RADAR: Relative Angular Divergence Across Representations

Xavier Cadet*
Dartmouth College
Hanover, NH 03755, USA
xfc@dartmouth.edu

Mateusz Nowak*
Dartmouth College
Hanover, NH 03755, USA
mateusz.m.nowak.th@dartmouth.edu

Peter Chin
Dartmouth College
Hanover, NH 03755, USA
peter.chin@dartmouth.edu

Abstract

Machine learning methods rely on data. However, gathering suitable data can be challenging due to availability constraints, cost, or the need for domain expertise. Expanding datasets with additional sources is a common response to limited data, yet this practice does not always improve downstream performance and can sometimes lead to a loss of performance, known as negative transfer. We propose RADAR, a simple, geometrically grounded metric for estimating cross-domain transferability in foundation models. RADAR analyzes the layer-wise evolution of representations by measuring angular alignments and relative changes in distance along layer-to-layer displacement trajectories, and by comparing empirical distributions of within-domain and cross-domain dynamics. We hypothesize that domain transferability is related to the divergence between these trajectory distributions. We evaluate the metric across multiple modalities, including cross-lingual sentiment classification with text embedding models and cross-domain image classification with foundation vision models. Across several settings, RADAR provides competitive predictive performance relative to existing transferability metrics on several vision and text benchmarks, with particularly strong results when domain transitions are smooth or cleanly separated. Our ablations further suggest that the effectiveness of transferability estimation depends on the geometry of the model's internal representation space, with different modalities favoring different topological formulations.

1 Introduction

In modern machine learning, expanding training datasets with auxiliary sources is a standard strategy to address data constraints. However, this practice does not guarantee improved downstream performance and can even lead to negative transfer. Identifying which source domains will yield positive transfer remains a difficult task, as empirically evaluating each candidate impact requires costly full-model retraining. To address this challenge, we propose RADAR (**R**elative **A**ngular **D**ivergence **A**cross **R**epresentations), a simple, geometrically grounded metric designed to efficiently estimate cross-domain transferability in foundation models using only frozen features.

Unlike baseline metrics that require pre-training on the target dataset or evaluate only the static output of a single layer, RADAR analyzes the dynamic, layer-wise structural evolution of representations

*Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

using frozen foundation models. We provide a high-level overview of the metric in Figure 1. The framework measures divergence by extracting intermediate feature representations from the model to map layer-to-layer displacement trajectories. These trajectories are then characterized by computing relative distance changes and angles between samples within the same domain versus samples across different domains. Based on these angles and relative distances, we fit high-dimensional Gaussian Mixture Models (GMMs) to model the continuous probability distributions of the within-domain and cross-domain dynamics. Finally, given these two distributions, we measure the overall structural shift between domains by computing the symmetric Kullback-Leibler (KL) divergence between them.

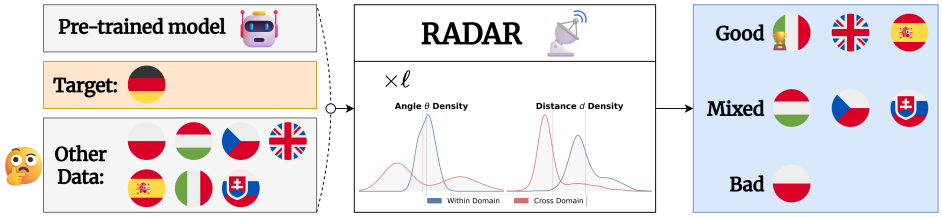


Figure 1: **Overview of the RADAR framework.** RADAR identifies optimal source data for transfer learning by extracting layer-wise (ℓ) geometric trajectories from pre-trained models. By computing the divergence in angle (θ) and relative distance (d) densities between within- and cross-domain pairs, it efficiently ranks auxiliary candidates without requiring computationally expensive fine-tuning.

We evaluate RADAR across both vision modalities (using CLIP [36] and DINOv3 [38] architectures) and text modalities (using Qwen3-Embedding [48] and EmbeddingGemma [43]) on benchmark datasets such as DomainNet [35], OfficeHome [42], PACS [25] for the vision modality, AmazonReview [34], and EuroEval [34] for the text modality, comparing it against state-of-the-art transferability metrics. We show that RADAR achieves top-three performance in seven out of ten configurations, securing the absolute best performance in six of them. Furthermore, we thoroughly evaluate the design choices of our algorithm—specifically focusing on domain subsampling, density estimation, and alternative optimal transport algorithms—demonstrating the overall robustness of our method. Our core contributions and findings are summarized as follows:

- **Dynamic, Layer-Wise Structural Evolution:** We propose RADAR—a geometrically grounded metric that evaluates dynamic representation shifts between different domains. By measuring the divergence between these dynamic, continuous distributions, we effectively capture both smooth domain transitions and distinctly separated domains.
- **Highly Competitive Predictive Performance:** RADAR consistently ranks as the best or second-best transferability metric against established baselines (such as LEEP [32], LogME [45], NCE [41], TransRate [13], and S-OTDD [33]) on text and vision modalities.
- **Robust and Efficient Estimation:** We demonstrate the robustness and high computational efficiency of our method through an extensive ablation study, focusing on hyperparameter choices, density estimation techniques, representation subsampling, and alternative optimal transport algorithms.

Ultimately, RADAR offers a competitive, cross-modal transferability metric that reliably predicts transfer gain rankings across diverse vision and text benchmarks — particularly when domain transitions are smooth or cleanly separated — while remaining computationally efficient relative to empirical- and optimal transport-based metrics.

2 Problem Statement

Domains and shared label space. Let $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ be a collection of K domains, each containing labeled data $\mathcal{D}_i = \{(x_j, y_j)\}_{j=1}^{N_i}$ drawn from a domain-specific joint distribution $P_i(X, Y)$. We assume all domains share a common label space $\mathcal{Y}$, with every class represented in every domain. One domain $\mathcal{D}_{\text{tgt}} \in \mathcal{D}$ is designated the *target*; the remaining $K - 1$ domains form the pool of *source*

Project page: `https://futuramistic.github.io/radar/`.

candidates, $\mathcal{D}_{\text{src}} = \mathcal{D} \setminus \{\mathcal{D}_{\text{tgt}}\}$. For instance, assuming we have three domains {real, paintings, sketches} dataset, we would like to know whether we should use paintings, and sketches to improve our performance on real images.

Blended training set. A *blend* is a subset $\mathcal{S} \subseteq \mathcal{D}_{\text{src}}$; the corresponding blended dataset is the union of its member domains,

$$\mathcal{D}_{\text{blend}}(\mathcal{S}) = \bigcup_{\mathcal{D}_{\text{i}} \in \mathcal{S}} \mathcal{D}_{\text{i}}. \tag{1}$$

There are 2^{K-1} possible blends. For instance, if our target domain is real, should we train with {{}, {paintings}, {sketches}, {sketches, paintings}}. Given a blend $\mathcal{D}_{\text{blend}}(\mathcal{S})$, we train a classifier on the union of the target data and the source blend, $\mathcal{D}_{\text{tgt}} \cup \mathcal{D}_{\text{blend}}(\mathcal{S})$, and evaluate it on $\mathcal{D}_{\text{tgt}}$. We can now naturally evaluate the target-only baseline as $\mathcal{S} = \varnothing$, yielding $A(\varnothing) = \text{acc}\big(f_{\mathcal{D}_{\text{tgt}}}, \mathcal{D}_{\text{tgt}}\big)$. We record two quantities as ground truth:

$$A(\mathcal{S}) = \text{acc}\big(f_{\mathcal{D}_{\text{tgt}} \cup \mathcal{D}_{\text{blend}}(\mathcal{S})}, \mathcal{D}_{\text{tgt}}\big), \tag{2}$$

$$\Delta(\mathcal{S}) = A(\mathcal{S}) - A(\varnothing). \tag{3}$$

A positive $\Delta(\mathcal{S})$ indicates that blending in the sources $\mathcal{S}$ improves over training on target data alone; a negative value indicates negative transfer.

Source selection problem. Computing $A(\mathcal{S})$ for every subset $\mathcal{S}$ requires 2^{K-1} full training runs, which is intractable for large K. Our goal is to predict $\Delta(\mathcal{S})$ — and thereby rank blend configurations — from *frozen features alone*, without any model retraining. Concretely, we seek a *blend similarity function*

$$\text{dsim}\big(\mathcal{D}_{\text{blend}}(\mathcal{S}), \mathcal{D}_{\text{tgt}}\big) \in \mathbb{R} \tag{4}$$

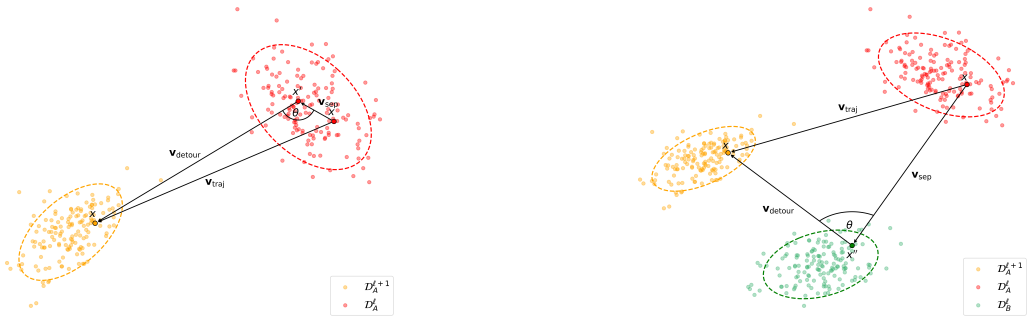
such that its ranking over all $2^{K-1} - 1$ non-empty blends is consistent with the ranking induced by $\Delta(\mathcal{S})$:

$$\text{dsim}\big(\mathcal{D}_{\text{blend}}(\mathcal{S}),\, \mathcal{D}_{\text{tgt}}\big) < \text{dsim}\big(\mathcal{D}_{\text{blend}}(\mathcal{S}'),\, \mathcal{D}_{\text{tgt}}\big) \implies \Delta(\mathcal{S}) > \Delta(\mathcal{S}'). \tag{5}$$

We evaluate the quality of a similarity function by its Spearman rank correlation ρ with the ground-truth Δ values across all blend configurations.

3 Relative Angular Divergence Across Representations as a Metric

Feature extraction. Let Φ denote a pre-trained foundation model, such as a text embedding or vision architecture. For an input sample x_i drawn from a specific domain $\mathcal{D}_{\text{d}} \in \mathcal{D}$, we define its intermediate feature representation extracted at layer l of Φ as $\mathbf{h}^{(l)}(x_i) \in \mathbb{R}^{H_l}$, where H_l is the dimensionality of the representation at layer l.



(a) Within-domain dynamic

(b) Across-domain dynamic

Figure 2: Geometric triangle of displacement vectors spanning layers l and $l + 1$

Angle and relative distance. Given two domains, $\mathcal{D}_{\mathrm{A}}$ and $\mathcal{D}_{\mathrm{B}}$, we compute empirical within-domain and cross-domain angle and distance distributions. For a given anchor sample $x \in \mathcal{D}_{\mathrm{A}}$ with label y, we draw a secondary intra-domain sample $x' \in \mathcal{D}_{\mathrm{A}} \setminus \{x\}$ and a cross-domain sample $x'' \in \mathcal{D}_{\mathrm{B}}$. To capture within-domain dynamics, we define a geometric triangle of displacement vectors spanning layers l and $l+1$ (see Figure 2):

$$\mathbf{v}_{\text{sep}}^{(l)}(x, x') = \mathbf{h}^{(l)}(x') - \mathbf{h}^{(l)}(x), \tag{6}$$

$$\mathbf{v}_{\text{detour}}^{(l)}(x, x') = \mathbf{h}^{(l+1)}(x) - \mathbf{h}^{(l)}(x'), \tag{7}$$

$$\mathbf{v}_{\text{traj}}^{(l)}(x) = \mathbf{h}^{(l+1)}(x) - \mathbf{h}^{(l)}(x). \tag{8}$$

This triad establishes the local spread — measuring the spatial separation between samples at a single layer ($\mathbf{v}_{\text{sep}}$) and the detour required if we substituted x with x' before transitioning to the next layer ($\mathbf{v}_{\text{detour}}$). Analogously, we compute the corresponding cross-domain triad to capture inter-domain divergence: the cross-domain separation $\mathbf{v}_{\text{sep}}^{(l)}(x, x'')$, the cross-domain detour $\mathbf{v}_{\text{detour}}^{(l)}(x, x'')$, and the direct trajectory of the anchor sample $\mathbf{v}_{\text{traj}}^{(l)}(x)$.

