Not All Instances Should Contribute Equally: Sparse Support Grounding via Unbalanced Transport for Heterogeneous MIL

Hojoon Kim
Korea University
hojoon@korea.ac.kr

Beomgu Kang
Korea University
bgkang1@korea.ac.kr

Hyunseok Seo*
Korea University
seoh@korea.ac.kr

Abstract

Multiple instance learning (MIL) learns from bag-level labels, but a bag's label is often supported by only a few hidden instances. In heterogeneous bags, these useful instances are rare, diverse, and surrounded by many weak or noisy observations. We propose Sparse Support Grounding, a MIL framework that learns instance influence directly. The key idea is simple: an instance should not matter just because it can be matched somewhere; it should matter when the match provides reliable label evidence. Our method forms a compact, bag-specific evidence support, compares instances with label-consistent supports, and uses unbalanced transport to estimate how much source mass each instance should retain. This retained mass becomes the strength of its grounding signal, so reliable evidence shapes training more than nuisance content, while the bag classifier still uses the full bag. Thus, transport is turned from a matching rule into a way of learning which instances should drive supervision. Experiments on synthetic MIL, whole-slide images, and clinical endomicroscopy show consistent gains, with the largest improvements when sparse diagnostic evidence is embedded in heterogeneous nuisance content.

1 Introduction

Many weakly supervised problems provide only bag-level labels, although the label may depend on a few unknown instances. Multiple instance learning (MIL) formalizes this setting [1–4] and is widely used in medical imaging, video analysis, and molecular prediction. The problem becomes especially difficult in heterogeneous bags, where useful evidence is rare and mixed with many weak or noisy instances. In this setting, a model should not only predict the bag label, but also learn which instances deserve the strongest learning signal.

Existing deep MIL methods improve bag prediction through attention aggregation, representative-region mining, prototype modeling, and inter-instance interaction modeling [4–10]. These methods mainly address how to summarize a bag or compare instances with representative patterns. We focus on a different question: after an instance is aligned to some evidence, how much should it matter? This distinction is important because a nuisance instance can still be matched somewhere, but it should not shape the training signal as strongly as truly label-relevant evidence.

We introduce *Sparse Support Grounding*, a MIL framework that treats instance contribution as a learnable quantity. The method builds a compact support from salient evidence in each bag and routes instances to label-consistent supports. It then uses entropic unbalanced optimal transport (UOT) [11–13] to turn instance–support compatibility into contribution strength. The transport plan decides where an instance aligns, while the retained source mass decides how strongly it shapes the grounding loss.

*Corresponding author.

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

This design gives bag-level supervision a sharper instance-level effect. Entmax forms sparse bag-specific supports, while bag prediction still uses the full-bag attention summary. Self-anchored same-class routing allows bags of the same class to share useful evidence without collapsing their supports into a single prototype. UOT then gives more influence to well-matched evidence and less influence to weakly matched nuisance instances.

We analyze this mechanism and validate it on controlled synthetic MIL, public whole-slide image, and clinical confocal laser endomicroscopy benchmarks. The largest gains appear when sparse evidence is embedded in heterogeneous nuisance content, matching the setting targeted by the proposed framework.

Contributions.

- We introduce an instance-influence view of heterogeneous MIL, where alignment and contribution strength are treated as separate learning problems.
- We propose *Sparse Support Grounding*, which converts bag-specific sparse supports and UOT retained mass into compatibility-weighted instance grounding.
- We show that UOT retained mass controls instance contribution, while self-anchored same-class routing preserves distinct evidence supports.
- We validate the framework on synthetic MIL, public WSI, and clinical CLE benchmarks, with the strongest gains under sparse heterogeneous evidence.

2 Related Work

MIL under heterogeneous evidence. Multiple instance learning (MIL) learns from bag-level labels when the label-relevant instances are unknown [2, 3]. Deep MIL methods improve bag prediction through attention pooling, representative-region mining, critical-instance modeling, and inter-instance interaction modeling [4–10]. These methods are effective at building predictive bag representations, but heterogeneous bags require an additional signal: the model must learn which instances should strongly influence training when useful evidence is rare or mixed with noisy instances. Sparse Support Grounding addresses this gap by learning not only a bag prediction, but also how much each instance should contribute to the grounding signal.

Bag-specific evidence supports. Prototype- and concept-based MIL methods compare instances with representative patterns learned across the dataset [14–16]. In contrast, our method builds a compact support separately for each bag, using the instances currently considered salient. This provides a bag-specific grounding target, allowing each instance to be compared with evidence from its own bag rather than only with global prototypes.

Sparse grounding. Sparse attention mechanisms such as sparsemax and entmax produce adaptive sparse weights [17, 18]. While sparsity is often used directly for bag prediction [19, 9, 10], we use it to define the support used by the grounding branch; the bag classifier still operates on the full-bag representation.

Transport-based influence. Unbalanced optimal transport (UOT) [20, 13, 21, 12, 11] relaxes the marginal constraints of balanced OT via KL penalties:

$$\min_{\pi \geq 0} \langle C, \pi \rangle + \rho_{\mathrm{src}} \, \mathrm{GKL}(\pi \mathbf{1} \| \boldsymbol{\mu}) + \rho_{\mathrm{tgt}} \, \mathrm{GKL}(\pi^{\top} \mathbf{1} \| \boldsymbol{\nu}) + \varepsilon \sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij} (\log \pi_{ij} - 1) \tag{1}$$

Here, i indexes flattened source-instance rows and j indexes target supports; in our discrete grounding problem below, those targets are the bag-specific proxies. Unlike balanced OT, which fixes each row sum $m_i = \sum_j \pi_{ij} = \mu_i$, UOT allows m_i to vary with instance–support compatibility. We use this retained mass as a learned contribution weight: well-aligned instances shape the grounding loss more than weak or noisy ones.

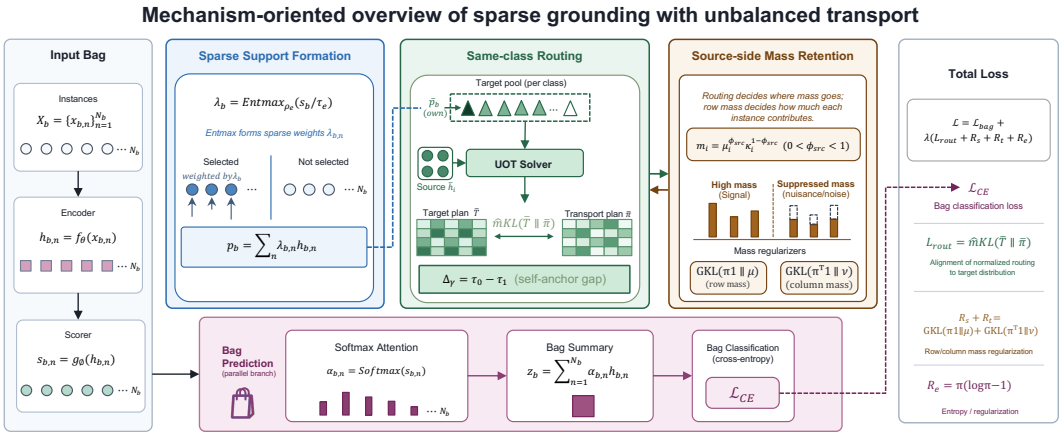


Figure 1: Overview of the proposed framework. Entmax constructs a sparse bag-specific proxy support, bag prediction uses the full-bag attention summary, and the grounding branch uses entropic UOT to separate routing from retained source mass.

