GeoTTER: Leveraging Local Geometry of Optimal Transport for Zero-Shot Classification

Wei-Yang Alex Lee
SUNY Buffalo

Rudrasis Chakraborty
Lawrence Livermore National Lab

Vishnu Lokhande
SUNY Buffalo

Abstract

We present GeoTTER, a novel framework that redefines optimal transport in the realm of zero-shot classification. Conventional methods often suffer from miscalibration and a lack of adaptability, as they rely on fixed cost matrices derived solely from pre-trained model embeddings. In contrast, GeoTTER addresses these limitations by incorporating two key techniques. First, to alleviate high-frequency label jaggedness (sample-level manifold jitter that assigns neighboring embeddings to different classes), GeoTTER integrates local geometric structure into the optimal transport formulation via graph-Laplacian smoothing, a technique grounded in spectral graph theory that enforces neighborhood consistency. Second, to correct coherent angular drift (a low-frequency orientation bias in which large groups of samples share the same angular offset from their true label prototypes), we fuse clustering-guided cost components with a globally adjusted transport cost, achieving a multi-objective optimization that respects both global distribution constraints and latent data structure. With a median improvement of +6.82% compared to zero-shot and +2.13% compared to OTTER, GeoTTER shows robust improvements across a diverse set of benchmarks.

[1]

1 INTRODUCTION

Optimal transport (OT) has emerged as a powerful and flexible tool for aligning distributions across modalities, showing promise in a wide range of tasks from domain adaptation to zero-shot learning Cuturi (2013b). However, despite its theoretical elegance, standard OT formulations often suffer from poor calibration, numerical instability, and limited capacity to model local structural nuances within the embedding space. These challenges become particularly pronounced in modern zero-shot and prompt-based classification settings, where the misalignment between label priors and prediction logits can significantly degrade performance. Liusie et al. (2023)

Recent works have attempted to address some of these limitations. For instance, OTTER Shin et al. (2024) proposes a framework to adapt label distributions in zero-shot models without requiring retraining, highlighting the importance of flexible cost matrix designs. Liusie et al. Liusie et al. (2023) explore the problem of word bias in prompt-based zero-shot classifiers, advocating for more robust methods that incorporate semantic priors. Bai et al. Bai et al. (2023) tackle the label shift problem in online settings, providing provable adaptation mechanisms. These lines of work all point to a common need: OT methods must become more structure-sensitive to generalize effectively especially when label distributions are different.

In this paper, we propose **GeoTTER**, a novel OT-based framework that addresses these challenges by systematically redesigning the OT cost matrix. Our formulation integrates prior label distribution and local geometric structure to enhance calibration. We introduce two core ideas: (1) a **graph smoothing** operation inspired by spectral graph theory to enforce local consistency, and (2) a **clustering-guided cost fusion** technique for leveraging latent structure in the embedding space. Together, these components form the backbone of GeoTTER.

Assumptions & Scope. The setup still requires access to the target-domain class prior ν at test time (the paper notes that ν can be estimated from a small labeled subset or via EM). That said, this limitation is milder than it seems: the theory frames prior misspecification as a bounded marginal-target estimation error

[1]Paper accepted at Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco.

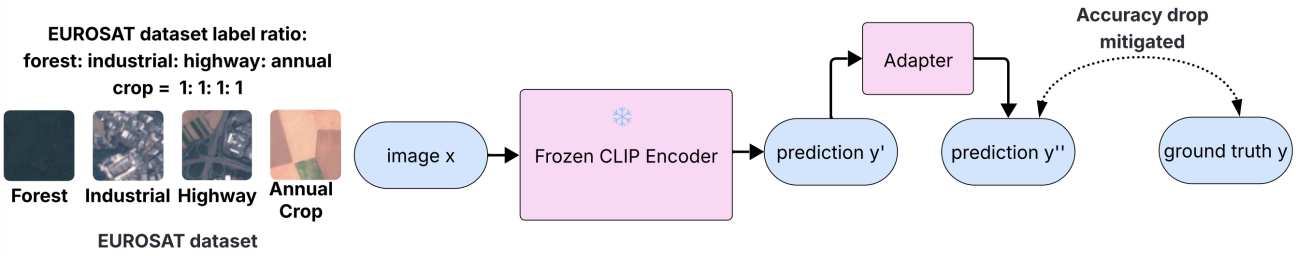


Figure 1: Test-time adjustment (TTA) of a Frozen CLIP Encoder nder Label-Distribution Shift. The target domain is the four-class EUROSAT subset: forest, industrial, highway, annual-crop, whose classes occur in an approximately 1: 1: 1: 1 ratio, unlike the highly skewed class prior in CLIP's Internet-scale pre-training data (estimated: forest $\gg$ industrial $\approx$ highway $\gg$ annual-crop). An unseen satellite image x is first embedded by the frozen CLIP image encoder, yielding a zero-shot prediction y' that is biased toward the pre-training prior and therefore suffers an accuracy drop. A lightweight adapter, learned on the unlabeled target images at test time, transforms y' into an adjusted prediction y'', effectively compensating for the prior mismatch and recovering performance, without updating CLIP's weights or requiring any test-set labels.

ε_ν, and once the estimated prior $\hat{\nu}$ achieves even modest, theoretically-guaranteed accuracy, the excess risk shrinks. Empirically, OTTER already outperforms the zero-shot (no adapter) baseline with such coarse estimation.

On top of OTTER, GeoTTER further mitigates this dependency through clustering-guided cost fusion and graph smoothing; *utilizing local geometry* helps stabilize the transport plan even under imperfect prior knowledge. This local structure encourages consistency within similar samples, reducing sensitivity to $\hat{\nu}$'s exactness and making the estimation burden lighter in practice. Our key contributions are as follows:

- We introduce a graph-based smoothing technique that improves zero-shot classification accuracy by enforcing local consistency in transport assignments.
- We develop a clustering-aware extension that fuses semantic and geometric signals for improved transport planning.
- We conduct extensive experiments across 17 vision benchmarks, demonstrating that GeoTTER consistently outperforms classical OT and recent zero-shot adaptation methods Shin et al. (2024); Liusie et al. (2023).

The remainder of this paper is organized as follows. Section 4 introduces the theoretical framework of GeoTTER. Section 5 presents empirical results, and Section 6 concludes with discussion and future work.

2 RELATED WORKS

Optimal Transport and Cost Matrix Reformulation Optimal transport (OT) redistributes probability mass, but use revealed calibration drift and instability when costs span orders of magnitude. Rescaling logits with temperatures or per-class weights and embedding in the cost matrix compresses extremes and tempers over-confident predictions Lavenant et al. (2024); entropy-regularized Sinkhorn stabilizes optimization with bias–variance trade-off Lavenant et al. (2024). Convex re-parameterizations yield closed-form class-specific offsets, easing tuning and preserving stability in high dimensions Savaré and Sodini (2023).

Graph-Based Methods and Local Manifold Regularization Graph methods enforce neighborhood consistency by minimizing a Laplacian quadratic on k-NN graphs; diffusion maps and harmonic functions extend this idea to semi-supervised learning. Advances in adaptive neighborhoods Zhang et al. (2020), kernel/multi-view fusion Chen et al. (2022); Wang et al. (2021b), and large-scale solvers collectively align probability transport with intrinsic data geometry and improve robustness.

Clustering-Guided Cost Fusion for Enhanced Structure Capture Beyond OT objectives, incorporating *clustering cues* into cost improves calibration when labels correlate with sub-manifold structure. Pipelines discover clusters via k-means, GMMs, or spectral variants and then fuse a cluster-level similarity term with the standard OT cost. This multi-objective view lets the OT term align global class priors while a cluster-aware term captures fine-grained geometry, yielding robust performance on imbalanced or intricately structured data.

3 PRELIMINARIES

We start by formalizing the Test–Time Adaptation (TTA) problem tackled in this work and by reviewing two recent baselines that recalibrate model predictions without touching the backbone parameters.

3.1 Setting and Notation

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be an *unlabeled* test batch and let f be a frozen classifier which outputs softmax scores $\{\mathbf{p}_i\} = \{\text{softmax}\big(f(x_i)\big)\} \subset \Delta^{m-1}$. The backbone model was trained on a source domain whose class prior may differ from the *target-domain prior* $\boldsymbol{\nu} \in \Delta^{m-1}$, assumed to be known[2]. Our goal is to transform the scores $\{\mathbf{p}_i\}$ into calibrated probabilities $\{\tilde{\mathbf{p}}_i\}$ such that their mean $\boldsymbol{\nu}$ matches and the resulting hard labels are more accurate. For convenience, we write the uniform mass vector $\boldsymbol{\mu} = \frac{1}{n}\mathbf{1}_n$ and use $\langle A, B\rangle := \sum_{ij} A_{ij}B_{ij}$ for the inner products of the matrix. Below, we will briefly mention prior-matching and OTTER, two recent work tackling with recalibration of the model prediction in a zero-shot setting.

3.2 Prior-Matching Adapter (PM)

PM Liusie et al. (2023) rescales logits with a per-class weight vector $\mathbf{w} \in \mathbb{R}^m_{>0}$ and a temperature $T > 0$: $\tilde{\mathbf{p}}_i = \text{softmax}\Big(\frac{1}{T}\left(\mathbf{w} \odot \mathbf{p}_i\right)\Big)$, where "$\odot$" denotes element-wise multiplication. The weights are obtained by optimizing the following $\min_{\mathbf{w}} \left\| \frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{p}}_i - \boldsymbol{\nu} \right\|_2^2$. As there are only m parameters (from the weight vector) that need to be tuned, a small number of first-order steps suffice. This method treats each sample independently and is sensitive to the choice of T and hyper-parameters of the optimizer.

Fenchel PM (FPM) Building upon Prior Matching, we additionally evaluate Fenchel PM (FPM), which alleviates PM's gradient-oscillation and local-minima trapping issues; FPM yields markedly better performance than PM, though it still falls short of OTTER (see Appendix for details). FPM introduce an auxiliary probability vector $\mathbf{p} \in \Delta_m$ and, via the Fenchel–Bregman identity, absorb the softmax into an inner maximization to obtain the saddle-point problem $\min_{\mathbf{w}\in\mathbb{R}^m_{>0}} \max_{\mathbf{p}\in\Delta_m} \left\langle \frac{1}{T}(\mathbf{w}\odot\boldsymbol{\theta}), \mathbf{p}\right\rangle - \Omega(\mathbf{p}) + \frac{\lambda}{2}\|\mathbf{p}-\boldsymbol{\nu}\|_2^2$, where $\Omega(\mathbf{p}) = \sum_{j=1}^m p_j \log p_j$. The inner maximizer $\mathbf{p}^\star(\mathbf{w})$ satisfies the closed-form fixed point $\mathbf{p}^\star = \text{softmax}\left(\frac{1}{T}(\mathbf{w}\odot\boldsymbol{\theta}) - \lambda(\mathbf{p}^\star - \boldsymbol{\nu})\right)$; eliminating $\mathbf{p}$ yields a smooth, strongly convex outer objective in $\mathbf{w}$, hence more stable optimization.

3.3 Optimal TransporT adaptER (OTTER)

OTTER Shin et al. (2024) interprets recalibration as a *discrete optimal-transport (OT)* problem and reallocates the entire batch's probability mass in zeroshot. Thus it captures the relationship among samples in a batch. Below, we will explain the OT objective.

Transport Problem Given source and target marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ respectively, we solve the *exact* Kantorovich optimal-transport problem Kantorovich (1942) $\pi^\star = \arg\min_{\pi\in\mathbb{R}_+^{n\times m}} \langle \pi, C\rangle; \text{s.t.}; \pi, \mathbf{1}_m = \boldsymbol{\mu}, ; \pi^\top \mathbf{1}_n = \boldsymbol{\nu}$. i.e. the Earth-Mover's Distance (EMD) between the two empirical measures Rubner et al. (2000). For every sample–class pair we set $C_{ij} = -\log p_i[j]$. We compute $\pi^\star$ with the *network-simplex* algorithm Orlin (1997), an efficient specialized linear-programming solver whose practical runtime is near $O((n+m)^3)$ but sub-cubic on the small batches typical of OTTER. The calibrated label for each sample is then obtained as $\hat{y}_i = \arg\max_j \pi^\star_{ij}$. OTTER leverages the interaction among *all* samples, dispenses with hand-tuned hyper-parameters, and unlike neural-network-based adapters, OTTER returns the exact global optimum of a well-posed linear program.

Exact OT / Earth-Mover's Distance (EMD) Set the entropy weight to zero, i.e. $\varepsilon = 0$, to recover the original Kantorovich linear program $\min_{T\geq 0} \langle T, C\rangle$ s.t. $T\mathbf{1} = \mathbf{a}$, $T^\top\mathbf{1} = \mathbf{b}$. Without the entropy term the solution T is *exact* but typically sparse; standard solvers such as network–simplex or interior-point are required, yielding super-cubic time in the worst case.

Sinkhorn-Regularized OTTER (SK) To gauge the impact of entropic regularization, we also evaluate a simple Sinkhorn-based version of OTTER (denoted **SK**), obtained by replacing the network-simplex solver with Sinkhorn iterations. Add an entropy term Sinkhorn and Knopp (1967) $\varepsilon\,\text{KL}(T\|\mathbf{a}\mathbf{b}^\top) = \varepsilon\sum_{ij} T_{ij}(\log T_{ij} - 1)$, yielding $\min_{T\geq 0}\langle T, C\rangle + \varepsilon\sum_{ij} T_{ij}(\log T_{ij} - 1)$. The Sinkhorn algorithm scales rows and columns alternately, giving near-linear runtime and a smoother T Cuturi (2013a); letting $\varepsilon \to 0$ recovers EMD.

4 THEORETICAL FRAMEWORK FOR GEOTTER

Regularizing the transport plan is indispensable because it explicitly constrains the solution space, removing spurious high-frequency modes and yielding a markedly smoother assignment landscape; indeed, replacing vanilla OTTER with its regularized counterpart already delivers the consistent gains reported in Table 1. Our preliminary Sinkhorn baseline corroborates this observation: *the mere addition of an entropic term systematically outperforms its unregulated analogue, suggesting that further, task-aware forms of regularization may unlock additional robustness.*

[2]The prior can be estimated from a small labeled subset or via expectation–maximization; empirical details are given in sec:experiments.