To quantify the structural shift at layer l, we compute the angular alignment $\theta^{(l)}$ between the separation and detour vectors using cosine similarity, as well as the relative distance $d^{(l)}$. For the within-domain pair within Euclidean space, these are given by:

$$\theta^{(l)}(x, x') = \arccos\left(\frac{\mathbf{v}_{\text{sep}}^{(l)}(x, x') \cdot \mathbf{v}_{\text{detour}}^{(l)}(x, x')}{\max(\|\mathbf{v}_{\text{sep}}^{(l)}(x, x')\|\|\mathbf{v}_{\text{detour}}^{(l)}(x, x')\|, \epsilon)}\right), \tag{9}$$

$$d^{(l)}(x, x') = \frac{\|\mathbf{v}_{\text{sep}}^{(l)}(x, x')\| + \|\mathbf{v}_{\text{detour}}^{(l)}(x, x')\| - \|\mathbf{v}_{\text{traj}}^{(l)}(x)\|}{\max(\|\mathbf{v}_{\text{traj}}^{(l)}(x)\|, \epsilon)}, \tag{10}$$

where $\epsilon = 10^{-8}$ serves as a lower bound for the denominators to prevent division by zero. We compute the corresponding cross-domain angular alignment $\theta^{(l)}(x, x'')$ and relative distance $d^{(l)}(x, x'')$ analogously. By comparing the within-domain and cross-domain distributions of these quantities, we measure the overall domain divergence at a given layer. We motivate the necessity of both angles and distances in Appendix C.1.

The relative-distance descriptor $d^{(l)}$ admits a clean geometric interpretation. By construction, $v_{\text{sep}} + v_{\text{detour}} = v_{\text{traj}}$ exactly (proven in Appendix K), so the triangle inequality guarantees $d^{(l)} \geq 0$, with $d^{(l)}$ measuring the normalized excess path length of the two-step detour over the direct trajectory. Geometrically, $d^{(l)} = 0$ when the three representation points are collinear (i.e., when $h^{(l)}(x')$ lies exactly on the line segment from $h^{(l)}(x)$ to $h^{(l+1)}(x)$), which corresponds to $\theta^{(l)} = 0$, since both conditions are equivalent to v_{sep} and v_{detour} being positively collinear as vectors.

Euclidean vs. other spaces However, we observe that some models inherently lie on the surface of a hypersphere. Therefore, we propose a version of this metric defined on **geodesic** distances and angles. Let $d_S(u, v) = \arccos\left(\frac{u \cdot v}{\|u\|\|v\|}\right)$ denote the great-circle distance between two representations. We first define the arc lengths of the spherical triangle sides:

$$s_{\text{sep}} = d_S(\mathbf{h}^{(l)}(x), \mathbf{h}^{(l)}(x')), \tag{11}$$

$$s_{\text{detour}} = d_S(\mathbf{h}^{(l)}(x'), \mathbf{h}^{(l+1)}(x)), \tag{12}$$

$$s_{\text{traj}} = d_S(\mathbf{h}^{(l)}(x), \mathbf{h}^{(l+1)}(x)). \tag{13}$$

Using the Spherical Law of Cosines, we define the geodesic angular alignment $\theta_{\text{geo}}^{(l)}$ and the relative geodesic distance $d_{\text{geo}}^{(l)}$ as:

$$\theta_{\text{geo}}^{(l)}(x, x') = \arccos\left(\frac{\cos(s_{\text{traj}}) - \cos(s_{\text{sep}})\cos(s_{\text{detour}})}{\max(\sin(s_{\text{sep}})\sin(s_{\text{detour}}), \epsilon)}\right), \tag{14}$$

$$d_{\text{geo}}^{(l)}(x, x') = \frac{s_{\text{sep}} + s_{\text{detour}} - s_{\text{traj}}}{\max(s_{\text{traj}}, \epsilon)}. \tag{15}$$

As these angles and relative distances are subsequently used to measure distributional differences, which are invariant to constant shifts, we directly use the interior spherical angle rather than aligning it with the Euclidean directional formulation. Moreover, we can project these polar descriptors into a **pseudo-Cartesian** space:

$$x_{\text{cart}}^{(l)}(x, x') = d_{\text{geo}}^{(l)}(x, x') \cdot \cos(\theta_{\text{geo}}^{(l)}(x, x')), \tag{16}$$

$$y_{\text{cart}}^{(l)}(x, x') = d_{\text{geo}}^{(l)}(x, x') \cdot \sin(\theta_{\text{geo}}^{(l)}(x, x')). \tag{17}$$

The optimal topology reflects a fundamental cross-modal divide: text representations strictly favor Euclidean space, whereas vision representations perform best in pseudo-Cartesian space. As projecting text embeddings into pseudo-Cartesian space causes severe performance degradation while vision models remain robust to either choice, we adopt **Euclidean** space as the most consistent unified formulation. We motivate this design choice in Appendix C.3.

Divergence algorithms and long trajectory To measure the divergence between the within-domain and cross-domain feature distributions, we evaluate four different metrics:

- **GMM + KL**: Using the collected angular alignments and relative distances, we fit Gaussian Mixture Models (GMMs) on both the within-domain metric set for $\mathcal{D}_{\text{A}}$ and the cross-domain metric between $\mathcal{D}_{\text{A}}$ and $\mathcal{D}_{\text{B}}$, yielding two continuous probability distributions, P and Q, respectively. We then measure the structural divergence using a weighted symmetric Kullback-Leibler (KL) divergence: $\frac{|\mathcal{D}_{\text{A}}|}{|\mathcal{D}_{\text{A}}|+|\mathcal{D}_{\text{B}}|} D_{\text{KL}}(P \parallel Q) + \frac{|\mathcal{D}_{\text{B}}|}{|\mathcal{D}_{\text{A}}|+|\mathcal{D}_{\text{B}}|} D_{\text{KL}}(Q \parallel P)$.
- **GMM + SWD**: Similarly, we fit GMMs on $\mathcal{D}_{\text{A}}$ and $\mathcal{D}_{\text{B}}$ to form continuous distributions, but compute the optimal transport metric between them using the Sliced Wasserstein Distance (SWD).
- **Sinkhorn Divergence**: To avoid the parametric assumptions of GMMs, we directly compute the entropic regularized Wasserstein distance between the empirical samples of $\mathcal{D}_{\text{A}}$ and $\mathcal{D}_{\text{B}}$. This provides a computationally efficient, differentiable approximation of optimal transport.
- **Gaussian Maximum Mean Discrepancy (MMD)**: Finally, we evaluate the non-parametric MMD using a Gaussian (Radial Basis Function) kernel, which measures the divergence by calculating the distance between the mean embeddings of $\mathcal{D}_{\text{A}}$ and $\mathcal{D}_{\text{B}}$ in a Reproducing Kernel Hilbert Space (RKHS).

GMM + KL is the only universally robust algorithm across modalities and is therefore adopted as the default, though Sinkhorn divergence proves highly competitive for text representations specifically. Full comparisons across divergence algorithms and GMM covariance parameterizations are provided in Appendices C.5 and C.4.

Another benefit of our method is its capacity for extended trajectory tracing. While we have previously described a single structural shift from layer $l \to l+1$, we propose evaluating our metric over longer sequences. In this paper, we capture extended trajectories spanning 13 layers (from $l-6 \to \cdots \to l \to \cdots \to l+6$). As each of the 12 intermediate layer transitions yields a pair of descriptors, concatenating these step-wise measurements produces a comprehensive 24-dimensional feature vector representing the structural evolution of the sample pair. The GMM is subsequently fit directly on this 24-D space to capture the holistic multi-layer divergence. Since total variation between domains cannot increase with depth (Appendix J), a single late layer can understate differences visible at earlier layers. This justifies multi-layer extraction but not our specific descriptor, which we support empirically (Appendix C).

Sampling strategy. Computing this metric exactly across domains $\mathcal{D}_{\text{A}}$ and $\mathcal{D}_{\text{B}}$ requires $l \times (|\mathcal{D}_{\text{A}}||\mathcal{D}_{\text{B}}| + |\mathcal{D}_{\text{A}}|^2)$ descriptor evaluations, necessitating GMM fitting on millions of data points. To ensure computational feasibility, we subsample N distinct pairs for both the within-domain and cross-domain sets, such that $N \ll |\mathcal{D}_{\text{A}}|^2$ and $N \ll |\mathcal{D}_{\text{A}}||\mathcal{D}_{\text{B}}|$. This bounds our computational complexity to $\mathcal{O}(l \cdot N)$ per ($\mathcal{D}_{\text{A}}$,$\mathcal{D}_{\text{B}}$) pair.

To guarantee balanced coverage, we employ stratified sampling across class pairs (c_i, c_j). Within these strata, we apply a distance-based weighted sampling strategy rather than uniform random selection. For a given sample x, let $\bar{d}(x) \in [0, 1]$ be its min-max normalized Euclidean distance to

its domain-specific class centroid. We define its scalar inlier weight as $w(x) = 1 - \bar{d}(x)$, inherently assigning higher sampling probabilities to samples closer to the distribution center.

To obtain a comprehensive representation of the domain manifold dynamics, we construct our N pairs by sampling two distinct subsets for both within-domain (x, x') and cross-domain (x, x'') sets:

- **Inlier-Inlier pairs:** The first subset is drawn by weighting the probability of selecting a pair proportional to $w(x) \cdot w(x')$.
- **Inlier-Outlier pairs:** The second subset captures the boundaries of the distribution by weighting the pair selection proportional to $w(x) \cdot (1 - w(x'))$.

These sampled pairs are used to compute step-wise angular alignments and relative distances, which are concatenated into the feature vectors used for final divergence estimation. Comparisons across sampling strategies are provided in Appendix C.2.

While we have formulated this metric in a pairwise manner between two domains, our approach naturally extends to multi-source settings. Specifically, the secondary domain $\mathcal{D}_{\mathrm{B}}$ can be constructed as a blended pool of multiple source domains (i.e., $\mathcal{D}_{\mathrm{B}} = \bigcup_{d \in \mathcal{S}} \mathcal{D}_{\mathrm{d}}$). This allows our sampling strategy and subsequent divergence metrics to directly capture the structural shift between a target domain and an aggregated multi-source blend, satisfying our core objective.

4 Results

4.1 Experimental Setup

We evaluate our proposed method across two distinct problem settings: image classification (§ 4.2.1) and sentiment classification (§ 4.2.2). To demonstrate the versatility of our metric, we employ vision foundation models for the former and text embedding models for the latter. Motivated by recent findings that a model's final output layer is not guaranteed to yield the optimal representation for downstream transfer [3], we extract features from every transformer block rather than relying solely on the final representation. To establish a ground-truth measure of transferability, we train a lightweight two-layer MLP directly on these layer-wise features, averaging performance across $r = 10$ random seeds to ensure statistical stability. For a given target domain, we construct this ground-truth baseline by evaluating the MLP on all $K - 1$ pairs formed by combining the target data with one of the auxiliary source domains.

Dataset For image classification, we employ three standard domain adaptation benchmarks: DomainNet [35], OfficeHome [42], and PACS [25]. To keep evaluation computationally tractable, we subset DomainNet to the ten classes with the highest sample counts, and OfficeHome to the 20 most populated classes to maintain a consistent scale across datasets. PACS is used in its entirety given its smaller size.

For sentiment classification, we use the cross-lingual AmazonReview dataset from MTEB [31] and EuroEval [34], treating each language as a distinct domain—a natural choice given that language boundaries define the distribution shift. For EuroEval, we consider eight languages: English, German, Polish, Czech, Slovak, Hungarian, Italian, and Spanish. For AmazonReview, we use all six available languages. Full dataset details are provided in Appendix E.

Encoding backbones For our encoding backbones, we use the CLIP vision encoder [36] and the DINOv3 model [38] for image classification. For sentiment classification, we employ the Qwen3-Embedding [48] and EmbeddingGemma [43] models. We provide the full model name as listed on Hugging Face with the URL in Appendix L.

Baselines As a standard baseline, we compute the ℓ_2 distance between the feature centroids of two domains, $\mathcal{D}_A$ and $\mathcal{D}_B$. This centroid distance is naturally expected to be negatively correlated with accuracy improvements on the target domain—specifically, a smaller distance between domain centroids should correspond to higher transfer accuracy gains.

To rigorously evaluate our method against the state-of-the-art, we compare it against a diverse suite of transferability metrics applied to the same frozen features. These include label-conditional metrics

such as LEEP [32] and NCE [41]; feature-label compatibility scores including LogME [45], HScore [2], regularized HScore [19], and TransRate [13]; as well as distribution-matching and optimal transport approaches like s-OTDD [33].

Metrics We evaluate predictive performance by the Spearman rank correlation (ρ) between each metric's score and the ground-truth accuracy gains. Similarity metrics positively correlated with accuracy (e.g., LEEP [32]) are negated ($-X$), so negative ρ indicates better prediction for all methods. For baselines, ρ is computed per layer against the transfer gains at the same layer and averaged. For RADAR, the divergence from each trajectory window is correlated with the transfer gains at all layers and averaged. We primarily report the Mean Correlation Improvement (MCI), $\text{MCI}(X) = \rho_X - \rho_{\text{BASE}}$, where BASE is the ℓ_2-centroid baseline; as with ρ, lower is better. We focus on MCI for two reasons. First, attainable ρ varies by close to an order of magnitude across benchmarks, so raw values are not directly comparable. Second, MCI directly measures whether a metric improves over a cost-free baseline. Raw Spearman correlations are reported in Table 2, and computational cost is analyzed in Appendix I.