CLE as a heterogeneous MIL setting. Confocal laser endomicroscopy (CLE) provides a practical heterogeneous MIL setting: a case may contain diagnostic frames together with blur, defocus, motion artifacts, blood contamination, and tissue-dependent appearance variation [22, 23]. Since only case-level labels are available, the model must emphasize reliable diagnostic frames while reducing the influence of weak frames. This makes CLE a natural testbed for Sparse Support Grounding.

3 Sparse Support Grounding for Heterogeneous MIL

3.1 Problem Setup

We consider a multiple instance learning dataset

$$\mathcal{D} = \{(X_b, c_b)\}_{b=1}^{B}, \qquad X_b = \{x_{b,n}\}_{n=1}^{N_b}, \qquad c_b \in \{1, \dots, C\}.$$

Here, B is the number of bags, X_b is the b-th bag, $x_{b,n}$ is its n-th instance, N_b is the number of instances in the bag, and c_b is the bag label among C classes.

In heterogeneous MIL, only a small subset of instances may truly support the bag label. The remaining instances can be weakly related, noisy, or irrelevant. We therefore need an instance-level signal that answers two questions at once: where an instance should align, and how strongly it should affect learning.

Sparse Support Grounding addresses this by using a bag-specific evidence support and the retained source mass of unbalanced transport. The routing distribution tells where an instance aligns. The retained mass tells how much that instance contributes.

3.2 Sparse Bag-Specific Evidence Support

An encoder f_θ maps each instance to an embedding

$$h_{b,n} = f_\theta(x_{b,n}) \in \mathbb{R}^d,$$

and a scorer g_ϕ assigns a salience score

$$s_{b,n} = g_\phi(h_{b,n}), \qquad s_b = (s_{b,1}, \dots, s_{b,N_b}) \in \mathbb{R}^{N_b}.$$

Here, $h_{b,n}$ is the instance embedding, $s_{b,n}$ is its salience score, and s_b collects all salience scores in bag b.

Bag prediction uses the full-bag attention summary

$$\alpha_{b,n} = \frac{\exp(s_{b,n})}{\sum_{m=1}^{N_b} \exp(s_{b,m})}, \qquad z_b = \sum_{n=1}^{N_b} \alpha_{b,n} h_{b,n}.$$

Here, $\alpha_{b,n}$ is the attention weight of instance n, and z_b is the bag representation used for classification.

The grounding branch uses the same scores to build a compact evidence support. We compute sparse entmax weights

$$\lambda_b = \mathrm{Entmax}_{\rho_e}(s_b/\tau_e),$$

where $\lambda_b = (\lambda_{b,1}, \ldots, \lambda_{b,N_b})$, ρ_e controls sparsity, and τ_e is a temperature. The bag-specific proxy is

$$p_b = \sum_{n=1}^{N_b} \lambda_{b,n} h_{b,n}, \qquad \bar{p}_b = \frac{p_b}{\|p_b\|_2}. \tag{2}$$

Here, p_b summarizes the salient evidence in bag b, and $\bar{p}_b$ is its normalized proxy. We understand this normalization on $\|p_b\|_2 > 0$.

For a mini-batch of M bags, we collect the proxies as

$$\mathcal{P} = \{\bar{p}_j\}_{j=1}^{M},$$

where each proxy $\bar{p}_j$ inherits the label $c(j)$ of its source bag. We flatten all mini-batch instances into rows $i \in \{1, \ldots, N\}$, with $N = \sum_{b=1}^{M} N_b$, and write $\bar{h}_i = h_i/\|h_i\|_2$ for the normalized embedding and $o(i) \in \{1, \ldots, M\}$ for the source-bag index of instance i. We also write $N_{o(i)}$ for the number of instances in the source bag of row i, so $N_{o(i)} = N_b$ when $o(i) = b$. Here again, $\|h_i\|_2 > 0$ is assumed.

3.3 Self-Anchored Same-Class Target

For each instance i, let c_i be the label of its source bag. The set of same-class proxies is

$$Q_{c_i} = \{j : c(j) = c_i\}.$$

We define a self-anchored same-class target

$$T_{ij} = \mathbf{1}\{j = o(i)\} + \gamma \mathbf{1}\{j \in Q_{c_i},\, j \neq o(i)\}, \qquad \gamma \in [0, 1],$$

and normalize it as

$$\bar{T}_{ij} = \frac{T_{ij}}{1 + (K_{c_i} - 1)\,\gamma},$$

where $K_{c_i} := |Q_{c_i}|$ is the number of same-class proxies. Here, T_{ij} is the target weight from instance i to proxy j, and γ controls how much same-class sharing is allowed. When $\gamma = 0$, an instance is grounded only to its own-bag proxy. When $\gamma = 1$, the target is uniform over same-class proxies. Intermediate values preserve the own-bag anchor while allowing same-class evidence sharing. The exact row-wise normalization formulas are given in Appendix C.1.

3.4 Unbalanced Transport as Instance Contribution

We align normalized instances $\{\bar{h}_i\}_{i=1}^{N}$ to the proxy pool $\mathcal{P}$ using cosine dissimilarity

$$C_{ij} = 1 - \bar{h}_i^{\top} \bar{p}_j. \tag{3}$$

An entropic unbalanced optimal transport solver returns a transport plan

$$\pi \in \mathbb{R}_{+}^{N \times M}.$$

From π, we define

$$m_i = \sum_{j=1}^{M} \pi_{ij}, \qquad \bar{\pi}_{ij} = \frac{\pi_{ij}}{m_i}. \tag{4}$$

Here, $\bar{\pi}_i$ is the routing distribution of instance i, and m_i is its retained source mass. The row normalization is understood for $m_i > 0$.

This separation is the main mechanism. The normalized plan $\bar{\pi}_i$ decides where the instance aligns. The retained mass m_i decides how strongly the instance contributes to the grounding loss.