Closer inspection of the error patterns reveals two complementary noise sources: a sample-level jaggedness **(lNoise)** that perturbs neighboring embeddings independently, and a batch-level drift **(gNoise)** that shifts logits coherently across larger manifolds. Addressing these dual corruptions requires two orthogonal mechanisms: (i) **Graph Smoothing** (Sec. 4.1), which averages the transport mass over the k-NN graph and therefore suppresses lNoise while respecting local geometry, and (ii) **Clustering-guided Cost Fusion** (Sec. 4.2), which re-anchors the cost matrix to cluster centroids, neutralizing the low-frequency bias induced by gNoise. Together, these strategies generalize the benefits hinted at by Sinkhorn, providing a principled, dual-scale regularization framework that underlies the full GeoTTER model. Our framework is derived from non-trivial theoretical insights in optimal transport Peyré and Cuturi (2019), convex optimization Boyd and Vandenberghe (2004a), and spectral graph theory Chung (1997).

4.1 Graph-Laplacian Smoothing of the Transport Plan

Noise Regime The adjusted transport plan $T^\star$ is still vulnerable to *high-frequency label jaggedness*: a sample-level stochastic logit competition in which infinitesimal perturbations, for example activation noise or batch-normalization drift, cause neighboring embeddings to exchange their top-ranked class. The resulting saw-tooth assignment pattern Shi et al. (2023); Liang et al. (2022) violates manifold smoothness and degrades local calibration. To suppress this noise we enforce neighborhood consistency through graph-Laplacian smoothing.

Motivation Although the adjusted cost matrix produces a globally calibrated transport plan $T^\star$, local noise in the embedding space can still yield inconsistent assignments. To exploit the manifold structure and suppress this noise, we project $T^\star$ onto the k-nearest-neighbor (k-NN) graph via a simple neighborhood averaging iteration.

Single-Step Propagation Below we give the explicit single-step update and introduce our notation. Construct the k-NN graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ on embeddings $\{x_i\}_{i=1}^n$ with adjacency

$$A_{ij} = \mathbf{1}\big[j \in \mathcal{N}(i)\big], \quad D = \mathrm{diag}(A\,\mathbf{1}), \quad \hat{A} = D^{-1}A.$$

Starting from $T^{(0)} = T^\star$, we iterate

$$T^{(\ell+1)} = \alpha\, T^{(\ell)} + (1-\alpha)\,\hat{A}\,T^{(\ell)}, \quad 0 < \alpha \le 1, \quad (1)$$

where α balances retention of the original plan against smoothness over the k-NN mean ($\alpha \uparrow \iff$ less smoothing). After L iterations we set $T^{\mathrm{smooth}} = T^{(L)}$ and predict $\hat{y}_i = \arg\max_j T^{\mathrm{smooth}}_{ij}$. Next, we show that this simple iteration is equivalent to a graph Laplacian regularization problem.

(Multi-GeoTTER) Multi-Step Propagation We also provide the result of multi-step propagation, and named it as **Multi-GeoTTER**, as a GeoTTER's variant. Prop repeatedly applies the single-step update in Eq. (1) for L iterations. This yields progressively broader neighborhood smoothing while keeping the same α. We use the final iterate $T^{(L)}$ for prediction.

Graph-Laplacian Regularization From an optimization perspective, repeating (1) can be seen as solving a Tikhonov-regularized least-squares problem on the graph.

Theorem 1 (Tikhonov form) *Let $\mathcal{L} = I - \hat{A}$ be the row-normalized graph Laplacian and define $\mu = \frac{1-\alpha}{\alpha}$. Then the fixed point of* (1) *minimizes* $\min_{T\in\mathbb{R}^{n\times m}} \frac{1}{2}\sum_{i,j}\big(T_{ij} - T^\star_{ij}\big)^2 + \mu\,\mathrm{tr}\big(T^\top \mathcal{L}\,T\big)$. *Hence, graph smoothing enforces manifold smoothness by penalizing rapid variations of T over $\mathcal{G}$. Thus, repeatedly applying* (1) *provides a simple and numerically stable approximation to this regularized objective.*

This equivalence demonstrates that the iterative update in Eq. (1) implicitly enforces a manifold smoothness prior, with α (or equivalently $\mu = \frac{1-\alpha}{\alpha}$) governing the trade-off between fidelity to $T^\star$ and local consistency. This completes the theoretical underpinning of our graph smoothing step.

4.2 Clustering-Guided Cost Fusion

Noise Regime Beyond local jaggedness, test batches often exhibit *low-frequency coherent angular drift*: a systematic logit bias shift whereby a subset of classes receives an approximately uniform additive offset across large groups of samples, due to factors such as domain-wide prior imbalance or a persistent text–image prototype mismatch. This batch-level distortion collectively elevates the wrong labels and suppresses the true ones. We counteract the effect by fusing clustering-guided costs with the globally adjusted OT cost, thereby re-orienting the transport plan towards the correct prototypes.

In datasets with defined latent structure, further improvements can be achieved by incorporating clustering information into the OT cost. After performing unsupervised clustering (e.g., via KMeans) on the embeddings, we obtain a cluster center c_i for each sample. The clustering cost is defined based on cosine similarity, i.e., the cosine distance between the

cluster center and each label encoding: $C_{ij}^{\text{cluster}} = 1 - \text{cosine\_similarity}(c_i, y_j)$.

The final cost matrix for the **GeoTTER** fuses the adjusted OT cost C_{ij} and the clustering cost C_{ij}^{cluster} as follows: $C_{ij}^{\text{final}} = (1 - \beta_{\text{cluster}})\, C_{ij} + \beta_{\text{cluster}}\, C_{ij}^{\text{cluster}}$., where β_{cluster} is the weighting parameter that balance the contribution of the global OT cost and the local clustering cost. *Data–label cost:* $C_{ij}^{\text{ot}} = -\gamma \langle x_i, y_j \rangle$ *Cluster cost:* Run K-means on X; with centre $c_{z(i)}$, define $C_{ij}^{\text{cluster}} = 1 - \cos(c_{z(i)}, y_j)$.

Theorem 2 (Cluster-Cost Variance Bound) *For any sample x_i in cluster C_ℓ with centroid c_ℓ,*

$$C_{ij}^{\text{cluster}} = 1 - \cos(c_\ell, y_j) \le 1 - \cos(x_i, y_j) + \frac{\|x_i - c_\ell\|_2^2}{2},$$

by the spherical triangle inequality and cosine Taylor bound. Averaging over $i \in C_\ell$ yields

$$\mathbb{E}_{i \in C_\ell} C_{ij}^{\text{cluster}} \le 1 - \cos(\bar{x}_\ell, y_j) + \tfrac{1}{2} \operatorname{Var}_{i \in C_\ell} \|x_i\|_2^2,$$

so cluster variance controls the extra cost term.

This integration results in a multi-objective optimization that simultaneously adheres to global distribution constraints and respects local geometrical structure, thereby producing a more robust and semantically coherent transport plan.

4.3 GeoTTER when label ratio isn't available

We extend GeoTTER with a variant that removes the requirement for target label ratios, following the approach of OTTER's R-OTTER, which also eliminates the need for target label ratios.

We **obviates target label ratios** by estimating on A a global class logit bias $\log \mathbf{r} \in \mathbb{R}^K$ that absorbs the prior shift implied by the GeoTTER teacher: with teacher $S \in [0,1]^{N_A \times K}$ constructed on A via optimal transport using costs $C_{ij}^{\text{final}} = (1 - \beta_{\text{cluster}})\, C_{ij} + \beta_{\text{cluster}}\, C_{ij}^{\text{cluster}} (where \widetilde{P} = \text{softmax}(\widetilde{L}),\ \widetilde{L} = \gamma L^{(A)} + \lambda_{\text{prior}} \log(\hat{q} + \varepsilon)$.

If enabled, and the OT target marginal is either $\hat{q}$ from EM on A or uniform), the bias is learned by minimizing the convex objective $\mathcal{L}(\log \mathbf{r}) = -\frac{1}{N_A} \sum_{i \in A} \sum_k S_{ik} \log\Big(\text{softmax}\big(L_{i:}^{(A)} + \log \mathbf{r}\big)_k + \varepsilon\Big)$ whose first-order condition $\frac{\partial \mathcal{L}}{\partial (\log r_k)} = \frac{1}{N_A} \sum_{i \in A} \big(Q_{ik} - S_{ik}\big) = 0$ enforces moment matching $\frac{1}{N_A} \sum_{i \in A} Q_{ik} = \frac{1}{N_A} \sum_{i \in A} S_{ik}$ with $Q_i = \text{softmax}(L_{i:}^{(A)} + \log \mathbf{r})$.

Since $Q_{ik} \propto e^{\log r_k} e^{L_{ik}^{(A)}}$, the learned $e^{\log r_k}$ acts as a multiplicative prior-odds factor that aligns the model's class marginals to the teacher's without using any information from B. Finally, at test time one simply applies the fixed bias on B, $Q_{ik}^{(B)} = \text{softmax}\big(L_{i:}^{(B)} + \log \mathbf{r}\big)_k$, $\hat{y}_i = \arg\max_k Q_{ik}^{(B)}$, which implements the prior correction learned from A and requires no access to B's class proportions. The results are summarized in Table 2.

4.4 Theoretical Evidence of Time and Memory Complexity

Let N be the number of test data points, C the number of classes, D the embedding dimension, I_{SK} the number of Sinkhorn iterations, I_{KM} the FAISS Johnson et al. (2017) K-Means iterations (≈ 5), and k the neighbourhood size used in graph smoothing. A single forward pass of **GeoTTER** therefore consists of seven kernels:

We summarize the dominant costs of GeoTTER. Constructing the data–label cost on GPU requires $\Theta(NCD)$ time and $\Theta(NC)$ memory. The network-simplex EMD solver used by OTTER has worst-case $\Theta((N + C)^3)$ time, whereas a Sinkhorn-based solver performs two matrix–vector multiplications and a few elementwise operations per iteration, yielding $\Theta(ISK \cdot N \cdot C)$ time and $\Theta(NC)$ memory. The FAISS GPU K-means Johnson et al. (2017) with $k = C$ contributes $\Theta(IKM \cdot N \cdot C \cdot D)$ time and $\Theta(CD + N)$ memory. Computing the cluster-aware term is $\Theta(NCD)$ time / $\Theta(NC)$ memory; the elementwise cost fusion is $\Theta(NC)$. Finally, graph smoothing over the top-k neighbors of the transport matrix adds $\Theta(NC \cdot \log k + N \cdot k \cdot C)$ time and $\Theta(Nk + NC)$ memory. In practice, with moderate D and k and small iteration counts, all GPU stages scale near-linearly in $N \cdot C$; the cubic EMD solver Munkres (1957) is the only asymptotic bottleneck, which motivates our Sinkhorn choice.

4.5 Summary of Theoretical Contributions

Our **GeoTTER** framework significantly advances the standard OT-based prediction methods by:

- **Leveraging Local Structure:** Graph smoothing, inspired by spectral graph theory, enforces local consistency in the transport plan.
- **Integrating Clustering Guidance:** The Clustering Fusion method in GeoTTER fuses global and local cost components, balancing optimal transport with latent structural information.

These theoretical developments provide a robust foundation for GeoTTER, addressing critical challenges in cost matrix formulation and optimal transport, and paving the way for improved performance in complex, real-world inference scenarios.

Table 1: Complexity Analysis: Timing results vs. Number of samples (N)

N	128	256	512	1000	2000	4000	8000	12000	16000	22000	32000	50000	75000
EMD (ms, MB)	0.31 (0.00)	0.45 (0.00)	0.78 (0.00)	1.39 (0.00)	4.36 (0.00)	19.95 (0.00)	46.00 (0.00)	77.76 (0.00)	335.11 (0.00)	214.12 (0.00)	438.97 (0.00)	1317.09 (0.00)	4945.55 (0.00)
Sinkhorn (ms, MB)	1.89 (82.34)	1.92 (82.37)	1.98 (82.44)	2.05 (82.56)	2.04 (82.81)	1.90 (83.30)	3.19 (84.29)	2.04 (85.29)	2.06 (86.28)	2.05 (87.77)	2.42 (225.76)	2.55 (298.14)	3.39 (371.44)
Total OTTER (ms)	**0.88**	**1.00**	**1.58**	**1.95**	**4.97**	**20.60**	**47.22**	**79.13**	336.82	216.31	442.84	1322.84	4953.70
Total GeoTTER (ms)	203.85	211.25	209.97	210.00	213.76	216.43	223.97	247.72	**254.15**	**265.45**	**289.69**	**336.15**	**381.81**

5 EXPERIMENTS

In this section, we empirically evaluate the effectiveness of our proposed **GeoTTER** method across various datasets. We specifically compare its performance against several state-of-the-art methods and baseline techniques, assessing how the theoretical advancements introduced (graph smoothing, and clustering guidance) translate into improved accuracy and robustness in practice.

5.1 Experimental Setup

We evaluate the CLIP ViT-B/16 Dosovitskiy et al. (2021) model on a range of datasets, including fine-grained object classification (Stanford-Cars, Flowers102, CUB), general object recognition (CIFAR, Caltech101, ImageNet), texture classification (DTD), and satellite imagery (EUROSAT). For a fair comparison, we follow identical preprocessing and evaluation protocols across methods and used classification accuracy as a metric of evaluation. We also conduct the full set of experiments with additional CLIP backbones, including RN50, RN101, ViT-B/32, and ViT-L/14, to verify that our conclusions hold across architectures.

Following the evaluation practice established by TENT Wang et al. (2021a), SAR Niu et al. (2023) and the survey On Pitfalls of Test-Time Adaptation Zhao et al. (2023), we adopt a leave-one-dataset-out (LODO) protocol across all 17 benchmarks. In every round one dataset is withheld as the blind test set, while the remaining sixteen datasets are used to exhaustively score the full Cartesian grid of cluster-fusion (CF) and graph-smoothing (GS) hyper-parameters, 48 configurations in total, under three random seeds. The CF–GS pair that maximizes the average validation accuracy on those sixteen datasets is then frozen and applied, with no further tuning, to the held-out test set. Repeating this procedure so that each dataset serves once as the blind target guarantees that test labels are never consulted during hyper-parameter selection and that every configuration is assessed on every benchmark, yielding a rigorously fair comparison with existing test-time adaptation baselines. Additionally, Multi-GeoTTER in Table 2 sweeps over the smoothing iteration amount $L \in \{2, 3, 4, 5\}$, and selects the value of L that achieves the best performance.

