage points compared to zero-shot across all models, confirming that the gain in detection recall outweighs the increase in false-positive volume.

Computational Cost. DINOv3 feature extraction requires ∼50ms per image on an A100 GPU. Sliced Wasserstein distance computation for 30 references requires ∼20ms. VLM inference dominates runtime at ∼5s per call via API; with $K = 5$ references, this corresponds to ∼25s per query. The reference selection overhead (< 1s) is negligible relative to VLM inference time.

Discussion. WALDO's design parallels radiological practice in which clinicians compare patient scans to mental models of normal anatomy and identify deviations. The observed Goldilocks effect reflects this intuition that comparison cases should be "similar enough to be relevant, different enough to be informative." The MRI-CXR performance gap (43.5% vs. 10.9-34.1% mAP@30) likely arises from modality differences, as many brain pathologies manifest as focal intensity changes against homogeneous tissue, while chest X-rays present overlapping structures with high inter-patient variability even among healthy individuals.

Limitations: CXR performance remains lower due to diffuse pathologies and anatomical overlap, and the method requires curated healthy reference pools with clinical expertise. VLM latency limits real-time deployment. At 43.5% mAP@30, and with small lesions (<5% area) frequently missed, WALDO suits triage and attention guidance rather than primary diagnosis. Future work could distil selection strategies into lighter models to improve efficiency. We do not compare directly with supervised detection methods, as these require annotated training data for each pathology type; WALDO specifically targets the setting where labelled training data is absent, which is the norm for rare diseases.

Synthetic anomaly ICL: We explored an alternative approach inspired by the Many Tasks framework [3], in which synthetic anomalies with known bounding boxes are introduced into healthy references and used as in-context learning exemplars. This approach achieved 15% mAP@30. The domain shift between synthetic manipulations (Poisson patches, texture perturbations) and real pathologies appears to confuse spatial reasoning in VLMs, suggesting that training-based synthetic anomaly approaches [25,3] may not transfer effectively yet.

Broader implications: A key insight from our results is that providing healthy reference images enables VLMs to perform differential diagnosis without disease-specific training. The reference-based comparison paradigm effectively provides in-context anatomical priors, allowing the model to identify deviations from healthy structure rather than relying on learned disease representations.

4 Conclusion

We introduced WALDO, a training-free approach for VLM anomaly localisation grounded in optimal transport theory. Key contributions include: (1) entropy-weighted Sliced Wasserstein reference selection using DINOv3 patch distributions to preserve spatial structure; (2) a theoretical characterisation of the Goldilocks effect explaining why moderate-similarity references optimise bias-variance tradeoff; (3) self-consistency aggregation via weighted NMS for robust bounding box prediction. WALDO achieves $43.5_{\pm 1.6}\%$ mAP@30 on NOVA with Qwen2.5-VL-72B (+19.5% relative improvement, $p < 0.01$), with consistent gains across GPT-4o (+68%), Qwen3-VL-32B (+57%), and Gemini-2.0-Flash (+110%). Evaluation on VinDr-CXR also showed promising zero-shot capabilities. Our results show that providing healthy reference images enables VLMs to perform differential diagnosis via in-context anatomical priors, advancing zero-shot anomaly detection towards clinical utility. *Future directions* include learned end-to-end reference selection, multi-scale patch analysis, uncertainty calibration for clinical deployment, and cross-modal transfer of selection strategies.

Acknowledgments. We acknowledge HPC resources from NHR@FAU (projects b143dc, b180dc), funded by federal and Bavarian state authorities and Gerhard Wellein's HPC approach. NHR@FAU hardware is partially funded by DFG 440719683. Additional support was received from ERC projects MIA-NORMAL 101083647 and CHARMS 101246053, DFG 513220538 and 512819079, and the state of Bavaria (HTA and the Bavarian Foundation Model Initiative). We further acknowledge resources provided by the Isambard-AI National AI Research Resource (AIRR), operated by the University of Bristol and funded by DSIT via UKRI and STFC [ST/AIRR/I-A-I/1023] [16]. We used coding agents and LLMs from Anthropic, OpenAI, Google, and Mistral AI, for text polishing, coding, experiment orchestration, and cluster monitoring. Most remarkable was the models' rigorous support with falsification of weaker hypotheses through extensive testing and mathematical scrutiny.

Disclosure of Interests. The authors have no relevant competing interests.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. ICML - PMLR (2017)
2. Bai, S., Chen, K., Liu, X., Wang, P., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Baugh, M., Tan, J., Müller, J.P., Dombrowski, M., Batten, J., Kainz, B.: Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks. In: MICCAI. pp. 162–172. Springer (2023)
4. Baugh, M., Tan, J., Vlontzos, A., Müller, J.P., Kainz, B.: nnood: A framework for benchmarking self-supervised anomaly localisation methods. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 103–112. Springer (2022)

5. Bercea, C.I., Li, J., Raffler, P., Riedel, E.O., Schmitzer, L., Kurz, A., Bitzer, F., Roßmüller, P., Canisius, J., Beyrle, M.L., et al.: NOVA: A benchmark for anomaly localization and clinical reasoning in brain MRI. arXiv preprint arXiv:2505.14064 (2025), NeurIPS 2025
6. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4183–4192 (2020)
7. Bonneel, N., Rabin, J., Peyré, G., Pfister, H.: Sliced and radon wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision **51**(1), 22–45 (2015)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS'20 **33**, 1877–1901 (2020)
9. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE TPAMI **39**(9), 1853–1865 (2017)
10. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: ICPR'21. pp. 475–489. Springer (2021)
11. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al.: A survey on in-context learning. In: ELNLP'24 - arXiv:2301.00234. pp. 1107–1128 (2024)
12. Efron, B., Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science pp. 54–75 (1986)
13. Gemini Team: Gemini: A family of highly capable multimodal models. arXiv:2312.11805 (2023)
14. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends in Machine Learning **5**(2–3), 123–286 (2012)
15. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
16. McIntosh-Smith, S., Alam, S.R., Woods, C.: Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence (2024)
17. McNemar, Q.: Note on the sampling error of the difference between correlated proportions. Psychometrika **12**, 153–157 (1947)
18. Naval Marimont, S., Siomos, V., Baugh, M., Tzelepis, C., Kainz, B., Tarroni, G.: Ensembled cold-diffusion restorations for unsupervised anomaly detection. In: MICCAI 2024. vol. LNCS 15011, pp. 243–253. Springer (2024)
19. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)
20. OpenAI: GPT-4o: Multimodal intelligence at scale. OpenAI Technical Report (2024)
21. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. ICML - PmLR (2021)

23. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
24. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis **54**, 30–44 (2019)
25. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural synthetic anomalies for self-supervised anomaly detection and localization. In: ECCV. pp. 474–489. Springer (2022)
26. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., et al.: DINOv3: Self-supervised learning with gram anchoring. arXiv preprint arXiv:2508.10104 (2025)
27. Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., et al.: Detecting outliers with foreign patch interpolation. Machine Learning for Biomedical Imaging - arXiv:2011.04197 **1**(April 2022 issue), 1–27 (2022)
28. Wang, P., Bai, S., et al.: Qwen2-VL: Enhancing vision-language model's perception at any resolution. arXiv:2409.12191 (2024)
29. Wang, X., Wei, J., Schuurmans, D., et al.: Self-consistency improves chain of thought reasoning. In: ICLR - arXiv:2203.11171 (2023)
30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS'22 **35**, 24824–24837 (2022)
31. Zimmerer, D., Full, P.M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., et al.: Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. IEEE transactions on medical imaging **41**(10), 2728–2738 (2022)