To see why m_i reflects instance–support compatibility, define

$$\varphi_{\mathrm{src}} = \frac{\rho_{\mathrm{src}}}{\rho_{\mathrm{src}} + \varepsilon} \in (0,1), \qquad K_{ij} = \exp(-C_{ij}/\varepsilon), \qquad \kappa_i = \sum_{j=1}^{M} K_{ij} v_j.$$

Here, ε controls the smoothness of the transport plan, ρ_{src} controls how freely source mass can change, K_{ij} is the transport kernel, and v_j is the target-side Sinkhorn scaling variable, equivalently the target-side dual variable in the factorized transport plan $\pi_{ij} = u_i K_{ij} v_j$. Thus κ_i is the effective compatibility between instance i and the proxy pool after target-side reweighting. Under the UOT source update,

$$m_i = \mu_i^{\varphi_{\mathrm{src}}} \kappa_i^{1-\varphi_{\mathrm{src}}},$$

where μ_i is the reference source mass. With a uniform source prior, larger κ_i gives larger m_i. Thus, instances that better match the learned supports keep more mass, while weakly matched instances keep less. Appendix B.2 gives the full derivation and the comparison with balanced OT, which cannot express this behavior because its source marginal constraint fixes each row sum in advance.

We convert retained mass into a within-bag contribution weight. Define

$$r_b := \sum_{k:o(k)=b} \exp(m_k), \qquad \hat{m}_i := \frac{\exp(m_i)}{M\, r_{o(i)}}. \tag{5}$$

Here, r_b is the bag-wise sum of exponentiated retained masses, so $\sum_{i:o(i)=b} \hat{m}_i = 1/M$ for every bag b. Thus, $\hat{m}_i$ gives larger weight to instances with larger retained mass while keeping the mini-batch contribution balanced across bags. Let

$$\bar{T}_i = (\bar{T}_{ij})_{j=1}^{M}, \qquad \bar{\pi}_i = (\bar{\pi}_{ij})_{j=1}^{M}.$$

The routing loss is

$$\mathcal{L}_{\mathrm{rout}} = \sum_{i=1}^{N} \hat{m}_i \mathrm{KL}(\bar{T}_i \| \bar{\pi}_i). \tag{6}$$

3.5 Training Objective

The bag classifier predicts $\hat{y}_b = q_\psi(z_b)$. The bag-level loss, grounding loss, and final objective are

$$\mathcal{L}_{\mathrm{bag}} = \frac{1}{M} \sum_{b=1}^{M} \mathrm{CE}(\hat{y}_b, c_b), \qquad \mathcal{L}_{\mathrm{grnd}} = \mathcal{L}_{\mathrm{rout}} + \mathcal{R}_s + \mathcal{R}_t + \mathcal{R}_e, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{bag}} + \lambda_{\mathrm{grnd}} \mathcal{L}_{\mathrm{grnd}},$$

$$\mathcal{R}_s = \mathrm{GKL}(\pi \mathbf{1} \| \boldsymbol{\mu}), \quad \mathcal{R}_t = \mathrm{GKL}(\pi^\top \mathbf{1} \| \boldsymbol{\nu}), \quad \mathcal{R}_e = \sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij} (\log \pi_{ij} - 1).$$

Here, $\mathrm{GKL}(p\|q) = \sum_k [p_k \log(p_k/q_k) - p_k + q_k]$. In our discrete instance-to-proxy grounding problem, $\boldsymbol{\mu} \in \mathbb{R}_+^N$ and $\boldsymbol{\nu} \in \mathbb{R}_+^M$ are the discrete source-instance and target-proxy mass priors, matching the generic UOT measures in Eq. (1). The UOT solver uses ρ_{src}, ρ_{tgt}, and ε internally to compute the transport plan π; these coefficients are absorbed into the plan and do not appear again in the surrogate loss. The terms $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ are instead evaluated on the resulting plan π as independent surrogate loss components: they supply gradient signals to the encoder and scorer, regularizing instance-side mass, proxy-side mass, and plan smoothness, respectively, without re-enforcing the transport marginals. The bag classifier always uses the full-bag representation z_b, while the grounding branch adds an instance-level signal whose strength is controlled by λ_{grnd}.

4 From Matching to Instance Influence

Sparse Support Grounding turns transport into an instance-influence mechanism. In heterogeneous MIL, every instance can be matched to some support, but not every match should affect training equally. Our key idea is to separate these two roles. The normalized transport row tells where an instance aligns, while the retained row mass tells how strongly that instance contributes. This separation is built from three parts: sparse evidence supports, self-anchored same-class routing, and UOT retained mass.

4.1 Sparse Supports Define the Evidence

Recall the bag-specific support proxy $\bar{p}_b$ defined in Eq. (2). It is constructed from entmax weights over the instances in bag b, so only instances assigned nonzero sparse weights contribute to the proxy.

Because entmax can assign exact zero weights, p_b summarizes selected evidence rather than averaging all instances or forcing every instance to match a single global prototype. This gives each bag a compact bag-specific evidence support.

For $1 < \rho_e < 2$, the entmax-weighted proxy map $\Theta \mapsto p_b(\Theta)$ is globally C^1. When $\|p_b(\Theta)\|_2 > 0$, the normalized proxy map $\Theta \mapsto \bar{p}_b(\Theta)$ is also C^1. This regularity enables smoothness on compact trajectory neighborhoods without fixed-support segmentation; see Appendices B.1 and B.3.

4.2 Self-Anchored Routing Defines Where Instances Align

Recall the normalized self-anchored target $\bar{T}_i$ defined above. For class c, let $K_c = |Q_c|$ denote the number of same-class proxies. The row-wise normalization of the self-anchored target can then be written as

$$\bar{T}_{ij} = \tau_0 \mathbf{1}\{j = o(i)\} + \tau_1 \mathbf{1}\{j \in Q_{c_i},\, j \neq o(i)\},$$

where

$$\tau_0 = \frac{1}{1 + (K_{c_i} - 1)\gamma}, \qquad \tau_1 = \frac{\gamma}{1 + (K_{c_i} - 1)\gamma}, \qquad \gamma \in [0, 1].$$

The own-bag advantage is

$$\Delta_\gamma = \tau_0 - \tau_1 = \frac{1 - \gamma}{1 + (K_{c_i} - 1)\gamma}.$$

When $\gamma < 1$, $\Delta_\gamma > 0$, so each instance is pulled most strongly toward its own-bag support while still sharing evidence with same-class supports. This preserves label consistency without collapsing all same-class bags into one prototype. Appendix C.1 analyzes the row-wise normalization and resulting anti-collapse field; trajectory-level and curvature results appear in Appendices D.3 and D.1, and Appendix D.4 establishes the proxy-level anti-collapse field under parameter-level gradient descent.

4.3 UOT Retained Mass Defines How Much Instances Contribute

Recall from Eq. (4) that each UOT row is decomposed into a normalized routing distribution $\bar{\pi}_i$ and a retained source mass m_i. The routing distribution decides where instance i aligns, whereas the retained mass decides how strongly it affects the grounding objective.