Configuration and hyperparameters All results use a single fixed configuration, with no per-dataset, per-backbone, or per-modality tuning: Euclidean geometry, a diagonal-covariance GMM with symmetric KL divergence, standardized features, and mixed sampling without replacement. We set $N = 32768$, $\ell = 6$, and $\tau = 2.0$. Detailed ablations are provided in Appendix C.6, where we show that RADAR remains stable across $N \geq 4096$, $\ell \in [2, 6]$, and $\tau \in [0.5, 4.0]$.

4.2 Assessing transferability

Table 1 presents the Mean Correlation Improvement (MCI) for various metrics across both vision and language modalities. In the visual domain, RADAR achieves the best predictive performance among all evaluated metrics on DomainNet for both CLIP and DINOv3 architectures, alongside the strongest results on OfficeHome for CLIP. Performance is more mixed across the remaining vision benchmarks, consistent with the domain-separation regime discussed in Section 4.2.1 and Appendix B. This strong performance on structured datasets extends to the textual domain, where RADAR yields the highest correlation on the EuroEval dataset for the EmbeddingGemma model. On the AmazonReviews dataset, RADAR achieves the best result among evaluated metrics for EmbeddingGemma while remaining competitive for the Qwen3-Embedding architecture, where gains are more modest.

Table 1: Mean Correlation Improvement (MCI, %pt.). Bold indicates the best performance, underline indicates second-best performance.

(a) Results on vision benchmarks.

	Mean Correlation Improvement (MCI) (%pt.) ↓					
	DomainNet [35]		**OfficeHome** [42]		**PACS** [25]	
Method	**CLIP**	**DINOv3**	**CLIP**	**DINOv3**	**CLIP**	**DINOv3**
LEEP [32]	<u>-8.63</u>	<u>3.44</u>	-4.52	3.92	-14.64	-33.67
H-Score [2]	0.92	19.13	8.77	21.98	-10.28	<u>-50.29</u>
Reg. H-Score [19]	6.03	17.11	8.13	26.95	-12.11	-48.89
LogME [45]	12.17	19.90	-4.41	18.67	-9.96	**-72.08**
NCE [41]	-4.88	4.15	-7.20	3.93	-13.61	-33.89
TransRate [13]	2.29	12.71	<u>-11.78</u>	<u>-3.12</u>	**-25.45**	-27.38
S-OTDD [33]	27.01	21.31	10.88	**-24.69**	<u>-23.92</u>	-37.50
RADAR (OUR)	**-9.71**	**-9.87**	**-16.62**	3.55	-0.27	-30.23

(b) Results on text benchmarks.

	Mean Correlation Improvement (MCI) (%pt.) ↓			
	EuroEval [34]		**Amazon Reviews** [31]	
Method	**Qwen3-Embedding**	**EmbeddingGemma**	**Qwen3-Embedding**	**EmbeddingGemma**
LEEP [32]	<u>-3.58</u>	3.94	1.76	<u>1.91</u>
H-Score [2]	9.27	11.75	-6.84	3.35
Reg. H-Score [19]	6.13	10.45	<u>-6.91</u>	3.35
LogME [45]	22.93	20.82	22.57	20.91
NCE [41]	-0.02	<u>1.88</u>	**-7.60**	5.71
TransRate [13]	3.12	4.10	12.47	19.48
S-OTDD [33]	6.19	18.71	15.63	19.69
RADAR (OUR)	**-7.13**	**-30.58**	3.21	**1.23**

Table 2 reports the raw mean Spearman correlation (ρ), complementing the MCI results in Table 1; as MCI subtracts a per-column constant, the rankings are identical. Excluding PACS, RADAR and NCE are the only metrics with negative ρ on every vision and text benchmark, and RADAR achieves the best ρ in 6 of the 10 benchmark–backbone settings, with its largest margin on EuroEval with EmbeddingGemma.

4.2.1 Case study: Image Classification

As presented in Table 1a, RADAR demonstrates strong predictive performance on two of the three evaluated cross-domain datasets. On DomainNet, our method is the only evaluated metric to achieve a negative MCI for the DINOv3 architecture, while also securing the top overall score for the CLIP model. Furthermore, on the OfficeHome dataset, RADAR establishes the best overall score for CLIP and remains highly competitive for DINOv3, achieving the third-best rank among all tested metrics.

Table 2: Raw Mean Spearman Rank correlation (%pt., ↓). Bold indicates the best performance, underline indicates second-best performance.

(a) Results on vision benchmarks.

	Mean Spearman Correlation (%pt.) ↓					
	DomainNet [35]		**OfficeHome** [42]		**PACS** [25]	
Method	**CLIP**	**DINOv3**	**CLIP**	**DINOv3**	**CLIP**	**DINOv3**
LEEP [32]	-51.65	-31.16	0.16	-3.42	-4.09	12.00
H-Score [2]	-42.10	-15.47	13.45	14.63	0.27	-4.62
Reg. H-Score [19]	-37.00	-17.50	12.80	19.61	-1.56	-3.23
LogME [45]	-30.86	-14.70	0.27	11.33	0.59	**-26.41**
NCE [41]	-47.91	-30.46	-2.53	-3.42	-3.07	11.78
TransRate [13]	-40.73	-21.90	-7.10	-10.47	**-14.90**	18.29
S-OTDD [33]	-16.01	-13.29	15.55	**-32.03**	-13.37	8.26
RADAR (OUR)	**-52.74**	**-44.47**	**-11.94**	-3.80	10.27	15.44

(b) Results on text benchmarks.

	Mean Spearman Correlation (%pt.) ↓			
	EuroEval [34]		**Amazon Reviews** [31]	
Method	**Qwen3-Embedding**	**EmbeddingGemma**	**Qwen3-Embedding**	**EmbeddingGemma**
LEEP [32]	-4.61	-1.66	-17.29	-11.27
H-Score [2]	8.24	6.15	-25.89	-9.84
Reg. H-Score [19]	5.10	4.85	-25.96	-9.83
LogME [45]	21.90	15.21	3.53	7.72
NCE [41]	-1.05	-3.72	**-26.64**	-7.47
TransRate [13]	2.10	-1.50	-6.58	6.29
S-OTDD [33]	5.16	13.11	-3.40	6.50
RADAR (OUR)	**-8.16**	**-36.18**	-15.83	**-11.96**

We observe a reduced performance of RADAR on PACS, nonetheless, this is predictable from the dataset's geometry. We show in Appendices A and B, that inter-domain centroid distances in PACS closely mirror ImageNet-C at Severity 3, placing it in an intermediate separation regime that is neither smooth enough for trajectory-based adaptation nor distinct enough for clean divergence estimation. This same regime is precisely where RADAR degrades when evaluated on ImageNet-C (Table 3 in Appendix A), confirming that the PACS result reflects a principled scope condition rather than a general weakness. When domains exhibit scattered, non-uniform separation at the input level, as is the case with PACS, the added utility of measuring deep geometric trajectories diminishes, allowing purely statistical or static cluster-based metrics to perform highly competitively.

4.2.2 Case study: Sentiment Analysis

Similar to the image classification results, RADAR demonstrates strong predictive performance in the text modality, achieving the best or near-best score in three of the four evaluated configurations, with more modest gains on the Qwen3-Embedding model for Amazon Reviews (Table 1b). On the EuroEval dataset, RADAR is the only evaluated metric to achieve a negative MCI for the EmbeddingGemma model, outperforming the next best baseline by over 30 percentage points. Furthermore, for the Qwen3-Embedding model on EuroEval, RADAR secures the top score among the three metrics capable of achieving a negative MCI. On the Amazon Reviews dataset, RADAR establishes the state-of-the-art result for the EmbeddingGemma model while maintaining a highly competitive score for Qwen3-Embedding. Crucially, metrics that performed competitively on image classification tasks — such as LogME, TransRate, and S-OTDD — struggle significantly in the text modality. This stark contrast underscores RADAR's robust, cross-modal applicability.

5 Related Work

5.1 Transferability Metrics

Situated within the broader domain adaptation literature [5, 28], domain transferability metrics aim to quantify the extent to which candidate source data can facilitate learning on a target task under distribution shift [44]. A critical objective for these metrics is predicting negative transfer [47], a phenomenon where incorporating source information actively degrades target performance. Indeed, a recent survey emphasizes the urgent need for reliable metrics capable of detecting such degradation scenarios, highlighting the importance of studying negative findings and counterexamples [4]. In the following sections, we categorize existing metrics based on their computational paradigms: those relying on the discrete, empirical computation of domain shift, and those formulated using optimal transport algorithms.

Empirical Metrics Early transferability metrics focused on empirically measuring the relationship between various tasks. Given a classifier trained on a source domain, LEEP [32] computes the average log-likelihood of the expected empirical predictor—a simplified model that performs prediction based on the expected empirical conditional distribution between source and target label sets. Analogous to LEEP, NCE [41] quantifies transferability via the negative conditional entropy between these label spaces. In contrast, the H-Score [2] is derived from information-theoretic principles and solves the Hirschfeld-Gebelein-Rényi maximum correlation problem [11, 7, 37]. This approach was later refined

by the Regularized H-Score [19], which addresses the instability of covariance matrix estimation through a shrinkage-based regularizer.

Conversely, LogME [45] employs Bayesian statistics to maximize the marginal likelihood of a linear classifier head. Another metric, TransRate [13], estimates the mutual information between labels and target features through coding rates [29], offering a more geometrically grounded transferability signal than entropy-based approaches. Despite their varying methodological foundations, all of these approaches evaluate transferability from static representations at a fixed network depth, without modeling how domain structure dynamically evolves across the model's hierarchy.

Optimal Transport-based Metrics To address the limitations of empirical label-dependent metrics, Optimal Transport (OT) based approaches aim to measure distributional distances directly by modeling the underlying data geometry. OTCE [39] treats the OT plan between input distributions as a joint probability distribution and leverages conditional entropy to quantify the gap between datasets. In follow-up work, F-OTCE [40] simplifies this process by removing the Wasserstein distance from the domain difference calculation, instead solving the OT between source and target distributions to compute negative conditional entropy via optimal coupling. Conversely, JC-OTCE [40] improves transferability estimation by incorporating label distances directly into the OT problem.

Optimal Transport Dataset Distance (OTDD) [1], built upon the hierarchical OT framework [46], proposes calculating label distances between the class-conditional distributions of two supervised learning tasks. This approach yields a dataset distance that accounts for both sample and label discrepancies. This framework was recently optimized by s-OTDD [33], which utilizes Moment Transform Projection and sliced Wasserstein distance to achieve (near-)linear computational complexity. Similarly prioritizing computational efficiency, WTE [27] embeds datasets into a vector space and computes Euclidean distances as a lightweight approximation of optimal transport.

Despite their theoretical rigor, all of these approaches focus exclusively on static feature distributions at a single representation level. Consequently, they fail to capture the dynamic, layer-wise geometric evolution that RADAR leverages to characterize model transferability.

5.2 Geometric Structure of Foundation Model Representations

The general inductive biases of deep neural networks often induce a *cone effect* [26], where the effective embedding space of pre-trained models is restricted to a narrow cone, and contrastive learning actively preserves the modality gap. While this work suggests that widening the gap can enhance zero-shot performance, the follow-up work, MG-CLIP [12], demonstrates the existence of an optimal gap between modalities. The authors show that maintaining a stable modality gap is essential to preserve pre-trained knowledge and prevent representational degradation. In the specific context of CLIP, simply including a domain in the training data does not guarantee improved generalization, highlighting sufficient domain diversity as a crucial prerequisite [23]. Furthermore, Cross The Gap [30] shifts the focus to intra-modal misalignment, establishing that the structural integrity of representations within a single modality is equally important to inter-modal alignment.

Beyond modality-specific structures, the Platonic Representation Hypothesis [18] posits that foundation models, regardless of their training objectives or modalities, converge toward a shared statistical model of reality. This suggests that the geometric properties of representation spaces reflect fundamental characteristics of the underlying data manifold rather than mere architectural artifacts. Consistent with this view, recent work demonstrates that DINO features naturally reside on the surface of a hypersphere [24]. These insights collectively motivate the design of RADAR: by measuring angular alignment and relative distance across network layers, RADAR captures domain-specific geometric deformations within these representation spaces to directly assess cross-domain compatibility.