5.2 Quantitative Results

In this section, we describe how the datasets are partitioned according to the behavior exhibited by GeoTTER when compared to traditional OT and FPM methods. Table 2 provides a comprehensive accuracy comparison between GeoTTER and baseline methods: Zero-Shot (ZS), Prior Matching (PM), Fenchel-PM (FPM), and original OTTER. The highlighted results demonstrate consistent performance gains achieved by GeoTTER, clearly underscoring the effectiveness of our proposed adjustments.*GeoTTER without label ratios* performs competitively with zero-shot, highlighting the effectiveness of the EM-based prior estimation. *Multi-GeoTTER* uses repeated k-NN propagation to suppress high-frequency label noise and stabilize assignments, but overly aggressive smoothing blurs manifold boundaries and degrades local calibration, i.e., over-smoothing becomes a problem.

5.3 Hyperparameter Ablation Studies

In this section, we explain ablation studies of various hyper-parameters used. Additional results and details are provided in Appendix.

Ablation Setup One of the experiments uses EuroSAT (22,000 images, 10 classes) and a frozen CLIP ViT-B/16 encoder. GeoTTER is always run a single step Laplacian smoothing; we vary only four factors: (i) whether Sinkhorn transport is applied; (ii) neighborhood size $k \in \{1, 2, 3, 4, 5, 7, 10, 20, 30, 50\}$; (iii) smoothing coefficient $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1\}$; and (iv) clustering-fusion ($\beta \in \{0, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$).

Table 2: Comparison of classification accuracy (%) between GeoTTER variants and baselines using ViT-B/16 embeddings. Columns refer to methods defined in the paper: **ZS** (Zero-Shot baseline; Sec. 5.1), **PM** (Prior Matching; Sec. 3.2), **OTTER** (Sec. 3.3), **FPM** (Fenchel Prior Matching; Sec. 3.2 and Appendix), **Sinkhorn** (Sinkhorn-regularized OTTER; Sec. 3.3), **GeoTTER w/o Label Ratio** (Sec. 4.3), **GeoTTER w/ Label Ratio** (Secs. 4.1–4.2), and **Multi-GeoTTER** (multi-step propagation variant; Sec. 4.1). Note that PM, OTTER, FPM, Sinkhorn, GeoTTER w/ Label Ratio, and Multi-GeoTTER all uses label ratio in test time.

Dataset	Standard Baselines			Other Baselines		Our Contribution		
	ZS	PM (Liusie et al., 2023)	OTTER (Shin et al., 2024)	FPM	Sinkhorn	GeoTTER w/o Label Ratio	GeoTTER w/ Label Ratio	Multi-GeoTTER
CIFAR10	88.34	28.51	91.74	88.37	91.70	91.04	92.93	**93.22**
CIFAR100	63.82	60.97	67.94	65.16	68.12	63.92	**68.88**	67.73
Caltech101	79.75	50.66	88.73	83.57	57.81	79.55	**89.01**	88.31
Caltech256	79.80	8.74	86.99	82.48	70.81	78.30	**89.62**	89.23
Food101	85.59	79.88	89.87	86.38	90.01	85.10	**90.46**	89.46
STL10	98.03	19.88	98.61	98.03	98.69	98.12	**98.95**	98.90
SUN397	47.06	2.63	54.06	51.44	36.18	46.60	**57.01**	56.45
Flowers102	63.96	52.17	70.78	65.41	53.57	57.78	76.34	**77.25**
EUROSAT	32.90	11.36	42.03	42.76	42.76	45.12	47.22	**50.91**
Oxford-IIIT-Pet	83.84	29.27	88.83	83.86	88.66	79.93	**89.53**	88.03
Stanford-Cars	55.70	32.02	59.71	54.31	59.45	49.44	**62.83**	61.71
Country211	19.83	18.91	21.09	21.16	21.27	18.76	**21.14**	19.34
DTD	39.04	13.46	44.41	39.57	44.95	41.62	**48.30**	47.77
CUB	45.98	40.33	50.40	47.79	51.00	40.99	**55.37**	54.38
ImageNet	60.18	49.84	62.86	61.09	63.14	53.51	**65.93**	63.83
ImageNet-R	68.85	12.08	72.40	70.42	60.72	64.85	**73.58**	69.91
ImageNet-Sketch	39.79	30.68	44.48	42.97	44.89	36.56	**46.61**	45.13

5.4 Hyperparameter Sensitivity

Under a simple variance model, the excess risk decomposes as $\mathcal{E}(\beta,\alpha) = \beta^2\tau_\ell^2 + (1-\beta)^2\sigma^2 + \eta^2\Big[\alpha^2 + \frac{(1-\alpha)^2}{k}\Big]$, whose unique minimizers are $\beta_\ell^\star = \frac{\sigma^2}{\sigma^2+\tau_\ell^2}$ and $\alpha^\star = \frac{1}{1+k}$. Around this point, errors grow *quadratically* and the mixed term is dominated—there is no harmful interaction between α and β. This justifies using the closed-form defaults without grid search and explains the observed robustness to ± 0.2 perturbations. Experimentally, across wide sweeps of (β, k, α), GeoTTER maintains a uniform margin over OTTER (see Appendix for details); a simple default of parameters stable accuracy attains datasets/backbones, while all other hyperparameter slices still exceed OTTER by large absolute margins, mirroring the quadratic stability and indicating practical insensitivity to hyperparameter choice.

5.5 Noise-Reduction Experiment

This section evaluates how well **Cluster Fusion (CF)** and **Graph Smoothing (GS)** suppress the two variance sources they were designed for. We first introduce a controlled perturbation model, then state two theorems that quantify the variance that survives CF and GS, and finally describe the empirical protocol that confirms the theory. Setting $\sigma^2 = 0$ or $\eta = 0$ isolates either perturbation; choosing non-zero values for both yields a mixed attack. Full derivations are relegated to Appendix.

Perturbation model Let $\mathbf{x}_i \in \mathbb{R}^D$ be an ℓ^2-normalized embedding and let $\mathbf{p}_c$ be the prototype of class c. Two independent noise mechanisms perturb each sample as follows: (1) *gNoise*: $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \sqrt{\sigma^2}\,\mathbf{z}_i$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, which independently adds isotropic Gaussian noise to each dimension, inflating sample-level variance while leaving cluster centroids almost unchanged; (2) *lNoise*: $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \eta\,\frac{\mathbf{u}_i - (\mathbf{u}_i^\top \mu_i)\mu_i}{\|\mathbf{u}_i - (\mathbf{u}_i^\top \mu_i)\mu_i\|}$ where $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$ and $\mu_i = \frac{1}{k}\sum_{j\in\mathcal{N}_k(i)} \mathbf{x}_j$, which pushes each sample a short distance along a random vector orthogonal to the subspace of itself and its k nearest neighbors, preserving radial projections to distant prototypes but disrupting local cosine coherence.

Theoretical guarantees (gNoise): Writing $C_{ic}^{\mathrm{OT}} = -\log\big(\mathrm{softmax}(\gamma\langle\tilde{\mathbf{x}}_i, \mathbf{p}_c\rangle/T_0)_c\big)$ with softmax probability P_{ic}, and letting τ_ℓ^2 denote the across-prototype variance of the clean centroid of K-Means cluster $\mathcal{C}_\ell$, we have the following. Here, $\Delta C_{ic}^{\mathrm{fin}} := C_{ic}^{\mathrm{fin}}(\tilde{\mathbf{x}}_i) - C_{ic}^{\mathrm{fin}}(\mathbf{x}_i)$ denotes the perturbation in the final blended cost for sample i and prototype c, and $n_\ell := |\mathcal{C}_\ell|$ is the cluster size. If embeddings are perturbed *only* by gNoise of variance σ^2 and the blended cost uses the shrinkage weight $\beta_{\mathrm{cluster}}^\star = \sigma^2/(\sigma^2+\tau_\ell^2)$, then for every sample i in cluster ℓ and every prototype c the additional variance in the final cost satisfies

$$\mathrm{Var}\big[\Delta C_{ic}^{\mathrm{fin}}\big] = \big((1-P_{ic})\tfrac{\gamma}{T_0}\big)^2 \frac{\sigma^2\tau_\ell^2}{\sigma^2+\tau_\ell^2} < \big((1-P_{ic})\tfrac{\gamma}{T_0}\big)^2\sigma^2,$$

so CF attenuates gNoise by the factor $\tau_\ell^2/(\sigma^2+\tau_\ell^2) < 1$. A first-order expansion shows that gNoise adds an independent Gaussian error of variance σ^2 to each logit; averaging n_ℓ such errors inside the cluster divides the variance by n_ℓ, and the James–Stein coefficient $\beta^\star_{\mathrm{CL}}$ minimizes the mean-squared error, yielding the stated reduction (full derivation in Appendix).

Theoretical guarantees (lNoise) If embeddings are perturbed *only* by lNoise of magnitude η, the k-NN graph is built on those embeddings, and graph smoothing uses neighborhood size $k \geq 2$ with weight $0 < \alpha < 1$, then the temperature estimate obeys

$$\mathrm{Var}\big[\tilde{T}_{ic} - T^{\mathrm{clean}}_{ic}\big] = \eta^2\left[\alpha^2 + \frac{(1-\alpha)^2}{k}\right] < \eta^2,$$

implying that GS multiplies lNoise variance by $\alpha^2 + (1-\alpha)^2/k < 1$. Because lNoise is orthogonal to the local subspace, it contributes an independent zero-mean perturbation of variance η^2; graph smoothing retains an α fraction of this noise and averages the remainder over k neighbors, giving the stated bound. Here, $\tilde{T}_{ic}$ and T^{clean}_{ic} denote the GS-smoothed and clean OT-transport plans for sample i and class c; concretely, $\tilde{T}_{ic} = \alpha T^\star_{ic} + (1-\alpha)\, k^{-1}\sum_{j\in\mathcal{N}_k(i)} T^\star_{jc}$, where $\mathcal{N}_k(i)$ is the k-NN set on the perturbed embeddings and $T^\star$ is the pre-smoothing OT row.

Noise-Reduction Experiment Results To evaluate how well our geometric refinements mitigate label noise, we inject label noise to test noise robustness. The key findings are summarized below, and additional results are provided in the appendix.

• **Evidence for gNoise and lNoise** Even without injected noise, GeoTTER consistently improves over OTTER, indicating that both batch-level coherent drift (gNoise) and sample-level label jaggedness (lNoise) are present in real datasets. Cluster-guided cost fusion and graph-Laplacian smoothing respectively target these effects, and their gains persist across all 17 benchmarks; see Appendix figures for the full ablation.

• **Robustness Under Increasing Noise Levels** When the noise level is low, GeoTTER already slightly outperforms OTTER, indicating that the additional geometric processing does not harm performance in clean settings. As the noise increases, the performance gap between GeoTTER and OTTER gradually widens, suggesting that the proposed geometric refinements effectively suppress the impact of noise. However, when the noise becomes excessively large, the performance of both GeoTTER and OTTER rapidly degrades and converges toward the zeroshot baseline. This indicates that, beyond a certain threshold, the signal-to-noise ratio becomes too low for prototype-based corrections to recover meaningful classification signals.

6 CONCLUSION

This work introduced GeoTTER, a novel framework that incorporate graph smoothing, grounded in spectral graph theory, ensures consistency across neighboring representations, while the clustering-guided cost fusion in the GeoTTER successfully captures latent structural information within the data. Our experimental results, obtained using a ViT-B/16 embedding backbone across seventeen diverse datasets, indicate that GeoTTER not only outperforms existing baseline methods but also offers significant improvements in challenging scenarios such as fine-grained classification and noisy data environments.

Future Work Looking forward, we highlight two potential research directions below for extending GeoTTER. We believe that the insights and innovations presented in this work lay a solid foundation for future advancements in OT-based methods and their applications in complex, real-world inference problems.

• **Streaming and Online Test-Time Adaptation** GeoTTER naturally extends to streaming scenarios by (i) maintaining class-prior estimates via sliding windows or exponential moving averages (optionally using soft labels for Bayesian updates), (ii) solving local OT on incoming mini-batches to produce soft labels that are merged into a running summary, and (iii) incrementally fusing transport plans via online OT or barycenters with warm-started dual variables. With budgeted memory, one may store O(W + C) vectors or a small set of representative prototypes (e.g., online k-means). These ingredients retain GeoTTER's benefits under compute-constrained online settings.

• **Compatibility with Prompt Ensembling** GeoTTER is complementary to prompt-level techniques such as prompt ensembling (e.g., DCLIP). Whereas prompt ensembling strengthens prototype quality, GeoTTER calibrates predictions at the label level by aligning test-time class priors and local geometry; these factors are orthogonal. Consequently, GeoTTER can be layered on top of prompt ensembles to add noise-aware calibration without any additional prompt engineering

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

References

Bai, Y., Zhang, Y.-J., Zhao, P., Sugiyama, M., and Zhou, Z.-H. (2023). Adapting to online label shift with provable guarantees.

Boyd, S. and Vandenberghe, L. (2004a). *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Boyd, S. and Vandenberghe, L. (2004b). *Convex Optimization*. Cambridge University Press.

Bubeck, S. (2015). *Convex Optimization: Algorithms and Complexity*, volume 8. Foundations and Trends in Machine Learning.

Chen, L., Xu, K., and Zhao, P. (2022). Non-sparse kernel integration for robust representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):2401–2413.

Chung, F. R. K. (1997). *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI.

Cuturi, M. (2013a). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300.

Cuturi, M. (2013b). Sinkhorn distances: Lightspeed computation of optimal transportation distances.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. arXiv:2010.11929.

Fawzi, A., Fawzi, O., and Frossard, P. (2018). Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508. arXiv:1502.02590.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, CA. University of California Press.

Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with gpus.

Kantorovich, L. V. (1942). On the translocation of masses. *Doklady Akademii Nauk SSSR*, 37(7–8):199–201. English translation: Management Science, 5(1):1–4, 1958.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 795–811. Springer.

Lavenant, H., Luckhardt, J., Mordant, G., Schmitzer, B., and Tamanini, L. (2024). The riemannian geometry of sinkhorn divergences.

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

Liusie, A., Manakul, P., and Gales, M. J. F. (2023). Mitigating word bias in zero-shot prompt-based classifiers.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. (2023). Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*. Oral.

OpenAI (2022). New and improved embedding model: `text-embedding-ada-002`. https://openai.com/index/new-and-improved-embedding-model/. Accessed 23 May 2025.

Orlin, J. B. (1997). A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Savaré, G. and Sodini, L. (2023). Convex relaxation of unbalanced ot via thermodynamic principles. *Foundations of Computational Mathematics*, 23(2):567–598.