Using the cosine cost C_{ij} in Eq. (3), the Gibbs kernel and effective compatibility of instance i are

$$K_{ij} = \exp(-C_{ij}/\varepsilon), \qquad \kappa_i = \sum_j K_{ij} v_j.$$

The UOT source update gives

$$m_i = \mu_i^{\varphi_{\mathrm{src}}} \kappa_i^{1-\varphi_{\mathrm{src}}}, \qquad \varphi_{\mathrm{src}} = \frac{\rho_{\mathrm{src}}}{\rho_{\mathrm{src}} + \varepsilon} \in (0, 1),$$

where μ_i is the source prior, ρ_{src} controls how much source mass can change, and v_j is again the target-side Sinkhorn scaling variable in $\pi_{ij} = u_i K_{ij} v_j$. With uniform μ_i, a larger κ_i gives a larger m_i. Well-matched evidence therefore keeps more mass and receives a stronger grounding signal, while weakly matched nuisance instances keep less.

Through the contribution weight $\hat{m}_i$ in Eq. (5), retained mass directly controls the instance-level strength of the routing loss in Eq. (6). Balanced OT fixes $m_i = \mu_i$, so it can decide only where fixed mass is routed; UOT also decides how much mass to retain. This is what turns transport from matching into instance contribution. The full derivation and balanced-OT comparison are given in Appendix B.2.

4.4 Summary

Entmax selects what evidence each bag provides. Self-anchored routing decides where each instance aligns. UOT retained mass decides how strongly each instance shapes the grounding loss. Together,

these parts convert bag-level supervision into instance-level learning signals that emphasize reliable evidence and reduce the effect of nuisance instances. The mini-batch extension of the descent and anti-collapse behavior is analyzed in Appendices E.2 and E.3, and an exact-gradient convergence guarantee on compact trajectory neighborhoods is established in Appendix E.1.

5 Evaluation on Sparse and Heterogeneous MIL

5.1 Experimental Protocol

We evaluate whether our support–routing–contribution framework improves bag-level prediction and instance-level evidence discovery under sparse heterogeneous evidence. We use synthetic MIL benchmarks from MNIST, FMNIST, and CIFAR10 [24–26]. Each bag has 100 instances, with 5% signal instances from classes 0–4 and 95% nuisance instances from classes 5–9.

All baselines use the same multiclass bag-classification setup, including identical splits, features, training budgets, and validation protocols. For RGMIL [27], OTSurv [28], TransMIL [7], and DTFD-MIL [8], we replace only the task-specific head with a C-way cross-entropy head and keep the original aggregation, prototype, representative, transport, or distillation design. ABMIL denotes the gated-attention model of Ilse et al. [4].

The selected baselines cover the key MIL design families most relevant to our hypothesis—attention aggregation, prototype-based grounding, representative-region mining, and transport-based matching—and therefore directly test the added value of the proposed support–routing–mass mechanism.

Instance-level accuracy is evaluated by passing each held-out signal instance (classes 0–4) individually through the frozen encoder and bag classifier as a singleton bag, yielding a C-class prediction via argmax; nuisance instances (classes 5–9) are excluded from this metric. For CLE, the same singleton-bag protocol is applied to the quality-annotated held-out frames using the bag-supervised encoder, without instance-level fine-tuning.

5.2 Synthetic Benchmarks

Table 1 reports bag-level and instance-level accuracy on MNIST, FMNIST, and CIFAR10. The proposed method is consistently strong across both metrics. On MNIST and FMNIST, where bag accuracy is already near saturation, our method remains competitive while improving instance-level grounding over the ABMIL backbone. This indicates that the grounding branch improves evidence discovery without weakening full-bag prediction.

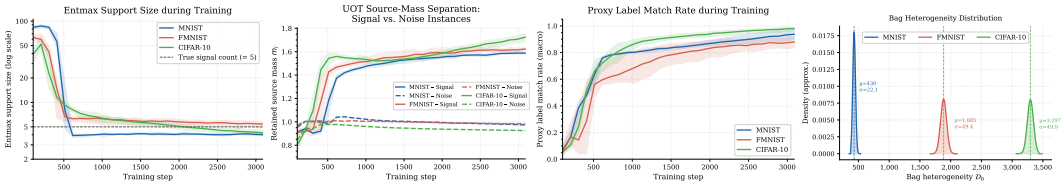
Table 1: Performance comparison across benchmark datasets with increasing visual complexity. MNIST, FMNIST, and CIFAR-10 are used as low-, moderate-, and high-heterogeneity proxies, respectively, reflecting their increasing intra-class variation, background complexity, and instance-level visual diversity.

	MNIST		**FMNIST**		**CIFAR-10**	
	Low visual heterogeneity		*Moderate visual heterogeneity*		*High visual heterogeneity*	
Model	Bag	Inst	Bag	Inst	Bag	Inst
ABMIL	98.71 ± 0.36	88.26 ± 0.65	93.24 ± 0.88	81.90 ± 0.12	84.80 ± 0.78	73.53 ± 1.70
CLAM-MB	$\mathbf{99.22 \pm 0.14}$	88.73 ± 0.70	93.34 ± 1.05	77.76 ± 2.08	84.14 ± 0.65	72.44 ± 1.08
DSMIL	98.31 ± 0.59	83.42 ± 3.25	88.88 ± 0.77	77.81 ± 1.11	76.84 ± 2.05	58.08 ± 3.83
OTSurv	77.24 ± 7.47	$\mathbf{91.83 \pm 1.90}$	67.22 ± 3.26	69.65 ± 4.85	49.80 ± 2.23	66.96 ± 0.73
RGMIL	98.69 ± 0.16	88.38 ± 0.47	91.62 ± 0.76	81.20 ± 0.89	80.24 ± 1.98	66.40 ± 2.02
TPMIL	98.24 ± 0.25	87.08 ± 1.00	88.30 ± 4.95	76.47 ± 5.60	83.36 ± 1.94	60.25 ± 3.65
Ours (UOT)	98.96 ± 0.34	90.52 ± 1.10	$\mathbf{93.88 \pm 1.38}$	$\mathbf{82.84 \pm 0.90}$	$\mathbf{88.44 \pm 0.96}$	$\mathbf{75.96 \pm 0.66}$

The largest gains appear on CIFAR10, the most visually heterogeneous benchmark. Our method improves bag accuracy from 84.80% with ABMIL and 84.14% with CLAM-MB to 88.44%, and instance accuracy from 73.53% and 72.44% to 75.96%. Because all benchmarks use the same bag length and 5% signal ratio, this result shows that the method is most useful when rare evidence is mixed with diverse nuisance content.

Table 2: Grounding-mechanism ablation. Check marks indicate active components. UOT adaptive mass denotes the relaxed UOT marginals together with their marginal and entropy loss terms ($\mathcal{R}_s$, $\mathcal{R}_t$, $\mathcal{R}_e$). Removing UOT-based contribution control hurts instance grounding most on CIFAR10, the most heterogeneous setting.