5.3 Perceptual Straightening and Walking Through a Network

The temporal straightening hypothesis suggests that biological visual systems facilitate the processing of dynamic stimuli by transforming highly curved visual trajectories into straightened paths within the brain's internal representation spaces [9, 8]. Recent applications of this hypothesis in AI-generated video detection utilize angles of inter-frame features to distinguish the low-curvature trajectories of real videos from the higher-curvature paths of synthetic ones [20]. Inspired by this foundation, we extend the formulation to domain transferability, treating the layer-wise progression of a deep neural

network as a representational trajectory. Incorporating relative distance changes between layers, we assess how the geometric structure of these paths reflects cross-domain structural shifts. This perspective is further motivated by recent findings that a model's most informative representations are not its final layer's [3], justifying our extraction of features across multiple transformer blocks.

6 Limitations

While RADAR demonstrates competitive performance across diverse datasets, several limitations remain. First, its predictive power degrades when domains lack clear, uniform geometric separation (e.g., PACS). This regime is, however, diagnosable in advance: layer-0 centroid distances, obtained from a single forward pass, place PACS alongside ImageNet-C-3 (Appendix B), so input-level separation can be checked before relying on RADAR. Second, the unified Euclidean formulation, while the most consistent choice across tested architectures, may not be optimal for future models operating on hyperspheres or other non-Euclidean manifolds, potentially requiring architecture-specific tuning. Finally, our reliance on GMMs for density estimation is a computational bottleneck: the lack of a standardized GPU implementation for GMM fitting limits scalability for real-time filtering tasks.

7 Future Work

We intend to extend RADAR with label-aware components, similar to TransRate and LogME, to further improve predictive performance. We also plan to explore dynamic weighting of layer positions to prioritize the most informative segments of the trajectory, which could clarify the conditions under which the metric fails, and to investigate formal guarantees linking RADAR's trajectory descriptor to downstream transfer performance. Finally, we aim to extend RADAR beyond classification and image or text data. Some tasks, such as retrieval, require only redefining the ground truth (e.g., recall@k), whereas others, such as detection and segmentation, require a per-region ground truth. As for additional modalities, we plan to study video and audio, to test whether these geometric principles generalize to temporal and spectral representations.

8 Conclusion

In this paper, we proposed RADAR, a geometrically grounded metric for predicting cross-domain transferability by measuring the KL divergence between angular and relative displacement distributions in foundation model feature spaces. Our evaluation demonstrates that layer-wise geometric structure carries a meaningful transferability signal across both vision and text modalities—a finding that most static, single-layer metrics fail to capture. We hope RADAR's cross-modal applicability and competitive performance encourage further investigation into the geometry of representation spaces as a principled basis for transfer learning decisions.

Acknowledgments and Disclosure of Funding

This research was funded by the Defense Advanced Research Projects Agency (DARPA), under contract W912CG23C0031.

References

[1] Alvarez-Melis, D., Fusi, N.: Geometric dataset distances via optimal transport. Advances in Neural Information Processing Systems **33**, 21428–21439 (2020)

[2] Bao, Y., Li, Y., Huang, S.L., Zhang, L., Zheng, L., Zamir, A., Guibas, L.: An information-theoretic approach to transferability in task transfer learning. In: 2019 IEEE international conference on image processing (ICIP). pp. 2309–2313. IEEE (2019)

[3] Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)

[4] Dong, H., Liu, M., Zhou, K., Chatzi, E., Kannala, J., Stachniss, C., Fink, O.: Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)

[5] Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020 pp. 877–894 (2021)

[6] Feydy, J., Roussillon, P., Trouvé, A., Gori, P.: Fast and scalable optimal transport for brain tractograms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–644. Springer (2019)

[7] Gebelein, H.: Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik **21**(6), 364–379 (1941)

[8] Hénaff, O.J., Bai, Y., Charlton, J.A., Nauhaus, I., Simoncelli, E.P., Goris, R.L.: Primary visual cortex straightens natural video trajectories. Nature communications **12**(1), 5982 (2021)

[9] Hénaff, O.J., Goris, R.L., Simoncelli, E.P.: Perceptual straightening of natural videos. Nature neuroscience **22**(6), 984–991 (2019)

[10] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)

[11] Hirschfeld, H.O.: A connection between correlation and contingency. Mathematical Proceedings of the Cambridge Philosophical Society **31**(4), 520–524 (1935). https://doi.org/10.1017/S0305004100013517

[12] Huang, L., Cao, X., Lu, H., Meng, Y., Yang, F., Liu, X.: Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3777–3786 (2025)

[13] Huang, L.K., Huang, J., Rong, Y., Yang, Q., Wei, Y.: Frustratingly easy transferability estimation. In: International conference on machine learning. pp. 9201–9225. PMLR (2022)

[14] Hugging Face: Embeddinggemma model card. `https://huggingface.co/google/embeddinggemma-300m` (2026), accessed: 2026

[15] Hugging Face: Model card: Clip. `https://huggingface.co/openai/clip-vit-base-patch32` (2026), accessed: 2026

[16] Hugging Face: Model card for dinov3. `https://huggingface.co/facebook/dinov3-vits16-pretrain-lvd1689m` (2026), accessed: 2026

[17] Hugging Face: Qwen3-embedding-0.6b. `https://huggingface.co/Qwen/Qwen3-Embedding-0.6B` (2026), accessed: 2026

[18] Huh, M., Cheung, B., Wang, T., Isola, P.: Position: The platonic representation hypothesis. In: Forty-first International Conference on Machine Learning (2024)

[19] Ibrahim, S., Ponomareva, N., Mazumder, R.: Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 693–709. Springer (2022)

[20] Internò, C., Geirhos, R., Olhofer, M., Liu, S., Hammer, B., Klindt, D.: AI-generated video detection via perceptual straightening. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), `https://openreview.net/forum?id=LsmUgStXby`

[21] Jiang, J., Shu, Y., Wang, J., Long, M.: Transferability in deep learning: A survey. arXiv preprint arXiv:2201.05867 (2022)

[22] Junguang, J., Baixu, C., Bo, F., Mingsheng, L.: Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library` (2020)

[23] Kempf, E., Schrodi, S., Argus, M., Brox, T.: When and how does clip enable domain and compositional generalization? arXiv preprint arXiv:2502.09507 (2025)

[24] Kumar, A., Patel, V.M.: Learning on the manifold: Unlocking standard diffusion transformers with representation encoders. arXiv preprint arXiv:2602.10099 (2026)

[25] Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization (2017), https://arxiv.org/abs/1710.03077

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

[27] Liu, X., Bai, Y., Lu, Y., Soltoggio, A., Kolouri, S.: Wasserstein task embedding for measuring task similarities. Neural Networks **181**, 106796 (2025)

[28] Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015)

[29] Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy data coding and compression. IEEE transactions on pattern analysis and machine intelligence **29**(9), 1546–1562 (2007)

[30] Mistretta, M., Baldrati, A., Agnolucci, L., Bertini, M., Bagdanov, A.D.: Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=VVVfuIcmKR

[31] Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2014–2037 (2023)

[32] Nguyen, C., Hassner, T., Seeger, M., Archambeau, C.: Leep: A new measure to evaluate transferability of learned representations. In: International conference on machine learning. pp. 7294–7305. PMLR (2020)

[33] Nguyen, K., Nguyen, H., Pham, T., Ho, N.: Lightspeed geometric dataset distance via sliced optimal transport. arXiv preprint arXiv:2501.18901 (2025)

[34] Nielsen, D.S., Enevoldsen, K., Schneider-Kamp, P.: Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual nlu tasks. In: Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025). pp. 561–572 (2025)

[35] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019)

[36] 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. PmLR (2021)

[37] Rényi, A.: On measures of dependence. Acta Mathematica Academiae Scientiarum Hungarica **10**(3), 441–451 (1959)

[38] Simoni, O., et al.: Dinov3: Self-supervised learning for vision at unprecedented scale. arXiv preprint arXiv:2508.10104 (2025)

[39] Tan, Y., Li, Y., Huang, S.L.: Otce: A transferability metric for cross-domain cross-task representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15779–15788 (2021)

[40] Tan, Y., Zhang, E., Li, Y., Huang, S.L., Zhang, X.P.: Transferability-guided cross-domain cross-task transfer learning. IEEE Transactions on Neural Networks and Learning Systems **36**(2), 2423–2436 (2024)

[41] Tran, A.T., Nguyen, C.V., Hassner, T.: Transferability and hardness of supervised classification tasks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1395–1405 (2019)

[42] Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5018–5027 (2017)

[43] Vera, H.S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S.R., Smoot, S., Naim, I., Zou, J., Chen, F., et al.: Embeddinggemma: Powerful and lightweight text representations. arXiv preprint arXiv:2509.20354 (2025)

[44] Wang, H., Wang, J., Zhao, Z., Tan, Y., Wu, Y., Liu, H., Yang, J., Zhang, E., Chen, X., Rong, Z., et al.: Understanding knowledge transferability for transfer learning: A survey. arXiv preprint arXiv:2507.03175 (2025)

[45] You, K., Liu, Y., Wang, J., Long, M.: Logme: Practical assessment of pre-trained models for transfer learning. In: International conference on machine learning. pp. 12133–12143. PMLR (2021)

[46] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., Solomon, J.M.: Hierarchical optimal transport for document representation. Advances in neural information processing systems **32** (2019)

[47] Zhang, W., Deng, L., Zhang, L., Wu, D.: A survey on negative transfer. IEEE/CAA Journal of Automatica Sinica **10**(2), 305–329 (2022)

[48] Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)

A RADAR robustness on ImageNet-C dataset

In this section, we examine the robustness of RADAR using the ImageNet-C dataset, which allows us to simulate synthetic domains with varying degrees of distribution shift by controlling corruption severity. Figure 3 illustrates these shifts by depicting a single sample across different severity levels. To quantitatively assess the resilience of various transferability metrics against these perturbations, we report MCI in Table 3 across three representative severity levels—1, 3, and 5.

Table 3: Performance comparison on the ImageNet-C dataset across increasing corruption severity levels (1, 3, and 5). While baseline statistical metrics like TransRate perform well under moderate synthetic noise (Severity 3), RADAR demonstrates exceptional robustness at extreme corruption levels (Severity 5) and competitive performance at low noise levels (Severity 1) for both architectures.

	Mean Correlation Improvement (MCI) (%pt.) ↓					
	ImageNet-C-1		**ImageNet-C-3**		**ImageNet-C-5**	
Method	**CLIP**	**DINOv3**	**CLIP**	**DINOv3**	**CLIP**	**DINOv3**
LEEP [32]	**7.61**	5.13	17.72	19.23	35.96	20.39
H-Score [2]	9.46	-1.10	12.26	-1.99	18.21	-11.38
Reg. H-Score [19]	12.44	**-1.82**	13.52	-1.36	24.33	N/A
LogME [45]	13.43	17.06	13.63	1.09	19.24	-19.51
NCE [41]	12.21	2.87	19.72	19.18	42.95	24.57
TransRate [13]	10.57	7.54	**5.73**	**-7.38**	**-25.68**	-25.07
S-OTDD [33]	19.20	20.61	22.37	22.60	4.83	-0.46
RADAR (OUR)	**7.61**	-0.25	20.69	8.82	1.98	**-30.16**

Under low corruption conditions (Severity 1), RADAR demonstrates highly competitive baseline stability. For the CLIP architecture, RADAR ties with LEEP for the best performance; similarly, for DINOv3, it maintains a robust third-place ranking that closely rivals top-performing statistical benchmarks such as Regularized H-Score. These results suggest that at minimal noise levels, our geometric trajectory approach effectively captures underlying domain shifts without being significantly derailed by minor perturbations.

As corruption increases to moderate levels (Severity 3), we observe a shift in metric efficacy. Metrics such as TransRate achieve the lowest MCI results across both architectures in this regime. In contrast, RADAR experiences a comparative reduction in predictive performance at this level. This suggests that moderate synthetic noise may temporarily disrupt the trajectory alignments upon which our method relies, creating a regime that favors metrics with a learnable feature component like TransRate.

However, the most striking results emerge under extreme corruption (Severity 5). As the input space becomes heavily degraded, traditional metrics experience catastrophic failures; for instance, LEEP and NCE both exceed an MCI of 35.0 for the CLIP architecture. In contrast, RADAR demonstrates exceptional robustness. For DINOv3, RADAR achieves the optimal score of -30.16, significantly outperforming all baselines. Similarly, for CLIP, it remains highly resilient, outperforming all statistical benchmarks except for TransRate.

These findings suggest a compelling structural advantage: while moderate noise can perturb local trajectories, severe corruption uniformly distorts the embedding space, causing static metrics to collapse. Because RADAR measures the *relative angular divergence* of how representations evolve across layers, it remains remarkably invariant to extreme spatial distortions, preserving its ability to accurately predict model transferability even when the underlying data is heavily corrupted. Nevertheless, we note that when the perturbations are moderate, the metric's performance degrades. This indicates that RADAR is highly competitive when domain shifts are either minimal (Severity 1) or extreme (Severity 5), but exhibits vulnerabilities to intermediate, unstructured perturbations (Severity 3).