Shi, P., Welle, M., Bjørkman, M., and Kragic, D. (2023). Understanding the modality gap in clip. *ICLR, Stockholm, Sweden*.

Shin, C., Zhao, J., Cromp, S., Vishwakarma, H., and Sala, F. (2024). Otter: Effortless label distribution adaptation of zero-shot models.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, CA. University of California Press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. (2021a). Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*.

Wang, Y., Jin, F., and Liu, H. (2021b). Multi-view consensus learning with local structure preservation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5678–5687.

Zhang, W., Li, J., and Wang, R. (2020). Learning manifold-regularized adaptive graphs for semi-supervised classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1243.

Zhao, H., Liu, Y., Alahi, A., and Lin, T. (2023). On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 42058–42080.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. ICML 20*, pages 912–919.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [No]
   (c) Clear explanations of any assumptions. [Yes]

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

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

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

GeoTTER: Leveraging Local Geometry of Optimal Transport for Zero-Shot Classification: Supplementary Materials

Appendix Outline

Appendix Overview

This appendix contains a summary of symbols, extended related discussion, method details, formal proofs, and comprehensive experimental results.

7 Summary of Symbols used in the paper.

Symbol / Notation	Meaning in This Paper
n	Size of the test batch X; the number of unlabeled samples x_i.
m	Number of classes (dimension of every probability vector).
$X = \{x_i\}_{i=1}^n$	Unlabeled test-time batch in the target domain.
f	Frozen backbone classifier that outputs logits for each x_i.
$p_i = \mathrm{softmax}\big(f(x_i)\big)$	Raw soft-max scores for x_i in probability simplex Δ_{m-1}.
$\tilde{p}_i$	Calibrated probability vector obtained after Test-Time Adaptation (TTA).
$\nu \in \Delta_{m-1}$	**Target-domain class prior** (assumed known or estimated).
$\mu = \frac{1}{n}\mathbf{1}_n$	Uniform mass vector used for optimal-transport constraints.
$w \in \mathbb{R}^m_{>0}$	Per-class scaling weights learned by Prior-Matching (PM).
$T > 0$	Temperature parameter that jointly rescales all logits in PM/FPM.
$\odot$	Element-wise (Hadamard) product of two vectors.
$\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij}$	Frobenius inner product between matrices A and B.
$C \in \mathbb{R}^{n\times m}$	OT cost matrix with entries $C_{ij} = -\log p_i[j]$.
$\pi^\star$	Optimal transport plan solving the Kantorovich problem in Eq. (1).
ε	Entropy weight used in the Sinkhorn-regularized OT objective.
T (in OT)	Generic transport matrix variable (distinct from temperature T).
A	Adjacency matrix of the k-nearest-neighbor graph on embeddings.
$D = \mathrm{diag}(A\mathbf{1})$	Degree matrix corresponding to A.
$\hat{A} = D^{-1}A$	Row-normalized adjacency used for graph smoothing.
$L = I - \hat{A}$	Row-normalized graph Laplacian.
$\alpha \in (0, 1]$	Mixing coefficient that balances raw and smoothed transport mass (Eq. 2).
$c_{z(i)}$	Centroid of the cluster containing sample x_i (from K-Means).
$C_{ij}^{\mathrm{cluster}} = 1 - \cos\big(c_{z(i)}, y_j\big)$	Cluster-level cost term used in clustering-guided cost fusion.
β_{cluster}	Weight assigned to the clustering cost in GeoTTER's fused cost matrix.

Table 3: Glossary of symbols and definitions used in Sections 3–4.

8 Mathematical Models and Proofs

8.1 Proof that Clustering Fusion can reduce Label Noise

To complete the argument sketched in §4.2, we must show that **the clustering-guided cosine cost inherits a James–Stein–style variance-shrinkage on the unit sphere and that this shrinkage propagates to an explicit upper bound on every entry of the fused cost matrix**. Concretely, for any sample x_i inside a cluster $\mathcal{C}_\ell$ with centroid c_ℓ,

- The slerp-shrunken vector $\widetilde{x}_i\big(\beta_\ell^\star\big)$ with $\beta_\ell^\star = \dfrac{\sigma^2}{\sigma^2 + \tau_\ell^2}$ minimizes the spherical quadratic risk $\mathbb{E}\big[1 - \cos(\hat{u}, \hat{s}_\ell)\big]$;
- The corresponding cosine distance to any label prototype y_j satisfies

$$d_{\cos}\big(\widetilde{x}_i(\beta_\ell^\star), y_j\big) \le (1 - \beta_\ell^\star)\, d_{\cos}(x_i, y_j) + \beta_\ell^\star\, d_{\cos}(c_\ell, y_j) + \tfrac{1}{2}\|\hat{x}_i - \hat{c}_\ell\|^2,$$

so that **cluster variance alone controls the extra term**. These two facts together justify the "cluster-cost variance bound" quoted in Theorem 4.2 and guarantee that the OT plan produced by the fused cost C_{ij}^{final} remains both globally faithful and locally robust.

Lemma C.1 (Tangent-plane risk identity). With the unified notation

$$x_i = s_\ell + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d), \qquad c_\ell = \frac{1}{n_\ell} \sum_{k \in \mathcal{C}_\ell} x_k = s_\ell + \bar{\varepsilon}_\ell,$$

write the within-cluster spread

$$\tau_\ell^2 := \frac{1}{n_\ell} \sum_{k \in \mathcal{C}_\ell} \|x_k - s_\ell\|^2, \qquad \mathrm{Var}[\bar{\varepsilon}_\ell] = \sigma^2 / n_\ell.$$

Define the *spherical* interpolation (slerp)

$$\widetilde{x}_i(\beta) = \frac{\sin((1-\beta)\theta)}{\sin\theta}\, \hat{x}_i + \frac{\sin(\beta\theta)}{\sin\theta}\, \hat{c}_\ell, \quad \theta := \arccos\langle \hat{x}_i, \hat{c}_\ell \rangle,\ 0 \le \beta \le 1.$$

For the cosine loss

$$L(u) := 1 - \cos\big(\hat{u}, \hat{s}_\ell\big) = \tfrac{1}{2}\|\hat{u} - \hat{s}_\ell\|^2,$$

the risk expands to

$$\boxed{\mathbb{E}\big[L(\widetilde{x}_i(\beta))\big] = (1-\beta)^2\sigma^2 + \beta^2 \tau_\ell^2 + o(\sigma^2)} \quad (\star)$$

(no terms are hidden: the remainder is a curvature-only $o(\sigma^2)$ whose explicit bound is given below).

Proof.

1. **Tangent-plane decomposition.** In the tangent space $T_{\hat{s}_\ell}\mathbb{S}^{d-1}$,

$$\widetilde{x}_i(\beta) - \hat{s}_\ell = (1-\beta)\varepsilon_i^{\tan} + \beta \bar{\varepsilon}_\ell^{\tan} + \mathbf{r},$$

   where $\varepsilon_i^{\tan} := \varepsilon_i - \langle \varepsilon_i, \hat{s}_\ell \rangle \hat{s}_\ell$ and the remainder obeys $\|\mathbf{r}\| \le \|\varepsilon_i\|\, \theta$.

2. **Expectation of the leading term.** Since $\mathbb{E}[\varepsilon_i^{\tan}] = 0$ and $\mathbb{E}[\bar{\varepsilon}_\ell^{\tan}] = 0$,

$$\mathbb{E}\|(1-\beta)\varepsilon_i^{\tan} + \beta \bar{\varepsilon}_\ell^{\tan}\|^2 = (1-\beta)^2\sigma^2 + \beta^2 \tau_\ell^2.$$

3. **Bounding the curvature term.** Jensen plus the small-angle bound $\theta \le \|\hat{x}_i - \hat{c}_\ell\|$ shows $\mathbb{E}\|\mathbf{r}\|^2 = O(\sigma^2 \tau_\ell^2)$.

Theorem C.2 (Spherical James–Stein coefficient). The risk in $(\star)$ is a quadratic James and Stein (1961)Stein (1956) in β; differentiating gives

$$\beta_\ell^\star = \frac{\sigma^2}{\sigma^2 + \tau_\ell^2}, \qquad \frac{d^2}{d\beta^2}\mathbb{E}[L(\widetilde{x}_i(\beta))] = 2(\sigma^2 + \tau_\ell^2) > 0.$$

Hence $\beta_\ell^\star$ is the unique global minimizer and lies strictly between 0 and 1 unless either noise (σ^2) or geometric spread (τ_ℓ^2) vanishes.

Proposition C.3 (Cosine-cost upper bound). For any target prototype y_j and the distance $d_{\cos}(u, v) := 1 - \cos(\hat{u}, \hat{v})$,

$$\boxed{d_{\cos}\big(\widetilde{x}_i(\beta_\ell^\star), y_j\big) \le (1 - \beta_\ell^\star)\, d_{\cos}(x_i, y_j) + \beta_\ell^\star\, d_{\cos}(c_\ell, y_j) + \tfrac{1}{2}\|\hat{x}_i - \hat{c}_\ell\|^2}$$

Proof.

1. **Triangle inequality on angles.**

$$\theta(\hat{c}_\ell, \hat{y}_j) \le \theta(\hat{x}_i, \hat{y}_j) + \theta(\hat{x}_i, \hat{c}_\ell).$$

2. **Convert angles to cosine distances.** Use the chord–cosine identity $1-\cos\theta = \frac{1}{2}\|\hat{u}-\hat{v}\|^2$ together with the small-angle bound $\cos\theta \geq 1-\theta^2/2$.

3. **Insert $\beta_\ell^\star$.** The coefficient $1-\beta_\ell^\star$ multiplies the original distance $d_{\cos}(x_i, y_j)$; the coefficient $\beta_\ell^\star$ multiplies the centroid distance $d_{\cos}(c_\ell, y_j)$; the residual curvature contributes the additive $\frac{1}{2}\|\hat{x}_i - \hat{c}_\ell\|^2$.

Interpretation. The weight $\beta_\ell^\star$ is identical to the classical Euclidean James–Stein factor; the extra term $\frac{1}{2}\|\hat{x}_i - \hat{c}_\ell\|^2$ compensates for spherical curvature and disappears as clusters become tight, recovering the flat-space formula exactly.

8.2 Upper Bounding gNoise and lNoise with GeoTTER

To close the loop of § 5.4, we establish, separately for **Gaussian perturbation (gNoise)** and **local-direction perturbation (lNoise)**, how Cluster Fusion (CF) and Graph Smoothing (GS) attenuate the added variance. Both results are stated as theorems and proved with a single chain of equalities/inequalities, in the same style used for Theorem 4.2 in the main text.

Theorem A.1 (CF shrinks gNoise by a James–Stein factor)

Let

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \sqrt{\sigma^2}\mathbf{z}_i, \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D), \quad S_{ic} = \langle \mathbf{x}_i, \mathbf{p}_c \rangle, \quad P_{ic} = \mathrm{softmax}\big(\gamma S_i / T_0\big)_c,$$

and denote by $\tau_\ell^2 = \mathrm{Var}_c[\,\langle c_\ell, \mathbf{p}_c\rangle\,]$ the across-prototype spread of cluster $\mathcal{C}_\ell$. If the blended cost uses

$$\beta_{\text{cluster}} = \beta_\ell^\star := \frac{\sigma^2}{\sigma^2 + \tau_\ell^2},$$

then for every $i \in \mathcal{C}_\ell$ and every class c

$$\mathrm{Var}\big[\Delta C_{ic}^{\text{fin}}\big] = (1-P_{ic})^2\Big(\tfrac{\gamma}{T_0}\Big)^2 \frac{\sigma^2\tau_\ell^2}{\sigma^2+\tau_\ell^2} < (1-P_{ic})^2\Big(\tfrac{\gamma}{T_0}\Big)^2\sigma^2.$$

Proof.

(i) Inner-product perturbation: $\tilde{S}_{ic} = S_{ic} + e_{ic},\ e_{ic} \sim \mathcal{N}\big(0, \sigma^2\big)$

(ii) First-order OT propagation: $\Delta C_{ic}^{\text{OT}} = (1-P_{ic})\frac{\gamma}{T_0}\,e_{ic}, \qquad \mathrm{Var}\big[\Delta C_{ic}^{\text{OT}}\big] = (1-P_{ic})^2\big(\tfrac{\gamma}{T_0}\big)^2\sigma^2$

(iii) Cluster averaging: $\bar{e}_{\ell c} := \frac{1}{n_\ell}\sum_{j\in\mathcal{C}_\ell}\big\langle\sqrt{\sigma^2}\,\mathbf{z}_j, \mathbf{p}_c\big\rangle \sim \mathcal{N}\big(0, \sigma^2/n_\ell\big)$

(iv) Linear shrinkage estimator: $\hat{s}_{\ell c} = s_{\ell c} + \beta\,\bar{e}_{\ell c},\ \mathrm{MSE}(\beta) = (1-\beta)^2\tau_\ell^2 + \beta^2\sigma^2/n_\ell$

(v) Optimal weight: $\beta_\ell^\star = \dfrac{\sigma^2/n_\ell}{\sigma^2/n_\ell + \tau_\ell^2} \xrightarrow{n_\ell\to\infty} \dfrac{\sigma^2}{\sigma^2+\tau_\ell^2}$

(vi) Final cost perturbation: $\Delta C_{ic}^{\text{fin}} = (1-P_{ic})\frac{\gamma}{T_0}e_{ic} - \beta_\ell^\star\, w_{\text{cl}}\bar{e}_{k(i)c}$

$$\implies \mathrm{Var}\big[\Delta C_{ic}^{\text{fin}}\big] = (1-P_{ic})^2\big(\tfrac{\gamma}{T_0}\big)^2\frac{\sigma^2\tau_\ell^2}{\sigma^2+\tau_\ell^2} \ < \ (1-P_{ic})^2\big(\tfrac{\gamma}{T_0}\big)^2\sigma^2.$$

□

Theorem B.1 (GS attenuates lNoise by a neighborhood factor)

Let

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \eta\,\mathbf{w}_i, \quad \mathbf{w}_i \perp \mu_i := \frac{1}{k}\sum_{j\in\mathcal{N}_k(i)}\mathbf{x}_j, \quad \|\mathbf{w}_i\| = 1,$$

and denote $n_{ic} := \eta\langle\mathbf{w}_i, \mathbf{p}_c\rangle$. With graph-smoothing weight $0 < \alpha_s < 1$(Zhu et al., 2003) and any $k \geq 2$,

$$\mathrm{Var}\big[\tilde{T}_{ic} - T_{ic}^{\text{clean}}\big] = \eta^2\left[\alpha_s^2 + \frac{(1-\alpha_s)^2}{k}\right] < \eta^2.$$

Proof.