Variant	Sparse support	Own-bag anchor	Same-class sharing	UOT adaptive mass	Mass-weighted routing	MNIST Inst.	FMNIST Inst.	CIFAR10 Inst.	Average Inst.
Full model, Ours	✓	✓	✓	✓	✓	**90.52**	82.84	**75.96**	**83.11**
Self-only target	✓	✓	–	✓	✓	89.63	**83.17**	73.58	82.13
Balanced OT	✓	✓	✓	–	–	89.01	82.48	69.34	80.28
Align-only	✓	✓	✓	–	✓	86.87	82.56	67.04	78.82



Entmax support | Source mass | Proxy-label match | Bag heterogeneity

Figure 2: Synthetic mechanism checks for support formation, mass retention, routing behavior, and empirical domain-level bag heterogeneity across the synthetic benchmarks. The rightmost panel reports a simple bag-level heterogeneity score; its definition is deferred to Appendix F.2.

The results follow the proposed design. Sparse supports define compact evidence, routing connects instances to label-consistent supports, and UOT retained mass controls instance contribution. Thus, reliable evidence receives a stronger grounding signal, while weak nuisance instances have less effect.

5.3 Mechanism-Isolation Ablations

Table 2 isolates the components that affect instance-level grounding, with instance accuracy as the main metric. UOT-based contribution control is most critical: replacing UOT with balanced OT lowers CIFAR10 instance accuracy from 75.96% to 69.34%, and the align-only variant drops to 67.04%. Thus, alignment alone is not enough; adaptive mass is needed to suppress weak or nuisance instances.

Same-class sharing also matters: removing it while keeping own-bag evidence reduces accuracy from 75.96% to 73.58%, showing that controlled sharing helps under diverse visual evidence. Additional sweeps appear in Appendix F.3 and Appendix F.4.

Figure 2 verifies the proposed mechanism: entmax yields compact adaptive supports, UOT assigns higher retained mass to signal instances, routing becomes class-consistent, and CIFAR10 exhibits stronger bag-level dispersion. During approximately the first 500–600 training steps, the initially broad support becomes more compact, signal and nuisance retained masses separate, and proxy–label matching improves; the trends then remain stable. Because the support is recomputed throughout training, the model can refine initially unreliable salience estimates rather than fixing the initial support. This is an empirical observation, not a general guarantee of optimization stability. Together with Table 2, these results support our design principle that heterogeneous MIL benefits from separating support formation, routing, and instance contribution control.

5.4 WSI Bag-Level Evaluation

We further evaluate the proposed method on TCGA Lung [29, 30] and BRACS [31], two slide-level WSI benchmarks with different diagnostic heterogeneity. In WSI, a slide label supervises many diverse patches, while label-relevant evidence may occupy only a small subset of regions. We report accuracy and macro-F1 at a fixed operating point, together with macro one-vs-rest AUC.

Table 3 shows that the proposed method achieves the highest mean performance on both datasets, including against TransMIL and DTFD-MIL. The gains are modest on TCGA Lung: relative to RGMIL, accuracy and F1 improve by 1.05 and 1.06 percentage points, respectively, while AUC is comparable (+0.07 points) and the reported variability ranges overlap. The gains become larger on BRACS, where fine-grained diagnostic heterogeneity is higher. This supports our claim that sparse,

Table 3: Slide-level performance on TCGA Lung and BRACS in terms of accuracy, AUC, and F1 (mean ± std over 4 folds for TCGA Lung and 5 random seeds for BRACS).

	TCGA Lung			**BRACS**		
	lower fine-grained diagnostic-spectrum heterogeneity			*higher fine-grained diagnostic-spectrum heterogeneity*		
Model	Acc	AUC	F1	Acc	AUC	F1
ABMIL	85.73 ± 0.84	93.04 ± 0.99	85.71 ± 0.81	59.30 ± 1.84	77.98 ± 1.37	57.10 ± 3.74
CLAM-MB	83.57 ± 1.54	93.48 ± 1.28	83.41 ± 1.61	63.49 ± 2.68	79.54 ± 2.13	59.28 ± 3.00
DSMIL	85.50 ± 1.52	93.18 ± 1.43	85.37 ± 1.62	53.93 ± 1.95	73.10 ± 2.90	52.19 ± 4.22
OTSurv	83.35 ± 2.35	92.77 ± 0.86	83.35 ± 2.35	55.81 ± 4.27	73.01 ± 3.01	51.97 ± 4.38
RGMIL	86.89 ± 1.37	94.34 ± 0.83	86.86 ± 1.39	56.98 ± 2.75	76.29 ± 1.33	56.31 ± 2.14
TPMIL	83.49 ± 2.29	92.07 ± 1.74	83.24 ± 2.59	60.23 ± 2.37	79.82 ± 1.45	55.31 ± 3.74
TransMIL	86.60 ± 4.74	93.38 ± 3.08	86.62 ± 4.83	57.78 ± 5.28	77.63 ± 3.70	55.35 ± 8.33
DTFD-MIL	84.74 ± 3.37	90.12 ± 3.23	84.72 ± 3.40	55.84 ± 3.73	78.42 ± 2.73	49.79 ± 5.26
Ours (UOT)	**87.94 ± 1.56**	**94.41 ± 0.65**	**87.92 ± 1.66**	**64.96 ± 1.39**	**80.47 ± 1.13**	**60.92 ± 1.07**

bag-specific grounding is broadly effective and becomes especially beneficial when label-relevant patches are diverse and unevenly distributed across a WSI.

The UOT branch adds training-time computation but is removed at inference, when prediction uses only the standard full-bag attention branch. Relative to the same attention-pooling backbone with the UOT branch disabled, it increases training time per epoch by approximately 37% on CIFAR10 and 45% on TCGA Lung, while increasing peak GPU memory by only approximately 1–2%. Full measurement and scaling details are provided in Appendix F.1.

5.5 CLE Instance-Critical Evaluation

In CLE, we follow the IRB-approved, de-identified clinical cohort and protocol of Byun et al. [23], which benchmark biopsy-confirmed brain-tumor cases from a prospective multicenter CLE study against frozen-section diagnosis. Bag accuracy is measured by 5-fold bag-level cross-validation, and instance accuracy on a separate quality-annotated held-out set without CLE-specific threshold tuning, following the singleton-bag protocol described in Section 5.1.

Table 4: CLE case-level and held-out frame-level accuracy (mean ± std over 5 folds).

Metric	ABMIL	CLAM-MB	DSMIL	OTSurv	RGMIL	TPMIL	Ours (UOT)
Bag	71.75±4.44	71.97±3.49	71.28±1.65	73.97±4.46	71.96±4.59	70.85±5.65	**77.92±5.47**
Instance	50.42±6.52	53.33±5.57	52.36±2.59	61.82±3.51	58.06±5.17	43.64±3.45	**67.39±3.96**

Table 4 shows that our method delivers the clearest gain at the frame level while also improving case-level accuracy. Across datasets, the empirical pattern is consistent with the proposed role separation, while also indicating that support formation should be interpreted as a heterogeneity-aware inductive bias rather than a universally dominant sparse choice. Broader sweeps for γ and transport hyperparameters are deferred to Appendices F.3 and F.4.