Figure 3: **Visualizing ImageNet-C synthetic corruptions.** Examples of various visual corruptions across increasing severities (0, 1, 3, and 5). The profound structural degradation at Severity 5 fundamentally scrambles patch-level token distributions in Vision Transformers. As shown in Table 3, RADAR successfully tracks these chaotic geometric trajectories where most of the standard statistical metrics fail.

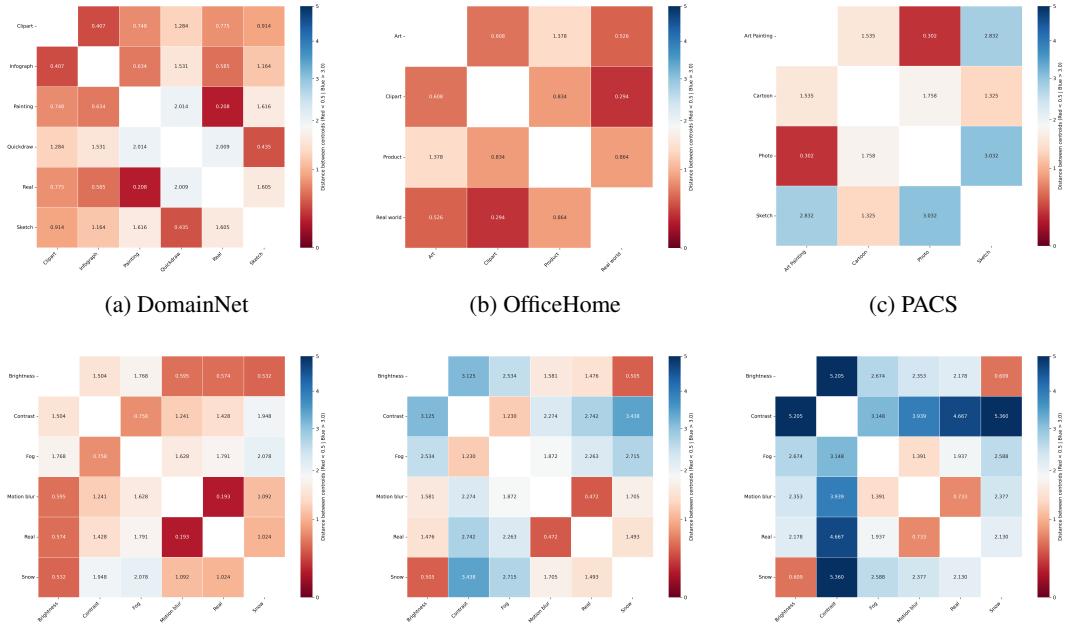
B Domain Separation — Comparing ImageNet-C and other datasets

Building upon the discussion in Appendix A, we investigate the domain separation within the first-layer feature spaces of both CLIP and DINOv3. This analysis allows us to assess the domain separability inherent in standard benchmarks—DomainNet [35], OfficeHome [42], and PACS [25]—relative to the varying corruption severities of ImageNet-C. The resulting feature-space visualizations are presented in Figures 4 and 5.

As demonstrated in Figure 4, the distances between domain feature centroids in DomainNet and OfficeHome closely resemble those of the low-severity corruptions in ImageNet-C (Severity 1). This indicates that the feature representations of these domains exhibit smooth transitions and modest distribution shifts. A notable exception is observed in the distances between the Quickdraw–Painting and Quickdraw–Real domain pairs; these demonstrate severe feature separation, mirroring the more pronounced spatial divergences characteristic of ImageNet-C-3 and ImageNet-C-5.

The PACS dataset, however, exhibits a markedly different pattern. Its inter-domain distances closely mirror those of ImageNet-C-3, indicating a moderate degree of domain separation. This offers a principled explanation for the reduced efficacy of our method on PACS: as the ImageNet-C experiments demonstrate, our approach performs optimally when domains either transition gradually or are highly separated in feature space. PACS appears to fall into an intermediate regime—analogous to ImageNet-C-3—characterized by partial overlap that is neither structured enough for smooth trajectory alignment nor distinct enough for clear separation.

An analysis of the inter-feature distances in the first layer of DINOv3, presented in Figure 5, corroborates our earlier findings in the CLIP feature space. We observe that DomainNet and OfficeHome features exhibit behavior consistent with the lower ImageNet-C corruption severities (1 and 3). In contrast, PACS exhibits a broader variance, resembling the distributions associated with severities 3 and 5. This pattern suggests that domain separation in PACS is inherently challenging regardless of the underlying architecture. Specifically, it consistently exhibits the ambiguous, intermediate spatial



(a) DomainNet (b) OfficeHome (c) PACS

(d) ImageNet-C-1 (e) ImageNet-C-3 (f) ImageNet-C-5

Figure 4: **Visualization of Layer 0 centroid distances across datasets for the CLIP model.** Distance matrices illustrate the spatial separation between domains right at the input embedding level. Warmer hues denote smaller centroid distances between domains, whereas cooler hues denote larger separations. At Layer 0, the domain distance distributions of DomainNet and OfficeHome closely mirror the low-severity corruptions of ImageNet-C-1. Conversely, PACS exhibits greater initial domain separation, more closely resembling the distribution seen in ImageNet-C-3.

overlaps (analogous to Severity 3) observed in the CLIP setting—a regime in which our method demonstrably struggles.

Critically, the scope of the separability condition is diagnosable in advance from Layer 0 centroid distances alone, suggesting a practical pre-screening step: compute inter-domain separation at the input embedding level before selecting a transferability metric.

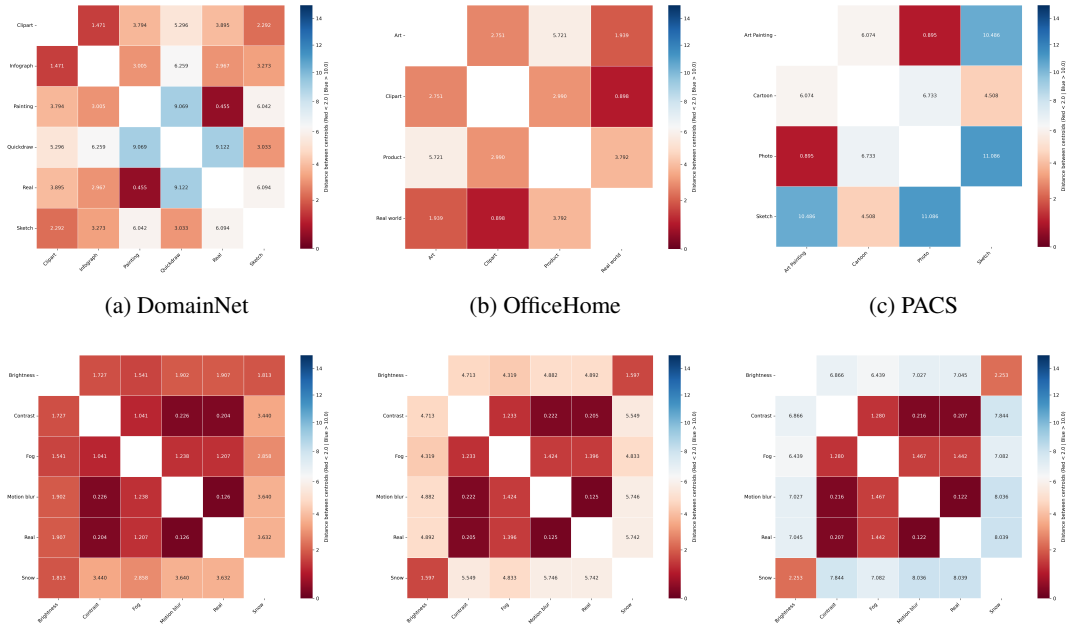
C Ablation Studies

C.1 Feature Representation and Standardization

We conduct an extensive ablation study to isolate the impact of specific geometric features and representation choices within RADAR. Table 4 presents the results of these geometric design choices across vision and text modalities.

We observe a universal synergy between distance and angular features for both modalities. Omitting either the angles or relative distances leads to severe performance degradation, with distance-only features even negatively impacting the CLIP and Qwen3-Embedding models. Notably, text representations exhibit a disproportionate reliance on angular features, yielding a total score of -25.10 alone compared to -4.68 for distance. This aligns with the widespread use of cosine-based contrastive objectives in language modeling, which inherently prioritizes angular topology.

Furthermore, we uncover a modality-specific divergence regarding feature standardization. While vision models suffer slightly from standardizing distance and angle scales, indicating that raw magnitude variance can contain useful transferability signals, text models seem to benefit from it. Standardizing text representations neutralizes magnitude noise, allowing the combined distance and angle metric to achieve the optimal overall score of -37.46, improving the MCI by roughly 4 percentage points over the unstandardized version. Despite the minor performance drop in vision models, we opt to apply feature standardization to both modalities to maintain a unified, modality-



(a) DomainNet (b) OfficeHome (c) PACS

(d) ImageNet-C-1 (e) ImageNet-C-3 (f) ImageNet-C-5

Figure 5: **Visualization of Layer 0 centroid distances across datasets for the DINOv3 model.** Distance matrices illustrate the spatial separation between domains right at the input embedding level. Warmer hues denote smaller centroid distances between domains, whereas cooler hues denote larger separations. At Layer 0, the domain distance distributions of DomainNet and OfficeHome closely mirror the ImageNet-C-1 and ImageNet-C-3 corruption severities. Conversely, PACS exhibits greater initial domain separation once more, closely resembling the distributions seen in ImageNet-C-3 and ImageNet-C-5.

Table 4: Feature ablations for vision (DomainNet) and text (EuroEval) representations. Both modalities exhibit a strong preference for combining Distance and Angle features, though text representations show heavy reliance on angular displacement, and only text benefits from layer-wise standardization.

(a) Ablation study on the vision models CLIP [36] and DINOv3 [38]. We observe both angles and distances are crucial.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	CLIP	DINOv3	Total
Features	✗	✓	✗	Uniform	✗	Euclidean	Diag	GMM + KL	-2.66	-5.62	-8.28
	✓	✗	✗	Uniform	✗	Euclidean	Diag	GMM + KL	1.34	-7.96	-6.62
	✓	✓	✗	Uniform	✗	Euclidean	Diag	GMM + KL	**-10.25**	-7.48	**-17.73**
	✓	✓	✓	Uniform	✗	Euclidean	Diag	GMM + KL	-8.85	**-8.31**	-17.16

(b) Ablation study on the text models Qwen3-Embedding [48] and EmbeddingGemma [43]. We observe standardization provides a massive boost to text modalities, unlike vision models.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	Qwen3-Embedding	EmbeddingGemma	Total
Features	✗	✓	✗	Uniform	✗	Euclidean	Diag	GMM + KL	-5.68	-19.42	-25.10
	✓	✗	✗	Uniform	✗	Euclidean	Diag	GMM + KL	7.42	-12.10	-4.68
	✓	✓	✗	Uniform	✗	Euclidean	Diag	GMM + KL	-4.24	-29.29	-33.53
	✓	✓	✓	Uniform	✗	Euclidean	Diag	GMM + KL	**-6.71**	**-30.75**	**-37.46**

agnostic framework, as the substantial gains in text evaluation far outweigh the slight degradation in vision.

C.2 Pairwise Sampling Strategies

Next, we evaluate the pairwise sampling used in our density estimation of angles and relative distances. Table 5 details the ablation comparing uniform sampling, positive (inlier-inlier) pair sampling, negative (inlier-outlier) pair sampling, and a mixture of both.

Table 5: Sampling ablations for vision (DomainNet) and text (EuroEval) representations. Both modalities demonstrate that "Mix" sampling — incorporating both positive (inlier-inlier) and negative (inlier-outlier) pairs — yields the strongest predictive correlation, proving that mapping both class centers and class boundaries is essential.

(a) Ablation study on the vision models CLIP [36] and DINOv3 [38]. We observe that Mixed sampling with replacement yields a marginal peak in total performance.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	CLIP	DINOv3	Total
Sampling	✓	✓	✓	Positive	✗	Euclidean	Diag	GMM + KL	-9.25	-8.76	-18.01
	✓	✓	✓	Negative	✗	Euclidean	Diag	GMM + KL	-8.45	-8.65	-17.10
	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	**-9.72**	-9.87	-19.59
	✓	✓	✓	Mix	✓	Euclidean	Diag	GMM + KL	-8.03	**-11.58**	**-19.61**

(b) Ablation study on the text models Qwen3-Embedding [48] and EmbeddingGemma [43]. Mixed sampling significantly outperforms isolated sampling, with "no replacement" yielding the best results.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	Qwen3-Embedding	EmbeddingGemma	Total
Sampling	✓	✓	✓	Positive	✗	Euclidean	Diag	GMM + KL	-7.03	-29.51	-36.54
	✓	✓	✓	Negative	✗	Euclidean	Diag	GMM + KL	-6.85	-28.19	-35.04
	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	-7.13	**-30.58**	**-37.71**
	✓	✓	✓	Mix	✓	Euclidean	Diag	GMM + KL	**-7.28**	-28.40	-35.68

The results clearly demonstrate that the "Mix" strategy strictly dominates isolated sampling across both vision and text modalities. Relying solely on positive pairs captures only intra-class compactness, whereas relying on negative pairs maps only inter-class separation. By integrating both, RADAR comprehensively captures the topological structure of the representation space—mapping both class centers and decision boundaries—which yields the strongest predictive correlations overall, surpassing uniform sampling.