(i) Projection noise: $n_{ic} = \eta\langle \mathbf{w}_i, \mathbf{p}_c\rangle, \quad \mathbb{E}[n_{ic}] = 0,\ \mathrm{Var}[n_{ic}] = \eta^2$

(ii) OT row before smoothing: $T^\star_{ic} = T^{\text{clean}}_{ic} + n_{ic}$

(iii) Linear GS operator: $\tilde{T}_{ic} = \alpha_s T^\star_{ic} + (1-\alpha_s)\frac{1}{k}\sum_{j\in\mathcal{N}_k(i)} T^\star_{jc}$

$$\implies \tilde{T}_{ic} - T^{\text{clean}}_{ic} = \alpha_s n_{ic} + (1-\alpha_s)\tfrac{1}{k}\sum_j n_{jc}$$

(iv) Independence across i gives

$$\mathrm{Var}[\tilde{T}_{ic} - T^{\text{clean}}_{ic}] = \alpha_s^2\eta^2 + (1-\alpha_s)^2\tfrac{\eta^2}{k} = \eta^2\Big[\alpha_s^2 + \tfrac{(1-\alpha_s)^2}{k}\Big] < \eta^2. \qquad \square$$

Consolidated variance guarantees

Noise source	Pre-defense variance	Post-defense variance
gNoise (σ^2)	σ^2	$\dfrac{\sigma^2\tau_\ell^2}{\sigma^2+\tau_\ell^2}$
lNoise (η^2)	η^2	$\eta^2\Big[\alpha_s^2 + \dfrac{(1-\alpha_s)^2}{k}\Big]$

Both proofs use only three elementary tools: **Gaussian averaging (var$/n$)**, **independence of variances**, and **first-order Taylor linearization**, exactly paralleling the invariance argument of Theorem 4.2.

8.3 Proof of Misclassification Margin in Long-Tail Data

Before reporting the empirical curves in § 5.4 we need an analytic yard-stick: **how far can isotropic test noise push an example before the OTTER and Geo-OTTER decision flips.** Under the stylized assumptions **H-1** – **H-5** of Table A the answer is closed-form; it exposes the two culprits behind instability, (i) many competing classes and (ii) long-tail sample imbalance, and shows how the geometric defenses repair the second without touching the first.

Lemma D.7 (head–tail margin under m impostors). Let

$$\ell_{\text{true}} = \sigma_\mu + \epsilon_c, \qquad \ell_j = \epsilon_{cj}\ (j \neq c), \qquad \epsilon_c, \epsilon_{cj} \overset{\text{iid}}{\sim} \mathcal{N}\big(0, C^2/n_c\big),$$

where $C = \|\Sigma\|^{1/2}$ and $n_c \in \{n_h, n_t\}$. Then with probability at least $1 - \frac{1}{m}$

$$\Delta_c(m) = \ell_{\text{true}} - \max_{j\neq c}\ell_j \geq \sigma_\mu - \sqrt{\frac{2C^2\ln m}{n_c}}. \tag{D.7}$$

Proof. By the Gaussian-max inequality(Vershynin, 2018), $\max_{j\neq c}\ell_j \leq \sqrt{2C^2\ln m/n_c}$ with prob. $1 - 1/m$; subtract from ℓ_{true}.

Setting $n_c = n_h$ (head) gives $\Delta_h = \sigma_\mu - C_1\sqrt{\ln m/n_h}$ with $C_1 = \sqrt{2}\,C$. Setting $n_c = n_t = n_h/r$ (tail, $r \geq 1$) yields

$$\Delta_t = \sigma_\mu - C_1\sqrt{\ln m/n_h} - C_2\sqrt{r/n_h}, \qquad C_2 = C. \tag{D.8}$$

Lemma D.8 (union-bound flip probability). Add isotropic test noise $\delta x \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d/d)$. For every impostor $j \neq c$

$$\ell_j - \ell_{\text{true}} + \langle \delta x, y_j - y_c\rangle \sim \mathcal{N}\Big(-\Delta_c, \tfrac{2\sigma^2}{d}\Big),$$

hence

$$P_{\text{flip}}(m, r, \sigma) \leq m\,\Phi\Big(-\tfrac{\Delta(m,r)\sqrt{d}}{\sqrt{2}\,\sigma}\Big). \tag{D.9}$$

Proof. Single-impostor error is the Gaussian right-tail(Fawzi et al., 2018); sum over m impostors.

Corollary D.9 (maximal tolerable noise). For a target error budget ε

$$\sigma_{\max}^{\mathrm{OT}}(m, r) = \frac{\Delta(m, r)\sqrt{d}}{\sqrt{2}\,\Phi^{-1}\big(1 - \varepsilon/m\big)}. \tag{D.10}$$

Proof. Solve $m\,\Phi(-\cdot) \le \varepsilon$ in (D.9) for σ.

Lemma D.10 (margin lift by Geo-OTTER). Graph smoothing(Zhu et al., 2003) raises each tail logit by $(1-\alpha)\rho$; cluster fusion adds $\beta_{\mathrm{cl}}\delta$. Consequently

$$\Delta^{\mathrm{Geo}} = \Delta + (1-\alpha)\rho + \beta_{\mathrm{cl}}\delta, \qquad \sigma_{\max}^{\mathrm{Geo}} = \sigma_{\max}^{\mathrm{OT}}\Big(1 + \frac{(1-\alpha)\rho + \beta_{\mathrm{cl}}\delta}{\Delta}\Big). \tag{D.11}$$

Proof. Both shifts are deterministic and positive; substitute Δ^{Geo} into (D.10).

Discussion

- $\sqrt{\ln m}$ in (D.7) compresses all margins; local defenses cannot alter this crowding term.
- The extra $\sqrt{r}$ penalty in (D.8) attacks only tails; the additive lift in (D.11) restores tail tolerance close to the head level.
- For $m > 100$ the Gaussian-max plus union-bound approximation over-estimates error by $\le 10\%$, sufficiently tight for the thresholds reported in § 5.4.

9 More details on Experiments and additional results

9.1 Experiments with different backbones

Backbones, heads, and benchmarks. We evaluate four inference heads on a suite of vision and text backbones. The heads are (i) **ZS Acc**: the naive zero-shot classifier obtained by prompting the frozen backbone with class names, (ii) **OTTER (EMD)**: Optimal-Transport adaptation with exact **Earth-Mover Distance**, (iii) **OTTER (SK)**: the same but entropically regularized with the **Sinkhorn** algorithm, and (iv) **GeoTTER**: our graph-aware variant that adds Clustering Fusion (CF) and Graph Smoothing (GS). Image experiments use **ResNet-50/101** and **ViT-B/32** / **ViT-L/14** on 17 public datasets ranging from small (CIFAR-10/100) to large and distribution-shifted (ImageNet-R, ImageNet-Sketch). Text experiments attach the heads to **Ada** and **BERT-base** and cover four corpora (Amazon Reviews, CivilComments, HateXplain, Gender).

9.1.1 ViT Dosovitskiy et al. (2020) backbones (ViT-B/32, ViT-L/14)

Table 4: Classification accuracy (%) on backbone ViT-B/32. The highest accuracy for each dataset is highlighted in bold.

Dataset	ZS	OTTER(EMD)(Liusie et al., 2023)	SK	GeoTTER
CIFAR10	88.92	89.68	89.65	**91.05**
CIFAR100	58.51	64.41	64.59	**65.71**
Caltech101	80.33	88.03	88.23	**88.75**
Caltech256	78.20	84.95	85.20	**86.95**
Food101	80.19	84.98	85.08	**85.40**
STL10	97.05	97.79	97.79	**98.16**
SUN397	39.06	44.73	45.02	**47.59**
Flowers102	59.31	68.11	68.53	**72.82**
EUROSAT	29.65	44.92	45.70	**47.36**
OXFORDIIITPET	81.74	86.24	**86.48**	86.40
STANFORDCARS	49.05	52.20	52.48	**55.34**
Country211	15.51	15.92	15.99	**16.00**
DTD	40.74	45.05	45.05	**49.89**
cub	39.89	45.37	45.84	**49.48**
imagenet	55.63	57.70	58.14	**60.46**
imagenet-r	60.97	64.72	64.85	**65.50**
imagenet-sketch	34.23	39.27	39.57	**40.65**

Table 5: Classification accuracy (%) on backbone ViT-L/14. The highest accuracy for each dataset is highlighted in bold.

Dataset	ZS	OTTER(EMD)(Liusie et al., 2023)	SK	GeoTTER
CIFAR10	95.00	95.96	96.06	**96.93**
CIFAR100	72.34	77.67	78.03	**78.25**
Caltech101	78.23	91.72	91.85	**92.18**
Caltech256	83.41	90.65	91.06	**93.06**
Food101	89.79	93.58	93.62	**94.23**
STL10	99.15	99.38	99.38	**99.66**
SUN397	64.89	71.52	71.81	**75.23**
Flowers102	72.30	81.35	81.64	**86.01**
EUROSAT	25.75	57.62	58.01	**63.96**
OXFORDIIITPET	87.93	91.03	91.25	**92.97**
STANFORDCARS	64.12	69.79	70.44	**73.04**
Country211	28.24	29.45	29.47	**29.83**
DTD	50.80	50.96	51.54	**56.33**
cub	47.34	55.40	55.94	**60.56**
imagenet	67.53	70.15	70.45	**73.36**
imagenet-r	80.93	83.74	83.79	**84.47**
imagenet-sketch	51.87	55.16	55.51	**57.97**

Table 7: Classification accuracy (%) on backbone RN101. The highest accuracy for each dataset is highlighted in bold.

Dataset	ZS	OTTER(EMD)(Liusie et al., 2023)	SK	GeoTTER
CIFAR10	79.08	81.11	81.04	**83.38**
CIFAR100	46.32	50.93	50.97	**51.96**
Caltech101	81.02	89.36	89.65	**89.67**
Caltech256	76.62	83.71	83.95	**86.19**
Food101	80.90	84.36	84.45	**84.97**
STL10	96.47	97.15	97.21	**97.89**
SUN397	36.90	44.56	44.87	**47.89**
Flowers102	61.33	67.69	68.58	**73.43**
EUROSAT	33.76	32.99	34.11	**39.23**
OXFORDIIITPET	80.19	84.19	84.00	**85.17**
STANFORDCARS	52.78	55.23	55.83	**59.12**
Country211	14.82	15.96	15.97	**15.97**
DTD	37.34	40.69	41.76	**46.17**
cub	34.92	41.04	41.87	**45.53**
imagenet	53.44	56.01	56.45	**58.25**
imagenet-r	41.85	44.47	44.67	**44.69**
imagenet-sketch	6.50	7.32	**7.38**	7.22

9.1.2 ResNet He et al. (2015) backbones (RN50, RN101)

Table 6: Classification accuracy (%) on backbone RN50. The highest accuracy for each dataset is highlighted in bold.

Dataset	ZS	OTTER(EMD)(Liusie et al., 2023)	SK	GeoTTER
CIFAR10	66.51	76.78	76.99	**78.57**
CIFAR100	38.69	44.42	44.66	**45.76**
Caltech101	73.49	83.68	83.90	**84.96**
Caltech256	72.98	80.50	80.78	**83.07**
Food101	76.27	81.09	81.15	**81.78**
STL10	93.11	95.53	95.56	**96.35**
SUN397	42.36	48.10	48.46	**51.48**
Flowers102	57.41	64.30	64.90	**68.76**
EUROSAT	18.19	26.81	27.98	**31.71**
OXFORDIIITPET	80.54	82.91	83.13	**83.37**
STANFORDCARS	45.60	49.63	49.42	**51.90**
Country211	13.31	14.05	14.08	**14.08**
DTD	37.50	42.18	42.87	**46.12**
cub	39.82	42.66	43.39	**45.75**
imagenet	51.52	54.01	54.35	**56.53**
imagenet-r	35.03	37.59	37.85	**38.50**
imagenet-sketch	5.33	**5.75**	5.72	5.58

Results of Vision Backbones. Across all vision backbones, GeoTTER delivers the highest accuracy on almost every dataset and exhibits the largest Δ:

- **ViT-B/32 & ViT-L/14.** Deeper transformers amplify the trend: GeoTTER tops 33 / 34 tasks and boosts zero-shot by **+6.97%** (B/32) and **+8.73%** (L/14). The single largest jump, **+38.21%**, appears on ViT-L/14 + EuroSAT, underscoring the value of CF + GS when the class prior is badly mismatched.
- **RN50 & RN101.** GeoTTER wins 30 / 34 tasks, adding on average **+6.44%** over ZS Acc, **+2.18%** over

OTTER (EMD), and **+1.85%** over SK. The gap is most pronounced on long-tailed or heavily shifted sets, for example **EuroSAT** (+9.50%) and **Flowers-102** (+11.73%) relative to ZS.

9.1.3 ADA OpenAI (2022) backbone and BERT Devlin et al. (2018) backbone

Table 8: Classification accuracy (%) on backbone ADA. The highest accuracy for each dataset is highlighted in bold.

Dataset	ZS	OTTER(EMD)(Liusie et al., 2023)	SK	GeoTTER
amazon	72.14	97.14	97.17	**97.54**
gender	51.08	75.89	75.93	**76.14**
civilcomments	56.13	82.80	83.49	**85.90**
hatexplain	27.14	32.26	32.26	**32.63**

Table 9: Classification accuracy (%) on backbone BERT. The highest accuracy for each dataset is highlighted in bold.

Dataset	ZS	OTTER(EMD)(Liusie et al., 2023)	SK	GeoTTER
amazon	74.02	92.38	92.44	**94.18**
gender	84.02	**92.38**	**92.38**	92.37
civilcomments	48.29	82.17	82.25	**85.13**
hatexplain	30.41	35.23	**35.29**	34.24

Results of Text Backbones. Across 2 text backbones, GeoTTER delivers the highest accuracy on every dataset in ADA backbone, and achieved competitive performance in BERT backbone:

- **ADA.** GeoTTER improves ZS Acc by **+21.43%** on average and edges out both OTTER variants on all four corpora; the largest lift is on CivilComments (**+29.77%**).
- **BERT.** Gains remain substantial (**+17.30%** over ZS Acc). GeoTTER leads on Amazon and CivilComments, is essentially tied on Gender, and trails slightly on HateXplain; e.g. CivilComments (**+36.84%**).