6 Conclusion

Heterogeneous multiple instance learning requires more than predicting a bag label from sparse evidence. The model also needs to learn which instances should strongly influence training. We addressed this problem with *Sparse Support Grounding*, which separates instance-level grounding into three roles: building a compact evidence support, routing instances to label-consistent supports, and using retained source mass to control instance contribution. In this view, UOT is not only a matching tool; its retained mass becomes the learned influence of each instance.

This role separation leads to consistent gains across synthetic benchmarks, whole-slide images, and clinical endomicroscopy. The improvements are strongest when useful evidence is rare and mixed with heterogeneous nuisance content. On the CLE benchmark, our method improves frame-level accuracy from 61.82% to 67.39% over the next best baseline, without frame-level supervision. Sparse support is designed for heterogeneous bags, where compact evidence targets are intended to keep the model focused on reliable instances.

Overall, the support–routing–mass view provides a simple design principle for weakly supervised learning: decide what evidence to compare against, where each instance aligns, and how strongly each instance should shape the learning signal.

7 Limitations

Sparse Support Grounding remains subject to the fundamental identifiability limits of bag-level MIL: without instance-level supervision, retained UOT mass should be interpreted as a model-dependent relative influence score, not as a calibrated instance label or causal evidence measure. The bag-specific proxy is also an approximation to latent evidence. When label-relevant instances are very rare, ambiguous, or visually close to nuisance variation, the proxy may capture only partial or biased support. Because proxies are constructed from instances available in the current mini-batch, the support pool is local and may underrepresent rare intra-class modes in highly diverse datasets. Finally, although we evaluate synthetic MIL, public WSI benchmarks, and clinical CLE data, broader validation across institutions, acquisition protocols, and label regimes is needed to establish generalization.

Acknowledgments

This work was supported, in part, by grants from the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean government (MSIT): the Information Technology Research Center (ITRC) program (IITP-2026-RS-2024-00436857) and the Artificial Intelligence Graduate School Program (Korea University) (No. RS-2019-II190079). This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Development of Unified Reasoning Technology Mimicking Human Cognition for Hierarchical Understanding and Unbounded Problem Solving (No. RS-2026-25522672). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2026-RS-2024-00338025) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant No. RS-2025-23872969).

References

[1] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3.

[2] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010. doi: 10.1017/S026988890999035X.

[3] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. doi: 10.1016/j.patcog.2017.10.009.

[4] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018. URL `https://proceedings.mlr.press/v80/ilse18a.html`.

[5] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

[6] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2021. doi: 10.1109/CVPR46437.2021.01409.

[7] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

[8] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18780–18790, 2022. doi: 10.1109/CVPR52688.2022.01824.

[9] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4055–4064, 2023. doi: 10.1109/ICCV51070.2023.00377.

[10] Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *Computer Vision – ECCV 2024*, volume 15111 of *Lecture Notes in Computer Science*, pages 125–143, Cham, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-73668-1_8.

[11] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314): 2563–2609, 2018. doi: 10.1090/mcom/3303.

[12] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/a9eb812238f753132652ae09963a05e9-Abstract.html.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

[14] Litao Yang, Deval Mehta, Sidong Liu, Dwarikanath Mahapatra, Antonio Di Ieva, and Zongyuan Ge. Trainable prototype enhanced multiple instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 1655–1665. PMLR, 2024. URL https://proceedings.mlr.press/v227/yang24d.html.

[15] Susu Sun, Dominique van Midden, Geert Litjens, and Christian F. Baumgartner. Prototype-based multiple instance learning for gigapixel whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15973 of *Lecture Notes in Computer Science*, pages 507–517, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-032-05185-1_49.

[16] Andrew H. Song, Richard J. Chen, Tong Ding, Drew F. K. Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11566–11578, 2024. doi: 10.1109/CVPR52733.2024.01099.

[17] André Martins and Ramón Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623. PMLR, 2016. URL https://proceedings.mlr.press/v48/martins16.html.

[18] Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1146. URL https://aclanthology.org/P19-1146/.

[19] Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4277–4285, 2015. doi: 10.1109/CVPR.2015.7299056.

[20] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

[21] Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

[22] Evgenii Belykh, Eric J. Miller, Alessandro Carotenuto, Arpan A. Patel, Claudio Cavallo, Nikolay L. Martirosyan, Debbie R. Healey, Vadim A. Byvaltsev, Adrienne C. Scheck, Michael T. Lawton, Jennifer M. Eschbacher, Peter Nakaji, and Mark C. Preul. Progress in confocal laser endomicroscopy for neurosurgery and technical nuances for brain tumor imaging with fluorescein. *Frontiers in Oncology*, 9:554, 2019. doi: 10.3389/fonc.2019.00554.

[23] Yoon Hwan Byun, Hyunseok Seo, Jae-Kyung Won, Boram Lee, Duk Hyun Hong, Sun Mo Nam, Jong Ha Hwang, Min-Sung Kim, Yong-Hwy Kim, Jang Hun Kim, Mi Ok Yu, Kyung-Jae Park, HoJoon Kim, Sunit Das, Doo-Sik Kong, Chul-Kee Park, and Shin-Hyuk Kang. AI augmented confocal laser endomicroscopy for rapid intraoperative diagnosis of brain tumors. *npj Digital Medicine*, 9(1):471, 2026. doi: 10.1038/s41746-026-02651-0.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[25] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747.

[26] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

[27] Zhaolong Du, Shasha Mao, Yimeng Zhang, Shuiping Gou, Licheng Jiao, and Lin Xiong. RGMIL: Guide your multiple-instance learning model with regressor. In *Advances in Neural Information Processing Systems*, volume 36, pages 35606–35618, 2023. doi: 10.52202/075280-1547. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/6feb9b30798abcfae937760d183605e1-Abstract-Conference.html`.

[28] Qin Ren, Yifan Wang, Ruogu Fang, Haibin Ling, and Chenyu You. OTSurv: A novel multiple instance learning framework for survival prediction with heterogeneity-aware optimal transport. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15974 of *Lecture Notes in Computer Science*, pages 439–449, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-032-05182-0_43.

[29] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014. doi: 10.1038/nature13385.

[30] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012. doi: 10.1038/nature11404.

[31] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images. *Database*, 2022:baac093, 2022. doi: 10.1093/database/baac093.