Furthermore, we evaluate the effect of sampling pairs with versus without replacement. Text modalities exhibit a strict preference for sampling without replacement, achieving an optimal total score of -37.71 compared to -35.68 with replacement, improving the MCI by roughly 2 percentage points. In those models, sampling with replacement risks reducing the geometric diversity of the sampled trajectories. Conversely, vision models exhibit high variance. While DINOv3 benefits from replacement, CLIP's performance degrades significantly. Given these dynamics, we conclude that mixed sampling *without* replacement offers the most stable and universally robust configuration across diverse foundation models.

C.3 Topological Space Formulation

As a subsequent evaluation, we analyze the topological space in which these angles and relative distances are computed. Table 6 compares the effects of computing the representations in Euclidean, Geodesic, and pseudo-Cartesian spaces prior to density estimation.

The data shows a stark cross-modal divide regarding the optimal underlying geometry. Text models exhibit a strict reliance on Euclidean space, achieving an optimal total MCI of -37.71, while projecting text embeddings into Geodesic or Cartesian spaces results in massive performance degradation.

Conversely, vision models achieve their highest total performance in a pseudo-Cartesian space. However, this total is heavily skewed by DINOv3, which uniquely thrives in Cartesian projection. CLIP, much like the text models, actually favors Euclidean space over the alternatives.

When establishing a unified, modality-agnostic configuration for RADAR, the Euclidean space emerges as the clear optimal middle-ground. It serves as the absolute best geometric space for both text models and CLIP, while remaining highly competitive for DINOv3. The severe degradation text

Table 6: Topological space ablations for vision (DomainNet) and text (EuroEval) representations. The data reveals a clear cross-modal divide: vision representations achieve peak predictability in pseudo-Cartesian space, whereas text representations heavily favor Euclidean space. However, Euclidean representation seems to be the best middle-ground.

(a) Ablation study on the vision models CLIP [36] and DINOv3 [38]. The Cartesian projection yields the best overall performance, largely driven by DINOv3.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	CLIP	DINOv3	Total
Space	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	**-9.72**	-9.87	-19.59
	✓	✓	✓	Mix	✗	Geodesic	Diag	GMM + KL	-6.74	-11.37	-18.11
	✓	✓	✓	Mix	✗	Cartesian	Diag	GMM + KL	-8.52	**-12.99**	**-21.51**

(b) Ablation study on the text models Qwen3-Embedding [48] and EmbeddingGemma [43]. Both text foundation models strongly degrade when moved away from Euclidean space.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	Qwen3-Embedding	EmbeddingGemma	Total
Space	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	**-7.13**	**-30.58**	**-37.71**
	✓	✓	✓	Mix	✗	Geodesic	Diag	GMM + KL	-4.79	-25.60	-30.39
	✓	✓	✓	Mix	✗	Cartesian	Diag	GMM + KL	-5.86	-24.31	-30.17

models experience in Cartesian and Geodesic spaces far outweighs the isolated gains observed in DINOv3, making Euclidean space the most robust universal standard.

C.4 GMM Covariance Parameterization

To evaluate how potential correlations between the angle and relative distance distributions impact density estimation, we ablate the design choice of the GMM covariance matrix. Table 7 presents this ablation for both vision and text representations, comparing Diagonal, Full, Tied, and Spherical parameterizations.

Table 7: GMM covariance ablations for vision (DomainNet) and text (EuroEval) representations. Both modalities show a strict preference for Diagonal covariance parameterization, avoiding the overfitting associated with Full covariance and the over-restriction of Spherical covariance.

(a) Ablation study on the vision models CLIP [36] and DINOv3 [38]. Full covariance severely degrades DINOv3's predictive correlation.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	CLIP	DINOv3	Total
Covariance	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	**-9.72**	**-9.87**	**-19.59**
	✓	✓	✓	Mix	✗	Euclidean	Full	GMM + KL	-8.12	-3.80	-11.92
	✓	✓	✓	Mix	✗	Euclidean	Tied	GMM + KL	-8.46	-6.00	-14.46
	✓	✓	✓	Mix	✗	Euclidean	Spherical	GMM + KL	-7.98	-8.47	-16.45

(b) Ablation study on the text models Qwen3-Embedding [48] and EmbeddingGemma [43]. Spherical covariance completely collapses the metric for Qwen3-Embedding.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	Angle	Standardize	Sampling	Replace.	Space	Covariance	Algorithm	Qwen3-Embedding	EmbeddingGemma	Total
Covariance	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	**-7.13**	**-30.58**	**-37.71**
	✓	✓	✓	Mix	✗	Euclidean	Full	GMM + KL	-3.21	-28.34	-31.55
	✓	✓	✓	Mix	✗	Euclidean	Tied	GMM + KL	-5.47	-26.68	-32.15
	✓	✓	✓	Mix	✗	Euclidean	Spherical	GMM + KL	0.38	-22.37	-21.99

Across both modalities, the Diagonal covariance structure strictly outperforms all alternative configurations, yielding the lowest MCI.

In the vision domain, utilizing a Full covariance matrix severely degrades predictive performance for DINOv3, almost halving its correlation improvement. This degradation indicates that the Full covariance matrix likely leads to overfitting on the sampled geometric trajectories.

Conversely, in the text domain, over-restricting the model proves equally detrimental. Applying a Spherical covariance matrix, which forces uniform variance across all dimensions, completely collapses RADAR's predictive power on the Qwen3-Embedding model, resulting in a degradation in MCI compared to the baseline. Ultimately, the Diagonal covariance strikes the optimal bias-variance tradeoff, with the Tied configuration emerging as the second-best option. The Diagonal covariance provides sufficient flexibility to capture the independent variances of angular and distance features without the parameter explosion and overfitting of a Full covariance matrix.

C.5 Optimal Transport and Divergence Metrics

Lastly, we ablate the algorithm used to measure the distributional shift between the source and target geometric representations. Table 8 compares our continuous density estimation approach (GMM + KL) against standard discrete optimal transport algorithms (Sinkhorn, Sliced-Wasserstein Distance) and kernel-based metrics (Maximum Mean Discrepancy).

Table 8: Optimal transport algorithm ablations for vision (DomainNet) and text (EuroEval) representations. While standard algorithms like Sinkhorn perform reasonably well on text embeddings, GMM + KL is the only algorithm that demonstrates robust, generalized predictive correlation across both vision and text modalities.

(a) Ablation study on the vision models CLIP [36] and DINOv3 [38]. Standard optimal transport metrics (SWD, Sinkhorn) collapse, while MMD experiences catastrophic failure.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	**Angle**	**Standardize**	**Sampling**	**Replace.**	**Space**	**Covariance**	**Algorithm**	**CLIP**	**DINOv3**	**Total**
Algorithm	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	**-9.72**	**-9.87**	**-19.59**
	✓	✓	✓	Mix	✗	Euclidean	Diag	SWD	3.61	-5.07	-1.46
	✓	✓	✓	Mix	✗	Euclidean	–	MMD (Gauss)	62.32	61.45	123.77
	✓	✓	✓	Mix	✗	Euclidean	–	Sinkhorn	2.81	-4.24	-1.43

(b) Ablation study on the text models Qwen3-Embedding [48] and EmbeddingGemma [43]. While MMD and Sinkhorn show competitive performance for specific architectures, GMM + KL provides the best overall correlation.

	Design choices								Mean Correlation Improvement (%pt.) ↓		
	Distance	**Angle**	**Standardize**	**Sampling**	**Replace.**	**Space**	**Covariance**	**Algorithm**	**Qwen3-Embedding**	**EmbeddingGemma**	**Total**
Algorithm	✓	✓	✓	Mix	✗	Euclidean	Diag	GMM + KL	-7.13	**-30.58**	**-37.71**
	✓	✓	✓	Mix	✗	Euclidean	Diag	SWD	6.07	-19.05	-12.98
	✓	✓	✓	Mix	✗	Euclidean	–	MMD (Gauss)	**-24.04**	-9.04	-33.08
	✓	✓	✓	Mix	✗	Euclidean	–	Sinkhorn	-13.86	-23.50	-37.36

The data reveals a stark difference in how vision and text modalities respond to these algorithms. In the text domain, discrete and kernel-based methods perform remarkably well, with Sinkhorn and MMD achieving a highly competitive total MCI. However, in the vision domain, these same methods collapse. Sinkhorn actively degrades predictive correlation for CLIP, while MMD experiences a catastrophic failure across both vision architectures. This indicates that operating directly on the empirical point clouds of vision trajectories introduces excessive noise, causing point-to-point matching to fail.

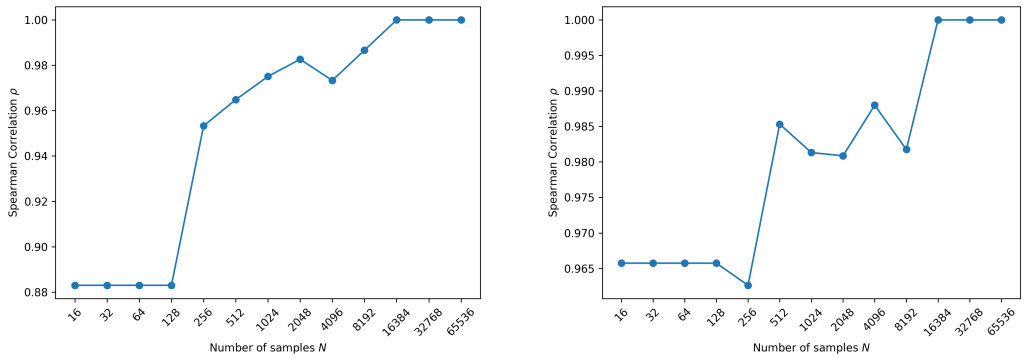
Fitting a GMM and computing the KL divergence emerges as the only universally robust algorithm. By mapping the discrete samples to a continuous probability density function, the GMM smooths out the empirical noise that derails other algorithms in vision models, while simultaneously capturing the precise structural alignments required to achieve optimal predictive correlation in text models. Consequently, *GMM + KL* is strictly required to maintain a unified, modality-agnostic framework.

C.6 Ablations on the robustness of hyperparameters

To demonstrate the robustness of our method to various hyperparameter configurations, we conduct a comprehensive hyperparameter ablation study. Specifically, we analyze the impact of the number of sampled pairs N, the window radius ℓ, and the sampling temperature τ on RADAR's overall performance.

C.6.1 Ablation on the RADAR Sample Size

In this subsection, we evaluate the stability of the RADAR metric across varying sub-sample sizes. As illustrated in Figure 6, we compare the metric's performance at lower sample counts against a high-fidelity baseline. Specifically, we report the Spearman rank correlation ρ between RADAR scores computed with smaller sample sizes N and those computed with a high-fidelity sample count of $N = 65536$.



(a) CLIP (b) DINOv3

Figure 6: **Effect of sample size (N) on RADAR metric stability.** The plots illustrate the Spearman rank correlation (ρ) of RADAR computed at various pair-wise sub-sample sizes ($N \in \{16, 32, 64, \ldots, 65536\}$) relative to a high-fidelity baseline computed with $N = 65536$ on the DomainNet [35] dataset.

We observe that for sample sizes $N \leq 128$, the CLIP model maintains a Spearman correlation exceeding 0.88, which increases to greater than 0.95 at $N = 256$. For the DINOv3 model, RADAR demonstrates even greater sample efficiency, achieving a strong correlation of 0.965 with a drastically reduced sample size of just $N = 16$. For both models, increasing N further leads to a steady performance increase, ultimately plateauing at $N = 16384$ where RADAR achieves a perfect correlation of 1.0 with the high-fidelity baseline of $N = 65536$. For our experiments, we choose $N = 32768$ to ensure evaluations fall safely within this stable regime.