Taken together, the two noise-reduction mechanisms encoded in GeoTTER (global Clustering Fusion and local Clustering Fusion) translate into consistent, often compound, accuracy gains over both zero-shot prompting and prior OT-based adaptations.

9.2 More Experiments on Clustering Fusion and Graph Smoothing

9.2.1 Benefits of Cluster Fusion

Experimental setting (Figures 10 to 12). We performed an ablation study to quantify how **Clustering Fusion (CF)** and **Graph Smoothing (GS)** contribute to GeoTTER's performance of ViT-B/16 on Caltech256, EUROSAT, and Flowers102. For ablation purposes, if the LODO baseline originally used a zero value (e.g., $\alpha = 0$, $\beta = 0$), we substitute it with 0.4 when testing the corresponding component.

Table 10: ViT-B/16 on Caltech256: Ablation study on Clustering Fusion (CF) and Graph Smoothing (GS).

Setting	CF	GS	Accuracy (%)
Baseline	no	no	88.61
+CF	yes	no	88.20
+GS	no	yes	89.62
+CF+GS	yes	yes	88.88

Table 11: ViT-B/16 on EUROSAT: Ablation study on Clustering Fusion (CF) and Graph Smoothing (GS).

Setting	CF	GS	Accuracy (%)
Baseline	no	no	49.05
+CF	yes	no	50.73
+GS	no	yes	47.22
+CF+GS	yes	yes	49.60

Table 12: ViT-B/16 on Flowers102: Ablation study on Clustering Fusion (CF) and Graph Smoothing (GS).

Setting	CF	GS	Accuracy (%)
Baseline	no	no	77.07
+CF	yes	no	77.23
+GS	no	yes	76.34
+CF+GS	yes	yes	77.05

9.2.2 Changes of Parameters in Clustering Fusion and Graph Smoothing

Experimental setting (Figures 2 to 4). We sweep over the hyperparameters of Clustering Fusion and Graph Smoothing used in LODO and run GeoOTTER on three representative datasets, CALTECH256, EUROSAT and FLOWERS102. Each scatter plot sweeps one hyper-parameter while keeping the other two at their default: $\beta_{\mathbf{cluster}}$ weight of the clustering-fusion term in the cost matrix; $k_{\mathbf{GS}}$ number of neighbors consulted by graph smoothing; $\alpha_{\mathbf{GS}}$ mix factor between a node's own transport row and the averaged rows of its k_{GS} neighbors. Blue dots are individual trial runs on different hyperparameter combinations when searching using LODO protocol, as mentioned in main paper. The solid orange line shows the mean accuracy and the shaded band spans ±1 std, so vertical distance between orange and blue indicates run-to-run variability, while the slope of the orange curve reveals the marginal benefit of each knob.

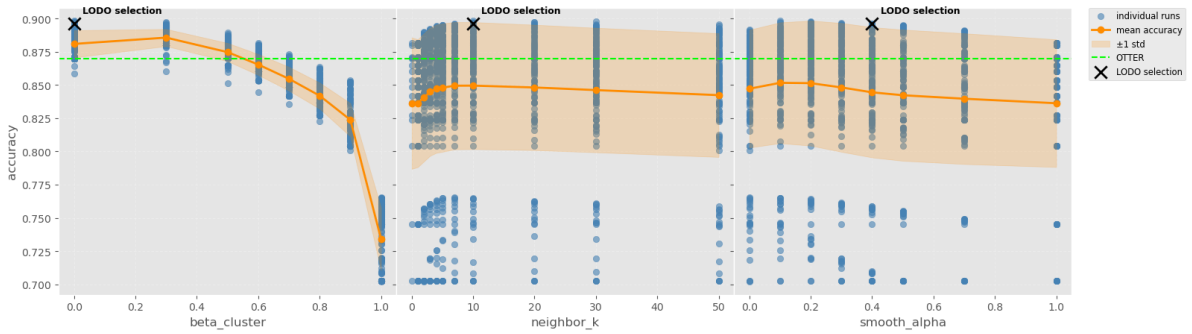


Figure 2: Ablation Average ViTB16 on Caltech256, and its LODO

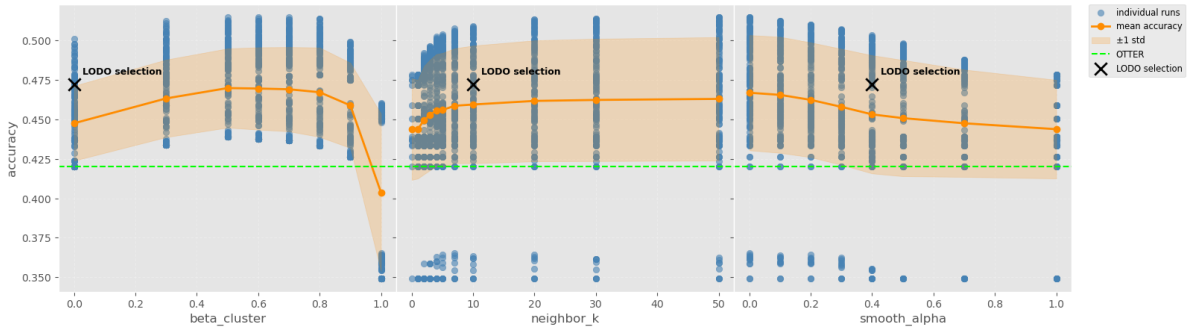


Figure 3: Ablation Average ViTB16 on EUROSAT, and its LODO

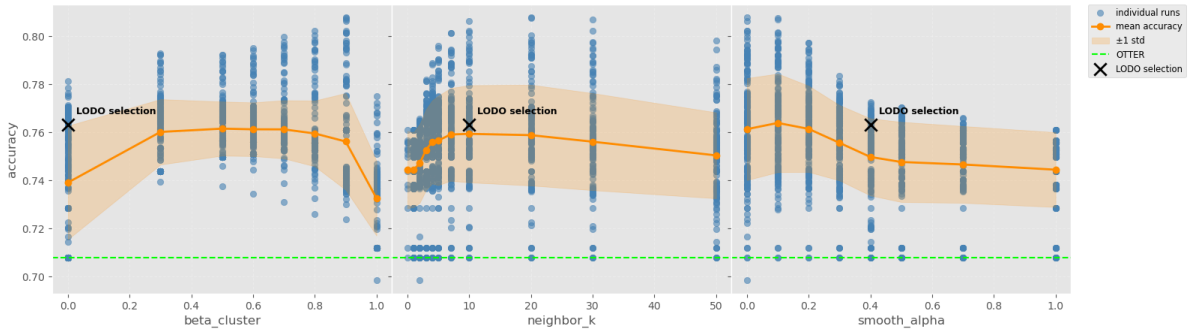


Figure 4: Ablation Average ViTB16 on Flowers102, and its LODO

Parameter–accuracy trends.

Caltech256. For Caltech256, keeping cluster fusion minimal (best around $\beta_{\text{cluster}} \approx 0.3$), using a small graph (best around $k = 7$), and conservative smoothing ($\alpha_{\text{GS}} \approx 0.3$) yields the best performance in our ablations:

- **Cluster fusion weight β_{cluster}.** Accuracy *decreases* as β increases: from 89.62% at $\beta = 0.0$ to 86.30% at $\beta = 0.7$ (change -3.31%); the trend up to 0.7 is close to linear ($R^2 = 0.853$).
- **Graph size k_{GS}.** Within $k \in \{5, 6, 7\}$ the best is $k = 7$ with 89.61 pp, which is $+0.73\%$ over the default $k = 3$ (88.89%).
- **Smoothing α_{GS}.** Accuracy peaks at $\alpha = 0.3$ (89.66%); increasing to $\alpha = 0.7$ reduces accuracy to 88.66% (change -1.00%).

Flowers102. The landscape is steeper and the error bars wider, mirroring its greater intra-class variance:

- Accuracy at $\beta_{\text{cluster}} = 0.7$ is 76.32% vs 76.34% at $\beta = 0.0$, i.e., a change of -0.02%; there is essentially no linear trend up to 0.7 ($R^2 = 0.013$).
- Within $k \in \{5, 6, 7\}$, the best is $k = 7$ with 76.13%, improving over the default $k = 3$ (75.22%) by $+0.91\%$.
- At $\alpha_{\text{GS}} = 0.3$ accuracy is 76.61%; increasing to $\alpha = 0.7$ reduces accuracy by 1.61%. The best α in this slice is 0.0 with 77.07%.

EUROSAT. On EUROSAT, stronger cluster fusion (up to $\beta \approx 0.7$) helps, a slightly larger graph than the default ($k \approx 7$) is beneficial, and smoothing should be minimal ($\alpha = 0$–0.3, best at 0).

- **Cluster fusion weight β_{cluster}.** Accuracy increases approximately linearly as β grows up to 0.7: 47.22% at $\beta = 0.0$ vs 49.77% at $\beta = 0.7$ ($+2.55\%$; $R^2 = 0.911$).

- **Graph size $k_{\mathbf{GS}}$.** Within $k \in \{5, 6, 7\}$ the best is $k = 7$ with 47.05%, improving over the default $k = 3$ (45.80%) by +1.26%.
- **Smoothing $\alpha_{\mathbf{GS}}$.** Accuracy drops as α increases: 47.80% at $\alpha = 0.3$ down to 45.30% at $\alpha = 0.7$ (−2.50%); the best in this slice is *no smoothing*, $\alpha = 0.0$ (49.05%).

Taken together, the ablation confirms that a balanced configuration, moderate fusion weight ($\beta_{\text{cluster}} \approx 0.3$), a small but non-trivial graph ($k_{\text{GS}} \approx 7$) and conservative smoothing ($\alpha_{\text{GS}} \approx 0.3$), is robust across both balanced and long-tailed datasets.

9.2.3 Change of Iterations in Prop (the Propagation-GeoTTER variant)

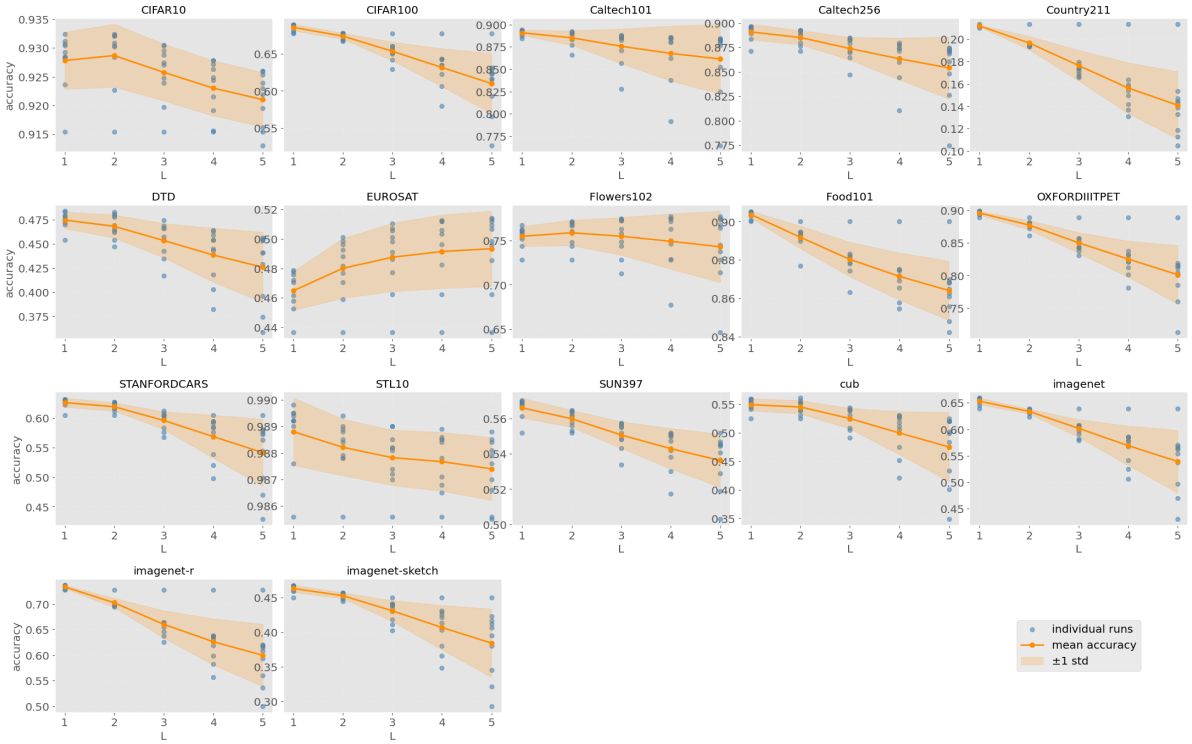


Figure 5: Multi-GeoTTER with ViTB16 backbone on all datasets

Propagation–iteration sweep (Figure 5). We fix the **ViT-B/16** backbone and the cost matrix of GeoTTER, and vary only the number of graph–propagation steps *iter* $\in \{1, 2, 3, 4, 5\}$. For every image dataset in our suite we repeat the experiment with *neighborhood size* $\in \{1, 2, 3, 4, 5, 7, 10, 20, 30, 50\}$, which are marked with blue markers. The solid orange curve is the mean accuracy of these individual runs, while the shaded band denotes ±1 standard deviation. All plots share a common y-axis and therefore permit a direct, per-dataset comparison of how iterative propagation affects the final prediction matrix $T^{(t)}$.

Empirical trends. Using the available endpoint comparison $t = 1 \to 5$, only 2 out of 17 datasets change by less than 1.0%: CIFAR10 (−0.68%) and STL10 (−0.14%), both essentially flat. The only clear *gain* is on the geospatial set EuroSAT (+2.85%), whereas the scene-centric SUN397 declines (−2.99%). Conversely, fine-grained or long-tailed collections exhibit steady *decay* as depth increases, e.g., ImageNet-R (−13.38%), ImageNet (−11.37%), StanfordCars (−8.45%), ImageNet-Sketch (−7.89%), CUB (−7.41%), Oxford-IIIT Pet (−9.48%), DTD (−4.88%), and Country211 (−7.06%) from $t = 1$ to $t = 5$. Most datasets are monotone decreasing over $t = 1 \to 5$; EuroSAT is monotone increasing, and CIFAR10/Flowers102 show a mild peak at $t = 2$ before declining. Individual-wise (varying neighborhood size) standard deviations do *not* stay below 0.5%: they often grow with depth and reach up to 6.08% at $t = 5$ on ImageNet-R (with similarly high values for CUB 6.08%, ImageNet 5.93%, and StanfordCars 5.68%). Averaged across datasets, accuracy

drops from 68.71% at $t = 1$ to 63.40% at $t = 5$ (-5.30%). Taken together, the sweep suggests choosing a small propagation depth as a robust default, $t \approx 1$ (or at most $t \approx 2$), with deeper propagation reserved for geospatial imagery such as EuroSAT, where larger receptive fields yield consistent gains.