[32] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983. doi: 10.1214/aos/1176346079.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

A Appendix Roadmap and Scope

A.1 Scope and standing assumptions

Parameter-space regularity claims in the main text and appendix assume globally C^1 network maps; smoothness and descent additionally require Assumptions 1 and 2 on $\mathcal{K}_{\mathrm{traj}}$. Transport priors are fixed and strictly positive. Unless explicitly frozen or stopped, transport derivatives use the exact coupled UOT solution. Descent guarantees use exact gradients and do not directly cover the ReLU or stop-gradient implementation. The proxy is defined as the entmax-weighted sum $p_b = \sum_{n=1}^{N_b} \lambda_{b,n} h_{b,n}$, which is globally C^1 on parameter space. All differential statements are therefore interpreted on the compact trajectory neighborhood $\mathcal{K}_{\mathrm{traj}}$ without fixed-support segmentation. The appendix establishes: (i) globally C^1 proxy regularity and L-smoothness on $\mathcal{K}_{\mathrm{traj}}$ (Appendices B.1 and B.3); (ii) the frozen-dual anti-collapse identity at symmetric same-class collapse states (Appendix C.1); (iii) the exact coupled anti-collapse identity under the weighted proxy pullback (Appendix C.2); (iv) Gauss–Newton curvature stability under bounded imbalance, with a full-Hessian extension only under an additional remainder assumption (Appendix D.2); (v) trajectory-level non-persistence of symmetric same-class proxy collapse at non-degenerate collapse states, with same-class diversity exceeding a fixed threshold at all but finitely many such iterations (Appendix D.3); and (vi) probabilistic mini-batch anti-collapse plus mini-batch descent under explicit bias and variance control (Appendices E.2 and E.3). Collectively, these results provide local full-batch and mini-batch controls on compact trajectory neighborhoods under the stated assumptions. They are not intended as claims about arbitrary global optimization behavior outside those neighborhoods, hard top-k equivalence, or unconditional mini-batch stationarity.

A.2 Main–appendix correspondence

The main text now presents the mechanism in four linked pieces. The sparse-support discussion points to Appendices B.1 and B.3, which establish the weighted proxy's global C^1 regularity and compact-neighborhood L-smoothness. The self-anchored routing discussion is grounded by the frozen-dual anti-collapse baseline in Appendix C.1 and the exact coupled refinement in Appendix C.2. The retained-mass discussion in Section 4.3 uses the row-mass identity and the balanced-OT comparison from Appendix B.2. Finally, the main-text summary sentence on compact-neighborhood convergence and mini-batch behavior is justified by Appendix E.1 together with Appendices E.2 and E.3, culminating in Corollary 5.

B Mathematical Prerequisites

B.1 Regularity of the entmax-weighted proxy

We record the regularity facts needed for the sparse proxy map under entmax-weighted aggregation. For each bag b,

$$p_b(\Theta) = \sum_{n=1}^{N_b} \lambda_{b,n}(\Theta) h_{b,n}(\Theta),$$
$$\lambda_b(\Theta) = \mathrm{Entmax}_{\rho_e}(\mathbf{s}_b(\Theta)/\tau_e),$$
$$\mathbf{s}_b(\Theta) := (s_{b,n}(\Theta))_{n=1}^{N_b}, \qquad \bar{p}_b = \frac{p_b}{\|p_b\|_2}.$$

The nonzero entmax weights describe the sparse active set, but the proxy itself is the weighted sum above rather than a uniform mean over active indices.

Lemma 1 (Entmax Jacobian bound)**.** *Let $1 < \rho_e < 2$, let $\lambda = \mathrm{Entmax}_{\rho_e}(z/\tau_e)$, and let $J_{\mathrm{ent}} := D_z \,\mathrm{Entmax}_{\rho_e}(z/\tau_e)$. Then J_{ent} is symmetric positive semidefinite and, for every $v \in \mathbb{R}^{N_b}$,*

$$v^\top J_{\mathrm{ent}} v = \frac{1}{\tau_e}\left(\sum_n \alpha_n v_n^2 - \frac{\left(\sum_n \alpha_n v_n\right)^2}{Z}\right) \le \frac{1}{\tau_e}\|v\|_2^2,$$

where $\alpha_n := \lambda_n^{2-\rho_e} \in [0,1]$ and $Z := \sum_k \alpha_k$. Hence

$$\|J_{\mathrm{ent}}\|_{\mathrm{op}} \le \frac{1}{\tau_e}.$$

Proof. The stated quadratic form is the standard closed-form Jacobian identity for entmax on a fixed ambient dimension. The second term is nonnegative, so

$$v^\top J_{\mathrm{ent}} v \le \frac{1}{\tau_e}\sum_n \alpha_n v_n^2 \le \frac{1}{\tau_e}\|v\|_2^2,$$

since $0 \le \alpha_n \le 1$. Symmetry and positive semidefiniteness are immediate from the same representation. □

Proposition 1 (Global C^1 regularity of the weighted proxy)**.** *Let $1 < \rho_e < 2$ and assume that $h_{b,n}$ and $s_{b,n}$ are globally C^1 in Θ. The map $\Theta \mapsto p_b(\Theta) = \sum_n \lambda_{b,n}(\Theta) h_{b,n}(\Theta)$ is globally C^1, with differential*

$$Dp_b[\delta\Theta] = \sum_{n=1}^{N_b}\sum_{m=1}^{N_b} (J_{\mathrm{ent},b})_{nm}\, h_{b,n}\, (Ds_{b,m})^\top \delta\Theta + \sum_{n=1}^{N_b} \lambda_{b,n}\, Dh_{b,n}[\delta\Theta],$$

where $J_{\mathrm{ent},b} := D\,\mathrm{Entmax}_{\rho_e}(\mathbf{s}_b/\tau_e)$. If $\|p_b(\Theta)\|_2 \ge \nu_p > 0$, then

$$\bar{p}_b(\Theta) = \frac{p_b(\Theta)}{\|p_b(\Theta)\|_2}$$

is also globally C^1, with $\Pi_{\bar{p}_b} := I - \bar{p}_b\bar{p}_b^\top$ denoting the orthogonal projector onto the tangent space at $\bar{p}_b$, and

$$D\bar{p}_b[\delta\Theta] = \frac{1}{\|p_b\|_2}\Pi_{\bar{p}_b} Dp_b[\delta\Theta].$$

Proof. Since both $\Theta \mapsto h_{b,n}(\Theta)$ and $\Theta \mapsto s_{b,n}(\Theta)$ are C^1 and entmax is C^1 as a map from score vectors to weights, the product rule gives the displayed differential. The normalization formula is the standard derivative of the map $p \mapsto p/\|p\|_2$. □

Fix a compact convex trajectory neighborhood $\mathcal{K}_{\mathrm{traj}}$, and define the uniform bounds

$$H_{\max} := \sup_{\Theta\in\mathcal{K}_{\mathrm{traj}}} \max_{1\le n\le N_b} \|h_{b,n}(\Theta)\|_2,$$
$$G_f := \sup_{\Theta\in\mathcal{K}_{\mathrm{traj}}} \max_{1\le n\le N_b} \|Dh_{b,n}(\Theta)\|_{\mathrm{op}},$$
$$G_g := \sup_{\Theta\in\mathcal{K}_{\mathrm{traj}}} \max_{1\le n\le N_b} \|Dg_\phi(h_{b,n}(\Theta))\|_{\mathrm{op}}.$$