C.6.2 Ablation on the RADAR window size

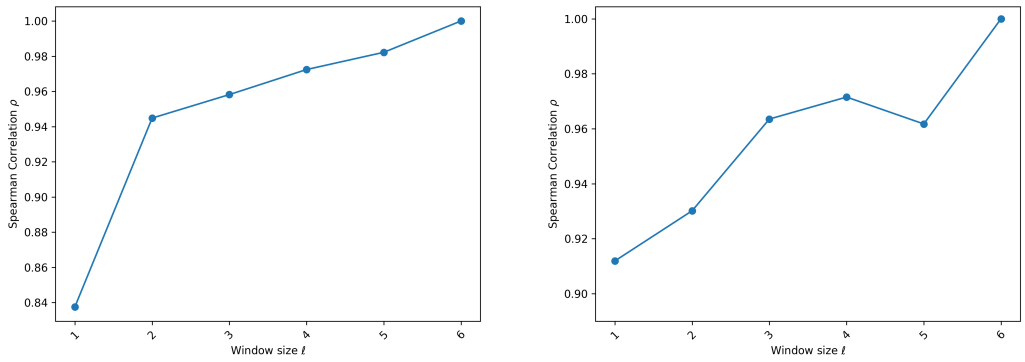
Furthermore, we ablate the window size parameter ℓ to determine its impact on the metric's predictive stability. Figure 7 demonstrates the Spearman rank correlation across different window radii, measured against a baseline RADAR configuration computed at $\ell = 6$.

Similar to the sample size N, we observe a strong correlation for both models even at the minimal window radius of $\ell = 1$. The CLIP model yields a correlation of approximately 0.84, while DINOv3 starts at roughly 0.91. As ℓ increases, the CLIP model experiences a sharp jump to roughly 0.95 at $\ell = 2$. Conversely, DINOv3 sees its steepest ascent at $\ell = 3$, reaching a Spearman correlation of approximately 0.97. Overall, performance for both models rises steadily as ℓ grows, with CLIP achieving a 0.98 correlation at $\ell = 5$. Although DINOv3 experiences a minor drop at $\ell = 5$, the metric remains highly correlated overall, demonstrating that longer, dynamic trajectories provide increasingly stable transferability estimates. Therefore, we set $\ell = 6$ as our default hyperparameter.

C.6.3 Ablation on the RADAR weighting temperature

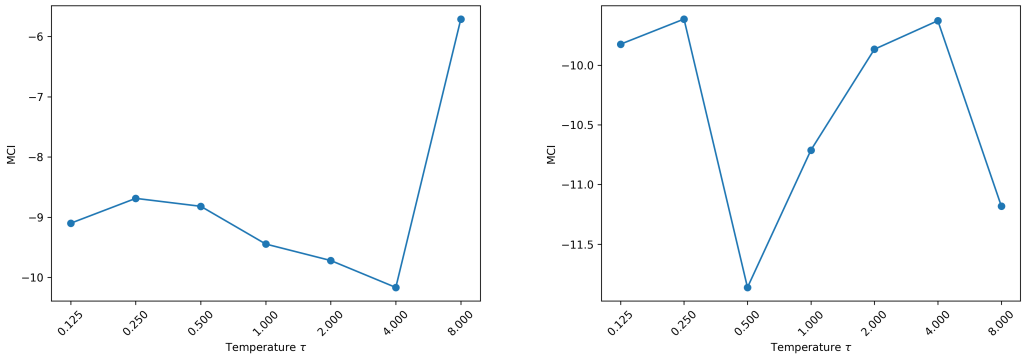
Finally, we ablate the RADAR temperature parameter τ, which controls the smoothness of the inlier-inlier and inlier-outlier sampling distributions, as detailed in Appendix D.1. Figure 8 demonstrates the Mean Correlation Improvement (MCI) across different temperature settings achieved on the DomainNet [35] dataset.

For the CLIP model, the MCI exhibits a slight initial degradation as τ increases from 0.125 to 0.250, followed by a steady improvement as τ ranges from 0.5 to 4.0. However, we observe a catastrophic performance collapse at $\tau = 8$, suggesting that oversmoothing the sampling distribution is highly



(a) CLIP (b) DINOv3

Figure 7: **Effect of the window size ($\ell \in \{1, 2, 3, 4, 5, 6\}$) on RADAR metric stability.** The plots illustrate the Spearman rank correlation (ρ) of RADAR computed at various window sizes relative to a high-fidelity baseline computed with $\ell = 6$ on the DomainNet [35] dataset.



(a) CLIP (b) DINOv3

Figure 8: **Effect of the temperature τ ($\tau \in \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$) on RADAR metric stability.** The plots illustrate the Mean Correlation Improvement (MCI) of the RADAR metric computed across various weighting temperatures on the DomainNet [35] dataset.

detrimental. In contrast, the DINOv3 model remains remarkably robust across the 0.125–8.0 range, with a total MCI variation of only 2.0 percentage points (pp.). Given that CLIP exhibits a much higher sensitivity—with a more than 4.0 pp. spread over the same range—we prioritize its stability and select $\tau = 2.0$ as our default configuration to avoid oversmoothing while maintaining peak performance across both architectures.

D Experiment parameters

D.1 RADAR parameters

As a first step of our metric, we apply layer-wise standardization to the geometric features. Specifically, we sample 2^{20} representational pairs to compute the baseline statistics, which are then used to normalize the angles and relative distances to a zero mean and unit variance. This standardization is critical as it ensures that the angular and distance components are evaluated on the same scale, preventing either feature from disproportionately dominating the Gaussian Mixture Model during density estimation.

Pairwise sampling is conducted using the weighted strategy detailed in Section 3. To reduce the variance of the sampling weights and encourage broader exploration, we apply temperature scaling:

$$w_{reg}(x) = \exp\left(-\frac{w(x)}{\tau}\right), \tag{18}$$

where the temperature hyperparameter τ is set to 2.0. Finally, to guarantee that every pair maintains a non-zero probability of selection, the resulting weights $w_{reg}(x)$ are strictly bounded within the interval $[0.1, 1.0]$.

For the density estimation step in RADAR, we employ a Gaussian Mixture Model (GMM) with strictly controlled parameters to ensure reproducibility. The number of mixture components, K, is dynamically configured as $K = \min(2C, 50)$, where C is the number of unique classes in the target dataset. Allocating twice as many components as classes serves a critical regularization purpose: it provides the GMM with sufficient structural flexibility to model multi-modal distributions within individual classes, effectively capturing both inlier-inlier and inlier-outlier pairs for computing the angles and relative distances. Initialization is performed via the k-means algorithm with a single run and capped at 100 Expectation-Maximization (EM) iterations. To maintain computational stability and enforce positive semi-definite covariance matrices, we apply a diagonal regularization constant of 1×10^{-3}. Finally, the covariance type remains configurable per experiment, and the random state is fixed.

For the remaining RADAR hyperparameters, we sample $N = 32768$ pairs and employ a window radius of $\ell = 6$. This configuration provides sufficient resolution to comprehensively capture the representations' multi-layer geometric trajectories. Importantly, RADAR exhibits strong robustness to these specific choices. A detailed ablation study confirming this stability is provided in Appendix C.6.

D.2 MLP probe parameters

To evaluate the quality of the learned representations, we train a Multi-Layer Perceptron (MLP) probe. The probe architecture consists of an initial linear layer that projects the input representations from the model's hidden dimension to an intermediate size of 512, followed by a ReLU activation function. To mitigate overfitting, we apply dropout regularization with a probability of $p = 0.1$. Finally, a second linear layer projects the intermediate features to the target number of classes.

The MLP probe is trained for a maximum of 100 epochs using a batch size of 256. We optimize the network using the AdamW optimizer with a learning rate of 0.01 and a weight decay of 1×10^{-4}. To mitigate overfitting, we employ early stopping with a patience of 10 epochs. To account for class imbalance during training, we apply a class-weighted Cross-Entropy loss. Specifically, the weight for each class c, denoted as w_c, is computed to be inversely proportional to its sample count N_c. To maintain the overall magnitude of the loss, these weights are normalized such that their sum equals the total number of classes C. Formally, the weight for class c is defined as:

$$w_c = C \frac{N_c^{-1}}{\sum_{j=1}^{C} N_j^{-1}}. \tag{19}$$

To ensure the robustness of our evaluation and account for initialization variance, we train the MLP probes over $r = 10$ distinct random seeds for each source-target pair. Across all runs, we observe rapid convergence, with early stopping consistently triggered before epoch 40.

E Datasets Used

To evaluate the cross-modal applicability of RADAR, we utilize diverse datasets across both vision (three benchmarks) and text (two benchmarks). While EuroEval [34] and PACS [25] are used in their entirety, we strategically subsample OfficeHome [42], DomainNet [35], and Amazon Reviews [31] to ensure tractable evaluation. Specifically, we restrict the size of DomainNet and OfficeHome by adopting sampling protocols standard in the domain adaptation literature [47]. Furthermore, we uniformly downsample the Amazon Reviews dataset from its original 256,000 samples per language to create a balanced, computationally efficient subset. Finally, to explicitly assess the metric's resilience to natural corruptions, we incorporate a subset of the ImageNet-C [10] dataset evaluated at varying severity levels (1, 3, and 5).

E.1 Image classification datasets

DomainNet [35] Dataset Details:

- **Dataset Splits:** Train: 15,874 | Validation: 3,175 | Test: 6,841
- **Domain Distribution (Train):** Clipart (1,889), Infograph (2,038), Painting (2,375), Quickdraw (3,500), Real (4,146), Sketch (1,926)
- **Class Distribution (Train):** Strawberry (1,524), Circle (1,172), Bread (1,609), Feather (1,620), Teapot (1,575), Whale (1,756), Windmill (1,649), Zebra (1,584), Tiger (1,732), Bird (1,653)

PACS [25] Dataset Details:

- **Dataset Splits:** Train: 8,977 | Validation: 1,014 | Test: 9,991
- **Domain Distribution (Train):** Art Painting (1,840), Cartoon (2,107), Photo (1,499), Sketch (3,531)
- **Class Distribution (Train):** Dog (1,555), Elephant (1,486), Giraffe (1,407), Guitar (1,000), Horse (1,384), House (846), Person (1,299)

OfficeHome [42] Dataset Details:

- **Dataset Splits:** Train: 6,031 | Validation: 606 | Test: 604
- **Domain Distribution (Train):** Art (1,034), Clipart (1,709), Product (1,720), Real World (1,568)
- **Class Distribution (Train):** Alarm Clock (299), Backpack (295), Bike (317), Bottle (338), Candles (330), Chair (363), Computer (303), Couch (268), Flipflops (270), Flowers (355), Folder (266), Helmet (298), Keyboard (291), Laptop (316), Monitor (310), Scissors (295), Sneakers (276), Speaker (290), TV (268), Telephone (283)

E.2 Sentiment analysis datasets

EuroEval [34] Dataset Details:

- **Dataset Splits:** Train: 8,029 | Validation: 2,048 | Test: 13,312
- **Domain Distribution (Train):** Czech (1,024), English (1,024), German (1,024), Hungarian (1,024), Italian (1,024), Polish (1,024), Slovak (1,024), Spanish (861)
- **Class Distribution (Train):** Negative (2,853), Neutral (2,360), Positive (2,816)

Amazon Reviews [31] Dataset Details:

- **Dataset Splits:** Train: 196,596 | Validation: 12,276 | Test: 12,276
- **Domain Distribution (Train):** Chinese (32,766), English (32,766), French (32,766), German (32,766), Japanese (32,766), Spanish (32,766)
- **Class Distribution (Train):** Negative (65,532), Neutral (65,532), Positive (65,532)

E.3 Synthetic datasets

Imagenet-C [10]

- **Dataset Splits:** Train: 16,140 | Validation: 3,228 | Test: 10,818
- **Domain Distribution (Train):** Brightness (2,699), Contrast (2,696), Fog (2,691), Motion Blur (2,687), Real (2,684), Snow (2,683)
- **Class Distribution (Train):** rock beauty (872), Mexican hairless (679), digital watch (800), Japanese spaniel (694), toy terrier (774), black-and-tan coonhound (658), English foxhound (678), space heater (903), otterhound (664), wire-haired fox terrier (879), Sealyham terrier (842), Petri dish (865), sweatshirt (837), hair spray (882), hatchet (801), throne (878), Irish water spaniel (872), power drill (817), letter opener (887), affenpinscher (858)

F Baseline Implementations

We implement the majority of our baseline methods by directly adapting the codebase from the Transfer Learning Library [21, 22]. For the s-OTDD baseline, we utilize the official implementation provided by the original authors, with minor modifications to interface properly with our standardized experimental framework.

Additionally, for the computation of the Sinkhorn and MMD (Maximum Mean Discrepancy) distances, we leverage the highly optimized `geomloss` library [6].

G Seed stability of RADAR

We report mean and standard deviation of RADAR's Spearman rank correlation across $r = 10$ random seeds for both backbone architectures on DomainNet [35] and EuroEval [34] in Table 9. This directly quantifies the variance introduced by RADAR's stochastic pair sampling and confirms the stability of the reported results.

Table 9: Mean, standard deviation, minimum and maximum of RADAR's Spearman rank correlation across $r = 10$ subsampling seeds on DomainNet [35] and EuroEval [34].