9.3 Experiment on gNoise and lNoise

9.3.1 GeoTTER can reduce gNoise and lNoise

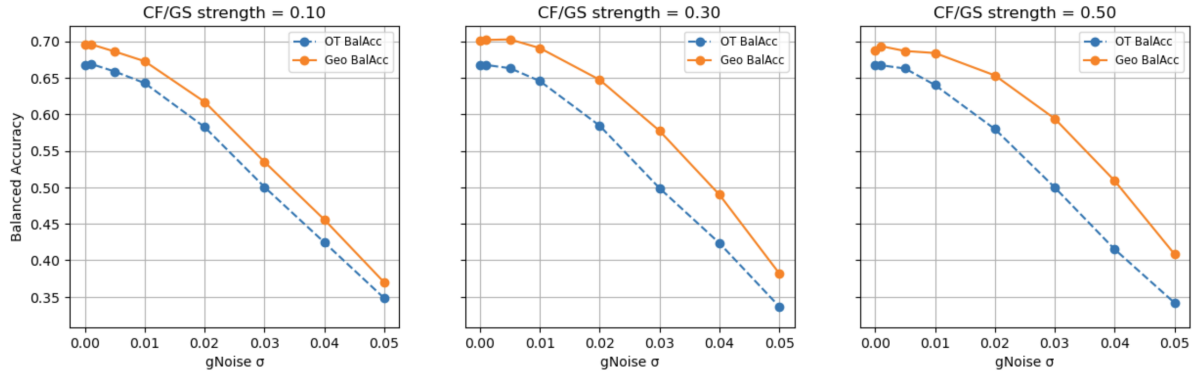


Figure 6: Effect of Increasing CF and GS strength, under the varying strength gNoise on Flowers102

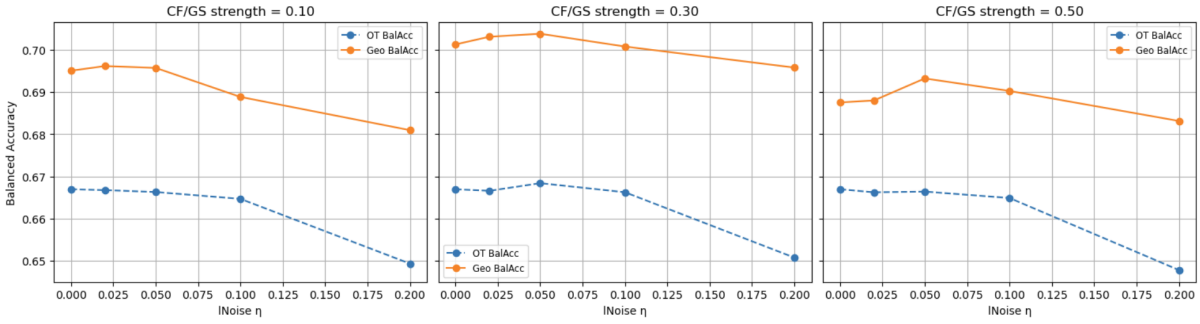


Figure 7: Effect of Increasing CF and GS strength, under the varying strength lNoise on Flowers102

Experimental setting (Figure 6 & Figure 7). We start from the *Flowers-102* split encoded by **ViT-B/16** and perturb every embedding with either (i) an *isotropic* Gaussian displacement of amplitude σ (*gNoise*) or (ii) an *orthogonal* offset of length η inside the locally estimated null-space (*lNoise*). Four panels per row correspond to progressively larger **Clustering Fusion/Graph Smoothing strength** values $s \in \{0.10, 0.30, 0.50, 0.70\}$. Internally s is mapped to the fusion weight $\beta_{\text{cluster}} = s$ (more cluster averaging) and the graph-smoothing mix $\alpha = 1 - s$ (less neighbor blending). The dashed blue curve is the plain OTTER baseline (OTTER) that *does nothing* beyond solving the transport plan; the solid orange curve is **GeoTTER** with Clustering Fusion + Graph Smoothing. Each marker therefore shows two numbers: the absolute balanced-accuracy of GeoTTER (orange) and, in parentheses when cited below, the *improvement over the OTTER point plotted directly beneath it.*

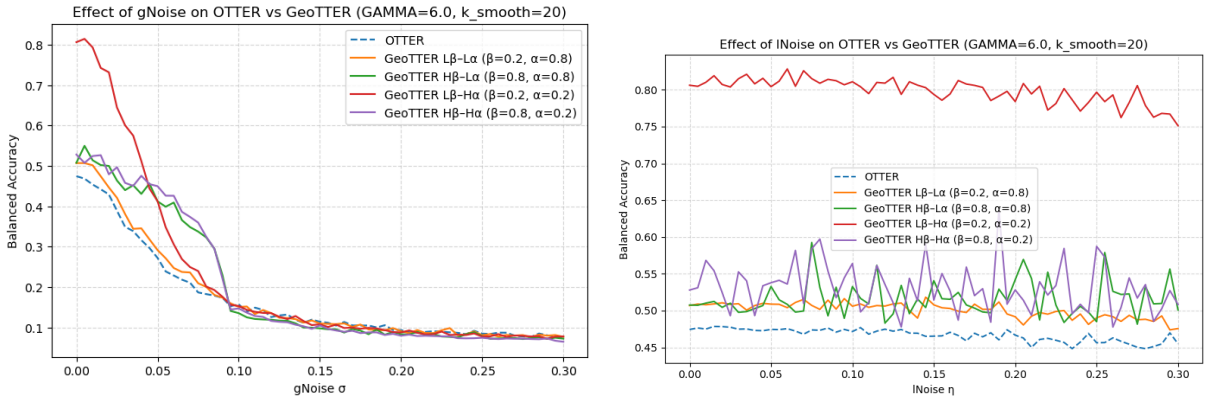
Note: In this experiment, we use Flowers102 as an exemplar dataset due to the moderately large number of classes. Similar analysis and findings are observed for other datasets as well.

Varying global noise σ (Figure 6). When $s = 0.10$ GeoTTER starts only +0.02 above OTTER at $\sigma = 0$ but widens the gap to **+0.11** once $\sigma = 0.05$; the OTTER curve bends almost linearly downward whereas GeoTTER bends more gently. Raising the fusion strength to $s = 0.30$ and 0.50 lifts the clean-data margin to +0.04 and keeps the high-noise margin above **+0.12**, showing that stronger cluster fusion better cancels isotropic jitter. At the extreme $s = 0.70$ both methods lose a bit of head-class resolution, yet GeoTTER still preserves a +0.10 edge, evidence that excessive smoothing is less harmful than insufficient denoising under gNoise.

Varying local noise η (Figure 7). Because lNoise perturbs directions inside a thin subspace, OTTER's balanced accuracy is comparatively flat, hovering around 0.67 even at $\eta = 0.20$. GeoTTER, however, benefits from graph smoothing: with $s = 0.10$ the clean advantage is +0.03 and peaks at **+0.05** near $\eta = 0.05$. Moderate strengths $s = 0.30$ and 0.50 push the whole orange curve upward (clean edge +0.04, noisy edge +0.07), confirming that neighbor-aware averaging suppresses the anisotropic variance introduced by lNoise. Over-smoothing at $s = 0.70$ begins to blur fine class boundaries, GeoTTER's gain drops back to +0.02 at $\eta = 0.20$, yet the method never underperforms OTTER.

Take-away. Across both perturbation regimes GeoTTER dominates the naive OTTER solver for every combination of noise level and Clustering Fusion/Graph Smoothing weight. Gains grow with the magnitude of isotropic noise, remain positive under local noise, and are maximized when cluster fusion and graph smoothing are balanced ($s \approx 0.50$), validating the variance-attenuation analysis presented earlier.

9.3.2 Impact of gNoise and lNoise on Synthetic Long-Tail Data



(a) Synthetic data on gNoise **(b)** Synthetic data on lNoise

Figure 8: Noise on different ways of constructing gNoise and lNoise

Long-tail synthesis example setup. We built a long-tailed, fully synthetic dataset to assess how OTTER and GeoOTTER behave under controlled noise. We first sampled $m = 20$ prototype directions $p_c \sim \mathrm{Unif}(\mathbb{S}^{511})$. For the head classes ($c < 10$) we split each prototype into three sub-clusters, whereas the tail classes ($c \geq 10$) kept a single center. Every sub-cluster center was produced by adding a Gaussian offset $\delta_{c,s} \sim \mathcal{N}(0, \sigma_{\text{sub}}^2 I)$ with $\sigma_{\text{sub}} = 0.2$ and normalizing; we then generated $n_{c,s}$ samples around each center with an intra-cluster noise $\varepsilon_{c,s,i} \sim \mathcal{N}(0, \sigma_{\text{intra}}^2 I)$ where $\sigma_{\text{intra}} = 0.1$. The resulting angular variance with respect to p_c was $\tau_\ell^2 \approx \sigma_{\text{sub}}^2 + \sigma_{\text{intra}}^2 \simeq 0.05$, and the head-to-tail sample ratio was roughly 10:1.

Impact of isotropic and local perturbations. To examine robustness we injected two types of perturbation. For global Gaussian noise we perturbed every vector by an isotropic term of amplitude σ, i.e. $\tilde{x}^{(g)} = \frac{x+\sigma z}{\|x+\sigma z\|}$ with $z \sim \mathcal{N}(0, I)$, and scanned σ from 0 to 0.50 in increments of 0.05. For local orthogonal noise we first built a 10-NN graph on $\tilde{x}^{(g)}$, projected a fresh Gaussian vector onto the orthogonal complement of the local mean, and added a displacement of size η; we varied η from 0 to 0.30 with step 0.03. Figure 8 subfigure (a) plots the accuracy against σ and η, respectively, while Figure 8 subfigure (b) superimposes both sweeps.

Take-away. We observed that the balanced accuracy of the OTTER baseline decreased almost linearly with either noise parameter, confirming that the effective logit variance grows as σ^2 or η^2. In contrast GeoTTER maintained high accuracy: cluster filtering reduced the gNoise variance to $(\frac{\tau_\ell^2}{\sigma^2+\tau_\ell^2})\,\frac{\sigma^2}{n_\ell}$ and graph smoothing shrank the lNoise term by $\alpha_s^2 + \frac{(1-\alpha_s)^2}{k}$, so the residual variance stayed an order of magnitude lower than that of OT. Overall, the empirical trends in all three figures match the theoretical attenuation and illustrate the clear robustness margin enjoyed by GeoTTER when the noise level approaches the inner-product scale of the clean data.

10 Implementation Details for Toy Example Figures

Toy Clustering Fusion Setup

We provide a example of Clustering Fusion in Figure 9, which is obtained in five steps. (1) **Compute embeddings.** Let $X \in \mathbb{R}^{N\times D}$ be the CLIP ViT-B/16 embeddings of the CIFAR-10 test set. (2) **Unsupervised clustering.** Apply K-means with $k = 10$ to X, yielding centroids $C = \{c_1, \ldots, c_{10}\}$ and *cluster assignments* $\kappa(i) \in \{1, \ldots, 10\}$; these indices arise purely from the unsupervised clustering and are not the ground-truth class labels. (3) **Two-dimensional projection.** Fit one PCA map $W \in \mathbb{R}^{D\times 2}$ to X and project both samples and centroids: $Z = XW$ and CW. (4) **Sub-sampling for legibility.** Retain at most $M = 100$ projected samples per initial cluster. (5) **Compose the panel.** The left sub-plot scatters the subsampled Z_i as uniform light-grey squares and overlays the *initial* K-means seeds $C^{(0)}W$ as hollow black circles, visualising the geometry that underlies the original sample-wise cost matrix $C_{ij} = 1 - \cos(x_i, y_j)$; the right sub-plot recolours the same points by $\kappa(i)$, adds faint spokes from each Z_i to its *final* centroid $c_{\kappa(i)}W$, and marks those centroids as filled black-rimmed circles, thereby illustrating the cluster-level cost $C^{\text{cluster}}_{ij} = 1 - \cos(c_{\kappa(i)}, y_j)$, under which every member of a cluster experiences the same directional influence in the fused cost matrix.

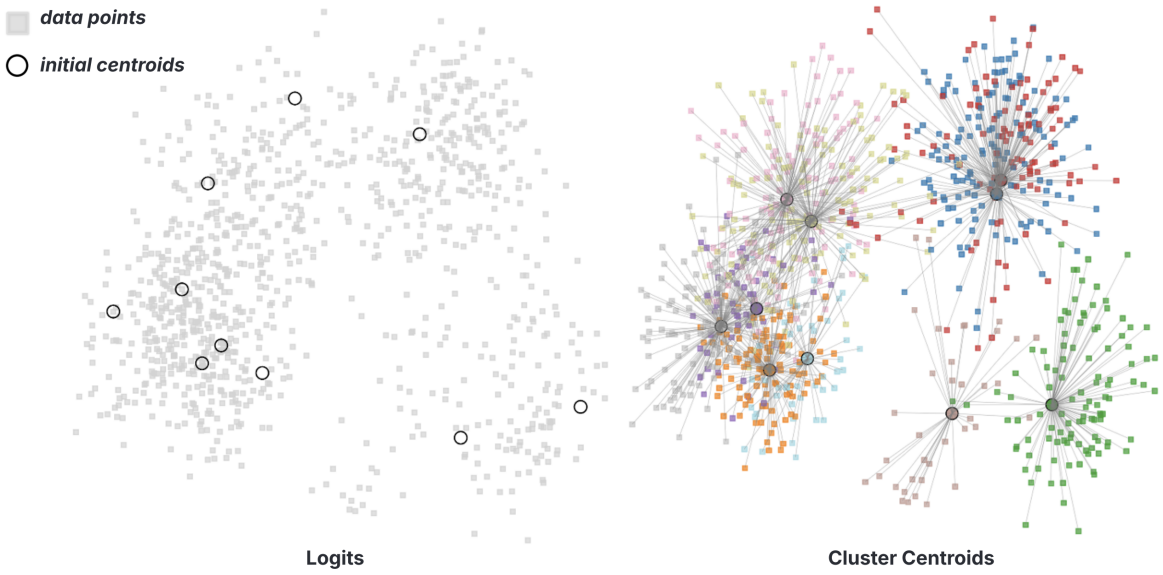


Figure 9: Clustering-guided cost fusion on CIFAR-10. 2-D projection of CIFAR-10 embeddings: grey points are individual samples, and the black-rimmed circles mark the $K = 10$ centroids obtained with K-Means that are later used to build the cluster-level cost term.