Dataset	Backbone	Mean ρ (%pt., $\downarrow$)		Std Dev	Min	Max
DomainNet	CLIP	-51.75	$\pm$	0.94	-53.25	-50.57
	DINOv3	-44.60	$\pm$	0.91	-46.09	-43.61
EuroEval	Qwen3-Embedding	-8.37	$\pm$	0.64	-9.62	-7.52
	EmbeddingGemma	-34.92	$\pm$	1.55	-36.80	-32.29

On DomainNet, RADAR exhibits strong stability with respect to the randomness of empirical pair sampling and GMM fitting. The standard deviations for the Spearman correlation (ρ) across all $r = 10$ independent seeds are consistently low for both vision architectures: ± 0.94 percentage points for CLIP and ± 0.91 for DINOv3, with narrow empirical spreads of $[-53.25, -50.57]$ and $[-46.09, -43.61]$, respectively.

RADAR exhibits broadly stable behaviour across both text architectures on EuroEval, though with some variation between models. The standard deviations are comparable to those observed in the vision domain for Qwen3-Embedding (± 0.64 percentage points), while EmbeddingGemma exhibits notably higher variance (± 1.55 percentage points), reflected in its wider empirical spread of $[-36.88, -32.29]$ compared to $[-9.62, -7.52]$ for Qwen3-Embedding.

These results suggest that RADAR's geometric trajectory descriptor consistently produces stable divergence estimates across independent subsampling runs, and that the stratified pair sampling strategy introduced in Section 3 provides sufficient coverage of the domain manifold without exhaustive enumeration.

Note: The mean ρ reported in Table 9 differs slightly from that reported in Table 2, as the two tables are computed over independent sets of subsampling seeds. Given the sampling variance of each model, this discrepancy is expected and reflects seed-to-seed variability rather than an inconsistency in the experimental setup. Nevertheless, the difference remains within 1 standard deviation reported in Table 9.

H Broader Impacts

Our work aims to yield positive societal impacts by enabling more data-efficient training paradigms. By strategically filtering training data, our method significantly reduces the computational overhead — and consequently, the environmental footprint, such as electricity and cooling water consumption — associated with training large-scale models. However, we acknowledge the dual-use nature of our research. As RADAR can be used to predict negative transfer, malicious actors could theoretically invert our metric's objective to intentionally degrade model performance, induce catastrophic forgetting, or systematically amplify harmful biases within foundation models. Mitigating this risk would require access controls on the metric's deployment and awareness of adversarial inversion when applying transferability metrics in safety-critical settings.

I Computational cost analysis

In Table 10, we report the wall-clock time required to compute each metric across all domain pairs in the DomainNet [35] dataset. Because RADAR relies on GMMs—which currently lack highly optimized GPU implementations—and several baseline metrics require a model trained on source data, we evaluate all methods under both CPU-only and GPU-accelerated configurations. For metrics necessitating a source model, we include the training time of an auxiliary MLP classifier head to ensure a fair comparison. Baseline metrics are evaluated using the standardized implementations from the Transfer Learning Library [21, 22], with the exception of s-OTDD [33], for which we utilize the official implementation. Due to implementation-specific compatibility constraints, s-OTDD [33] results are reported for CPU execution only. Regarding our own method, we omit the CPU timing for the Sinkhorn-based variant as its primary utility lies in GPU acceleration; its GPU performance is included to demonstrate the latent optimization potential of our metric.

Table 10: Computational cost analysis comparing RADAR to baseline metrics. Time represents the wall-clock time in seconds (↓) to evaluate all the domain pairs over the whole network, averaged over $r = 10$ runs. Evaluated on an AMD EPYC 9124 CPU and a single NVIDIA RTX 6000 Ada Generation GPU.

Hardware	LEEP	LogME	NCE	TransRate	s-OTDD	RADAR (Ours)	RADAR (Sinkhorn)
CPU Only	894.62	2803.74	916.64	1193.73	>3600	**644.56**	–
GPU	208.01	1411.28	**204.05**	378.09	N/A	454.20	253.53

Remarkably, RADAR is the most efficient metric in CPU-only environments, outperforming all baselines including LEEP [32], NCE [41] and s-OTDD [33]. While our approach involves training a GMM, strategic subsampling of the data allows us to bypass the longer training times associated with the classifier heads required by many traditional benchmarks. Although the layer-wise trajectory analysis of our vanilla method introduces a slight overhead on the GPU, the RADAR (Sinkhorn) variant significantly mitigates this, achieving runtimes comparable to the fastest statistical benchmarks. These results demonstrate that RADAR offers a superior balance of predictive power and practical scalability, with significant room for further hardware-specific acceleration.

J Data Processing Inequality and Multi-Layer Advantage

A natural question is whether the multi-layer trajectory descriptor used by RADAR provides any advantage over a single-layer divergence computed at the network's final representation. The data processing inequality (DPI) for total variation gives a precise answer: divergences present at intermediate layers can only be *suppressed*, never amplified, by subsequent deterministic layer maps. This justifies extracting features from intermediate transformer blocks rather than relying on the final-layer representation alone.

Assumption J.1 (Frozen constant-width transformer)**.** Let $\Phi : \mathcal{X} \to \mathbb{R}^H$ be a frozen, pre-trained foundation model with L residual blocks of constant ambient dimension H. The layer-l feature map $h^{(l)} : \mathcal{X} \to \mathbb{R}^H$ satisfies $h^{(0)}(x) = E(x)$ and $h^{(l)}(x) = T_{l-1}(h^{(l-1)}(x))$, where each $T_l : \mathbb{R}^H \to \mathbb{R}^H$ is deterministic and measurable, and features are uniformly bounded: $\|h^{(l)}(x)\| \le B$ for all l, x. The pushforward feature distribution at layer l for domain D is $\mu_D^{(l)} := (h^{(l)})_{\#} P_D^X$.

Proposition J.2 (Layer-wise total variation is non-increasing)**.** *Under Assumption J.1, for any two domains A, B and any pair of layer indices $l \le l'$ within the network,*

$$d_{\mathrm{TV}}\left(\mu_A^{(l)}, \mu_B^{(l)}\right) \;\ge\; d_{\mathrm{TV}}\left(\mu_A^{(l')}, \mu_B^{(l')}\right). \tag{20}$$

Proof. By Assumption J.1, the composition

$$T_{l:l'} \;:=\; T_{l'-1} \circ T_{l'-2} \circ \cdots \circ T_l \;:\; \mathbb{R}^H \to \mathbb{R}^H \tag{21}$$

is a deterministic measurable map satisfying $\mu_D^{(l')} = (T_{l:l'})_{\#}\, \mu_D^{(l)}$ for $D \in \{A, B\}$. The data processing inequality for total variation states that for any measurable map f and any pair of probability measures μ, ν on a common space,

$$d_{\mathrm{TV}}(f_{\#}\mu,\, f_{\#}\nu) \;\le\; d_{\mathrm{TV}}(\mu, \nu). \tag{22}$$

Applying this with $f = T_{l:l'}$, $\mu = \mu_A^{(l)}$, and $\nu = \mu_B^{(l)}$ yields the claim. □

Corollary J.3 (Final-layer representations underestimate domain divergence)**.** *For any intermediate layer $l < L$,*

$$d_{\mathrm{TV}}\left(\mu_A^{(l)}, \mu_B^{(l)}\right) \;\geq\; d_{\mathrm{TV}}\left(\mu_A^{(L)}, \mu_B^{(L)}\right). \tag{23}$$

Consequently, any transferability metric that operates exclusively on single-layer features (e.g., centroid distance, LogME, TransRate computed at the network output) systematically underestimates the domain divergence that may be visible at any earlier layer.

Remark J.4 (Implications for transferability estimation). Proposition J.2 formalizes an observation made empirically by recent work on foundation-model representations: the most discriminative features for downstream tasks are not always located at the network's output [3]. Final-layer normalization and the inductive bias toward task-aligned features can suppress domain-specific structure that remains observable at intermediate depths.

This has two consequences for the design of transferability metrics:

1. **Single-layer baselines are biased.** Metrics that compare only final-layer feature distributions (such as l_2 centroid distance or s-OTDD) inherit a one-sided bias: they may report small divergence even when the underlying domains are meaningfully distinct at earlier layers.
2. **Multi-layer extraction is justified.** Multi-layer extraction captures the full divergence profile across depth, recovering divergence signals that are monotonically suppressed as depth increases and would be invisible to any single-layer metric.

The trajectory description leveraged by RADAR utilizes this insight by integrating geometric information across a window of radius $\ell = 6$ centered at l (from $l - 6 \to \cdots \to l \to \cdots \to l + 6$). The empirical justification for the specific window size $\ell = 6$ is provided by the ablation in Appendix C.6.2 and Figure 7, which demonstrates that Spearman correlation increases monotonically with ℓ across both vision architectures.

Remark J.5 (Scope of the inequality). Proposition J.2 bounds total variation, not the full RADAR KL divergence, and concerns marginal feature distributions at single layers, not the joint trajectory law. It establishes that single-layer final-only metrics are biased, but does not by itself establish that the specific $(\theta^{(l)}, d^{(l)})$ trajectory descriptor is optimal among multi-layer alternatives. The empirical case for the geometric descriptor over alternative multi-layer summaries is made through the ablations in Section C.

K Proof of the Displacement Triangle

Proposition K.1 (Triangle closure)**.** *For all $x, x' \in \mathcal{X}$ and all $l \in \{0, 1, \ldots, L-1\}$,*

$$v_{\mathrm{sep}}^{(l)}(x, x') \;+\; v_{\mathrm{detour}}^{(l)}(x, x') \;=\; v_{\mathrm{traj}}^{(l)}(x). \tag{24}$$

Proof. Direct expansion of Definitions (6)–(8):

$$\begin{aligned} v_{\mathrm{sep}}^{(l)}(x, x') + v_{\mathrm{detour}}^{(l)}(x, x') &= \left[h^{(l)}(x') - h^{(l)}(x)\right] + \left[h^{(l+1)}(x) - h^{(l)}(x')\right] && (25)\\ &= h^{(l+1)}(x) - h^{(l)}(x) && (26)\\ &= v_{\mathrm{traj}}^{(l)}(x). && (27) \end{aligned}$$

The intermediate term $h^{(l)}(x')$ cancels, leaving the direct displacement from layer l to layer $l + 1$ at anchor x. □

L Base models

We describe the exact base model configurations used for feature extraction and our main results.

For image classification:

- **DINOv3 [38]**: facebook/dinov3-vits16-pretrain-lvd1689m [16]
- **CLIP [36]**: openai/clip-vit-base-patch32 [15]

For sentiment analysis:

- **Qwen3-Embedding [48]**: Qwen/Qwen3-Embedding-0.6B [17]
- **EmbeddingGemma [43]**: google/embeddinggemma-300m [14]

M Declaration of LLM usage

We declare that LLMs were used for spell-checking, editing, formatting, and paraphrasing. The proofs in Appendix J were additionally proposed and reviewed with LLM assistance and independently verified by the authors. LLMs were also used for programming aid, providing initial, skeleton implementations that were subsequently refined by the authors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide an empirical justification of our claims in Section 4 on both cross-lingual sentiment classification and cross-domain image classification, backed with ablations and method description. Our contributions are clearly delineate in the introduction and reflected in the paper's abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our limitations are presented in Section 6. We discuss the current three limitations (RADAR's degradation, unified formulation vs. fine-grained adaptation, GMM GPU support) and provide a case study in the Appendix B of our degradation on datasets that lack clear, uniform geometric separation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendices J and K provide formal proofs of the multi-layer advantage and displacement triangle closure, respectively, building on the provided Assumption J.1 of a frozen, constant-width transformer, with accompanying remarks and corollaries.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full dataset details, model descriptions, and evaluation parameters, alongside a complete specification of RADAR's hyperparameter configuration to ensure reproducibility. Full details are provided in Appendices C, E and D and Section 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We intend to release the code upon acceptance, including the full preprocessing pipeline.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The RADAR hyperparameters are well documented, and dataset splits are provided in Appendix E. We provide the description of our base models in the Appendix L. We provide the experimental setting under Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the sampling stability of our metric in Table 9 in Appendix G, summarizing the mean, standard deviation, and empirical range of RADAR's Spearman rank correlation across $r = 10$ independent subsampling seeds. Reporting equivalent variance

statistics for all baselines across all backbone architectures and domain pairs would be computationally prohibitive.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the exact machine specification used in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our study does not involve human subjects or violate privacy, consent, or copyright concerns. All datasets used are publicly available, and all the external used packages are readily available without any licensing issue. Broader societal impacts and safety considerations are discussed in Appendix H.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader societal impacts of malicious negative transfer and safety considerations are discussed in Appendix H.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: RADAR is a transferability metric that operates on frozen features from existing foundation models — it does not release new pretrained models, scraped datasets, or generative systems that would warrant safeguards. The broader impacts section acknowledges the dual-use risk are described in Appendix H.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, the four backbone models, and the Transfer Learning Library and the geomloss library are credited. All work is properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our declaration of LLM usage is provided in Appendix M. We acknowledge that LLMs were used solely as assistive tools and do not constitute any novel or non-standard component of this work; however — we provide the declaration for transparency.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.