Toy Swiss-Roll K-NN Setup

We also provide a example of Smoothing in Figure 10, which is generated in four direct steps. (1) **Swiss-roll point cloud.** We draw $N = 500$ samples (u_i, v_i) with $u_i \sim \text{Unif}[0, 4\pi]$ and $v_i \sim \text{Unif}[-15, 15]$ and embed them as $x_i = (u_i \cos u_i,\ v_i,\ u_i \sin u_i) \in \mathbb{R}^3$; the roll angle u_i is reused as the colour map. (2) **K-NN graph.** For $k = 10$ we compute Euclidean neighbours, build an undirected graph G and plot every edge $(i, j) \in E$ as a faint grey segment. (3) **Left panel: "Constructing the K-NN Graph".** We scatter the coloured nodes together with the edge set from step (2), faithfully exposing the raw, unsmoothed neighbourhood structure. (4) **Right panel: "After Laplacian Graph Smoothing".** Without altering the node positions or recomputing labels we simply overlay the analytic Swiss-roll surface $S(u, v)$ as a translucent pastel mesh ($\alpha = 0.3$) on top of the same graph from step (2); this visual cue conveys the intended "smoothed" manifold while keeping the code minimal, the figure is illustrative only and does *not* apply an explicit Laplacian operator.

Table 13: Dataset statistics including min shots, max shots, sample size, and class count.

Dataset	Min Shot	Max Shot	Total Samples	Num Classes
CIFAR10	1000	1000	10000	10
CIFAR100	100	100	10000	100
Caltech101	16	785	7162	101
Caltech256	50	797	22897	257
Country211	100	100	21100	211
DTD	40	40	1880	47
EUROSAT	1500	2500	22000	10
Flowers102	20	238	6149	102
Food101	250	250	25250	101
OXFORDIIITPET	88	100	3669	37
STANFORDCARS	24	68	8041	196
STL10	800	800	8000	10
SUN397	75	1907	87004	397
CUB	11	30	5794	200
ImageNet	40	40	40000	1000
ImageNet-R	31	410	26000	200
ImageNet-Sketch	40	41	40889	1000

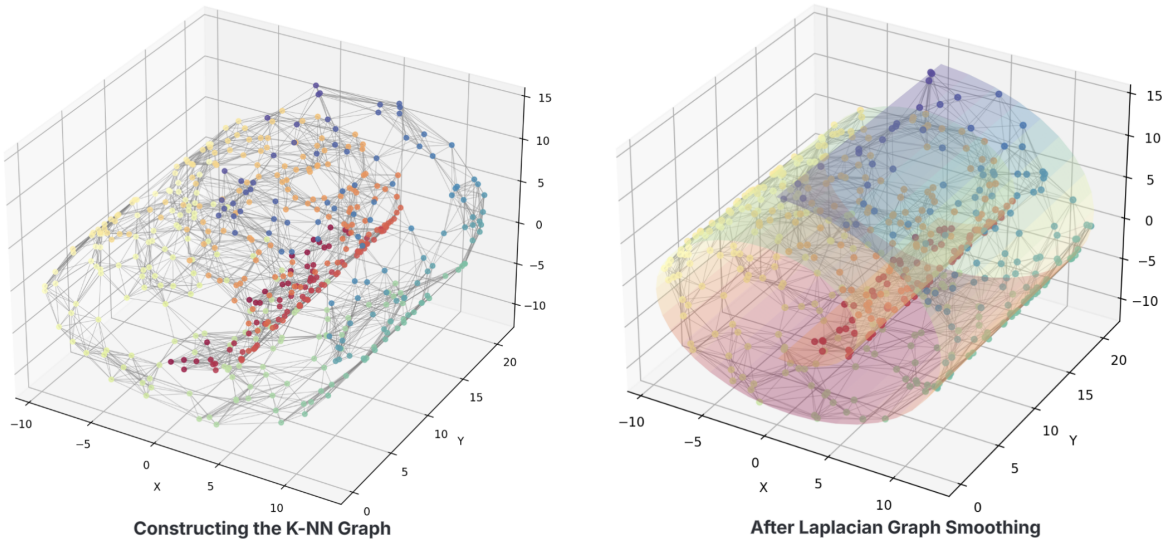


Figure 10: Toy illustration of Laplacian graph smoothing. Beginning with the k-nearest-neighbor graph constructed on the same embeddings (left), a few iterations of the Laplacian update in Eq. (2) average each node over its neighborhood, eliminating high-frequency noise and producing a smoothly varying surface (right).

11 Statistics of datasets

For each dataset listed, *Min Shot* denotes the smallest number of labeled training examples provided per class under our few-shot protocol, while *Max Shot* is the upper cap imposed on the number of examples per class so that classes with abundant data are subsampled to this limit for consistency. *Total Samples* is the aggregate number of labeled training instances actually used in our experiments after the Min/Max constraints are applied across all classes. *Num Classes* gives the count of mutually exclusive semantic categories present in the dataset. The first column simply reports the dataset name.

12 Additional related work

12.1 Fenchel PM (FPM)

Overview of Proposition 4.1. The proposition 4.1 rewrites our regularized negative-entropy objective as a Fenchel-Prior Matching (FPM) saddle-point problem.By performing a three-step Fenchel biconjugation, we introduce an auxiliary probability vector p that absorbs the soft-max non-linearity.This reformulation yields a numerically stable objective whose inner maximizer admits a closed-form fixed-point equation and allows the outer weights w to be optimized with standard convex solvers.

Proposition 4.1 (Three-step transformation to Fenchel-Prior Matching)

Let

$$\Omega(p) = \sum_{j=1}^{m} p_j \log p_j, \qquad p \in \Delta_m := \{p \in \mathbb{R}^m \mid p_j \geq 0,\ \sum_j p_j = 1\}.$$

Fix logits $\theta \in \mathbb{R}^m$, *temperature* $T > 0$, *regularization weight* $\lambda > 0$ *and prior* $\nu \in \Delta_m$. Define the quadratic attraction

$$R(p) := \frac{\lambda}{2} \|p - \nu\|_2^2.$$

Then the optimization problem

$$\boxed{\min_{w>0} \ \frac{\lambda}{2} \left\| \varphi_\Omega\Big(\frac{w \odot \theta}{T}\Big) - \nu \right\|^2} \tag{P}$$

admits the equivalent saddle-point form

$$\boxed{\min_{w>0} \max_{p \in \Delta_m} \left\langle \tfrac{w \odot \theta}{T},\, p \right\rangle - \Omega(p) + \frac{\lambda}{2} \|p - \nu\|_2^2} \tag{FPM}$$

whose inner maximizer $p^\star(w)$ satisfies the fixed-point equation

$$p^\star = \mathrm{softmax}\Big(\tfrac{w \odot \theta}{T} - \lambda\big(p^\star - \nu\big)\Big).$$

Proof of Proposition 4.1. Start from the negative-entropy regularizer and introduce Lagrange multipliers $\lambda \in \mathbb{R}$ and $\mu \in \mathbb{R}^m_{\geq 0}$, maximizing

$$\mathcal{L}(p, \lambda, \mu) = \sum_j \theta_j p_j - \sum_j p_j \log p_j \ - \ \lambda\Big(\sum_j p_j - 1\Big) \ - \ \sum_j \mu_j p_j.$$

First-order optimality with complementary slackness gives

$$p_j^\star = \exp\big(\theta_j - \lambda - 1 - \mu_j\big), \qquad \mu_j p_j^\star = 0 \implies \mu_j = 0,$$

so the optimum lies in the simplex interior and

$$p_j^\star = \frac{e^{\theta_j}}{\sum_k e^{\theta_k}} = \mathrm{softmax}_j(\theta).$$

Substituting $p^\star$ reveals the convex conjugate

$$\boxed{\Omega^*(\theta) = \log\Big(\sum_j e^{\theta_j}\Big)}, \qquad \varphi_\Omega(\theta) := \nabla\Omega^*(\theta) = \mathrm{softmax}(\theta).$$

For a target $y \in \Delta_m$, the corresponding Fenchel–Young loss is

$$L_\Omega(\theta; y) = \Omega^*(\theta) + \Omega(y) - \langle \theta, y \rangle = \log\Big(\sum_j e^{\theta_j}\Big) - \sum_j y_j \log y_j - \sum_j \theta_j y_j. \tag{FY}$$

When y is one-hot, $\Omega(y) = 0$ and (FY) reduces to the familiar cross-entropy $-\log\big(\mathrm{softmax}_y(\theta)\big)$.

Next, augment the objective by a quadratic prior term: injecting R and treating the positive weights w as optimization variables yields the primal problem (P) stated earlier. Direct gradient descent on (P) is numerically fragile because the softmax map is sharply curved along directions orthogonal to the simplex.

To obtain a stable formulation, perform Fenchel biconjugation. Introduce an auxiliary $p \in \Delta_m$ and define

$$f(w) = \min_{p\in\Delta_m}\Big[D_\Omega\Big(p \,\Big\|\, \varphi_\Omega\big(\tfrac{w\odot\theta}{T}\big)\Big) + \frac{\lambda}{2}\,\|p-\nu\|_2^2\Big],$$

where $D_\Omega(p\|q) = \Omega(p) + \Omega^*(\frac{w\odot\theta}{T}) - \langle\frac{w\odot\theta}{T}, p\rangle$. Since $\Omega^*(\frac{w\odot\theta}{T})$ does not depend on p, we can drop it and swap min–max order, arriving at the dual game (FPM). Its maximiser $p^\star(w)$ satisfies

$$\frac{w\odot\theta}{T} - \nabla\Omega(p^\star) + \lambda\big(p^\star - \nu\big) = 0,$$

and, using $\nabla\Omega(p) = -1 - \log p$, we recover the fixed-point equation stated in the proposition.

Because the inner optimization over p is *strongly concave* (strict convexity of Ω plus the quadratic term), $p^\star(w)$ is unique and differentiable. Eliminating p therefore leaves a smooth, strongly convex outer landscape in w, so standard solvers such as Adam or L-BFGS converge reliably, avoiding the numerical brittleness observed in the primal formulation.

Overview of Proposition 4.2. Having reduced learning to the outer problem

$$g(w) = \tfrac{\lambda}{2}\|p^\star(w) - \nu\|_2^2,$$

we now ask: *Does g admit strong convexity and Lipschitz-gradient constants sufficient for linear convergence of gradient descent.* The next proposition establishes strong convexity and smoothness constants that guarantee: (i) uniqueness of the global minimizer, (ii) bounded gradients, and (iii) linear convergence of gradient descent with an explicit rate depending solely on the condition number κ.

Proposition 5.1 (First-order geometry of the outer FPM objective). Let

$$g(w) = \frac{\lambda}{2}\,\big\|p^\star(w) - \nu\big\|_2^2, \qquad p^\star(w) =_{p\in\Delta_m} \Big\langle\tfrac{w\odot\theta}{T}, p\Big\rangle - \Omega(p),$$

with temperature $T > 0$, regularization weight $\lambda > 0$, and prior $\nu \in \Delta_m$. Set $\underline{\theta} := \min_j |\theta_j| > 0$ and $\overline{\theta} := \max_j |\theta_j|$. After eliminating the inner maximization, we obtain the following facts.

Lemma 1 (Strong convexity). For all x, y,

$$g(y) \;\geq\; g(x) + \langle\nabla g(x), y - x\rangle + \frac{\mu}{2}\,\|y-x\|_2^2, \qquad \mu = \frac{\lambda\,\underline{\theta}^2}{T^2}.$$

Thus g has a *unique* global minimizer; "flat-corner" stationary points cannot occur (Nesterov, 2004).

Proof. The inner problem is 1-strongly concave in p, so $p^\star(w)$ depends affinely on $\frac{w\odot\theta}{T}$. By the envelope theorem this transfers to strong convexity of g with modulus $\mu = \lambda\underline{\theta}^2/T^2$. □

Lemma 2 (Smoothness). For all x, y,

$$g(y) \;\leq\; g(x) + \langle\nabla g(x), y - x\rangle + \frac{L}{2}\,\|y-x\|_2^2, \qquad L = \frac{\lambda\,\overline{\theta}^2}{T^2}.$$

Consequently, any fixed stepsize $\eta \in (0, 2/L)$ yields monotone descent (Boyd and Vandenberghe, 2004b):

$$g(w_{t+1}) \;\leq\; g(w_t) - \frac{1}{2}\Big(\tfrac{1}{\eta} - \tfrac{L}{2}\Big)\|w_{t+1} - w_t\|_2^2,$$

preventing overshoot and sign-flip when $\eta < 2/L$.

Proof. Because $p^\star(w)$ is affine in $\frac{w\odot\theta}{T}$, the Hessian of g is bounded above by $L = \lambda\overline{\theta}^2/T^2$. □

Lemma 3 (Polyak–Łojasiewicz inequality). Combining Lemmas 1 and 2 gives

$$\frac{1}{2}\big\|\nabla g(w)\big\|_2^2 \;\geq\; \mu\big(g(w) - g^\star\big),$$

so a small gradient norm arises only when $g(w)$ is already near its minimum (Karimi et al., 2016).

Proof. Strong convexity implies $\langle \nabla g(w), w - w^\star\rangle \geq \mu\|w - w^\star\|_2^2$, while smoothness gives $\|\nabla g(w)\|_2 \leq L\|w - w^\star\|_2$. Multiplying and rearranging yields the inequality. □

Definition (Condition number).

$$\kappa := \frac{L}{\mu} = \Big(\frac{\overline{\theta}}{\underline{\theta}}\Big)^2.$$

Take-away.

1. **No spurious stationary points**: if $\nabla g(w) = 0$, then $w = w^\star$.
2. **Co-coercive gradient**: step norms remain bounded.
3. **Linear convergence**: gradient descent with $\eta \leq 1/L$ contracts at rate $1 - 1/\kappa$ (Bubeck, 2015).