We also write

$$G_s := \sup_{\Theta \in \mathcal{K}_{\mathrm{traj}}} \|D\mathbf{s}_b(\Theta)\|_{\mathrm{op}},$$
$$R_b(\Theta) := \big\|[h_{b,1}(\Theta) - p_b(\Theta), \ldots, h_{b,N_b}(\Theta) - p_b(\Theta)]\big\|_{\mathrm{op}},$$
$$R_{\max} := \sup_{\Theta \in \mathcal{K}_{\mathrm{traj}}} R_b(\Theta).$$

Lemma 2 (Proxy Jacobian bound)**.** *Under the compact-neighborhood bounds above and Lemma 1,*

$$\|Dp_b\|_{\mathrm{op}} \le C_p := \frac{R_{\max} G_s}{\tau_e} + G_f, \qquad \|D\bar{p}_b\|_{\mathrm{op}} \le C_{\bar{p}} := \frac{C_p}{\nu_p}.$$

Proof. For the scorer path, the simplex constraint gives $\sum_n D\lambda_{b,n}[\delta\Theta] = 0$, hence

$$D^{\mathrm{score}} p_b[\delta\Theta] = \sum_{n=1}^{N_b} \big(h_{b,n} - p_b\big) D\lambda_{b,n}[\delta\Theta].$$

Writing the centered feature matrix as $H_b^{\mathrm{ctr}} := [h_{b,1} - p_b, \ldots, h_{b,N_b} - p_b]$, this becomes

$$D^{\mathrm{score}} p_b = H_b^{\mathrm{ctr}} J_{\mathrm{ent},b} D\mathbf{s}_b,$$

so

$$\|D^{\mathrm{score}} p_b\|_{\mathrm{op}} \le \|H_b^{\mathrm{ctr}}\|_{\mathrm{op}} \|J_{\mathrm{ent},b}\|_{\mathrm{op}} \|D\mathbf{s}_b\|_{\mathrm{op}} \le \frac{R_{\max} G_s}{\tau_e}.$$

For the encoder path,

$$\left\| \sum_n \lambda_{b,n} Dh_{b,n} \right\|_{\mathrm{op}} \le \sum_n \lambda_{b,n} \|Dh_{b,n}\|_{\mathrm{op}} \le G_f.$$

Adding the two bounds gives C_p. The normalized-proxy bound follows from the factor $\|p_b\|_2^{-1} \le \nu_p^{-1}$. □

The centered form in this proof shows that scorer-path sensitivity is controlled by the centered feature spread R_b rather than the absolute scale $N_b H_{\max}$, so the proxy map becomes smoother when the instances in a bag cluster near their proxy.

Lemma 3 (Proxy Hessian bound)**.** *Assume additionally that $h_{b,n}$ and $s_{b,n}$ are C^2 on $\mathcal{K}_{\mathrm{traj}}$, and that the entmax Jacobian is Lipschitz on $\mathcal{K}_{\mathrm{traj}}$. Then there exists a finite constant $C_p^{(2)}$ such that, wherever the second derivatives exist (almost everywhere; equivalently, Dp_b and $D\bar{p}_b$ are Lipschitz with the constants below),*

$$\|\nabla^2_\Theta p_b\|_{\mathrm{op}} \le C_p^{(2)}, \qquad \|\nabla^2_\Theta \bar{p}_b\|_{\mathrm{op}} \le \frac{C_p^{(2)}}{\nu_p} + \frac{2C_p^2}{\nu_p^2}.$$

The second term is the normalization-curvature contribution coming from the quadratic $Dp_b \otimes Dp_b$ terms.

Proof. Differentiate the formula in Proposition 1 once more. The scorer path produces terms involving the Lipschitz variation of the entmax Jacobian and the network Hessians, while the encoder path contributes the second derivatives of $h_{b,n}$. All such terms are uniformly bounded on the compact neighborhood, yielding $C_p^{(2)}$. Differentiating the normalization map $p \mapsto p/\|p\|_2$ adds one linear term of size $C_p^{(2)}/\nu_p$ and one quadratic term of size $2C_p^2/\nu_p^2$. □

B.2 UOT row-mass identity and functional role separation

We record the UOT row-mass identity and balanced-OT comparison formally, and then analyze the functional role of each surrogate component.

Proposition 2 (UOT row-mass identity specialized to instance grounding)**.** *Let the transport plan have factorized form*

$$\pi_{ij} = u_i \exp(-C_{ij}/\varepsilon) v_j = u_i K_{ij} v_j, \qquad K_{ij} := \exp(-C_{ij}/\varepsilon),$$

and define

$$m_i := \sum_{j=1}^{M} \pi_{ij}, \qquad \kappa_i := (Kv)_i = \sum_{j=1}^{M} K_{ij} v_j.$$

Assume the source-side UOT update

$$u_i = \left(\frac{\mu_i}{\kappa_i}\right)^{\varphi_{\mathrm{src}}}, \qquad \varphi_{\mathrm{src}} = \frac{\rho_{\mathrm{src}}}{\rho_{\mathrm{src}} + \varepsilon} \in (0, 1).$$

Then

$$m_i = \mu_i^{\varphi_{\mathrm{src}}} \kappa_i^{1-\varphi_{\mathrm{src}}}.$$

Hence, for uniform source prior $\mu_i \equiv \mu_0$,

$$\frac{m_i}{m_{i'}} = \left(\frac{\kappa_i}{\kappa_{i'}}\right)^{1-\varphi_{\mathrm{src}}}.$$

In particular, row mass is monotone in the effective compatibility score κ_i.

Proof. Since $m_i = \sum_j \pi_{ij} = u_i \sum_j K_{ij} v_j = u_i \kappa_i$, substituting the UOT update yields

$$m_i = \left(\frac{\mu_i}{\kappa_i}\right)^{\varphi_{\mathrm{src}}} \kappa_i = \mu_i^{\varphi_{\mathrm{src}}} \kappa_i^{1-\varphi_{\mathrm{src}}}.$$

The ratio formula follows immediately under $\mu_i \equiv \mu_0$, and monotonicity follows because $1 - \varphi_{\mathrm{src}} > 0$. □

Corollary 1 (Balanced transport fixes source-side row mass)**.** *If the source marginal is enforced exactly, as in balanced OT, then*

$$m_i = \mu_i \qquad \textit{for all } i.$$

Thus balanced OT can redistribute a row's mass across columns but cannot change the retained source mass used to weight the routing term.

Proof. Balanced OT enforces $\pi \mathbf{1} = \boldsymbol{\mu}$ exactly, so each row sum equals the prescribed source prior. □

Functional role separation across surrogate loss components. The implemented surrogate $\mathcal{L}_{\mathrm{grnd}}$ uses the routing term $\mathcal{L}_{\mathrm{rout}}$ rather than the exact generalized-KL row objective. This makes the functional split between components especially transparent. The row-mass identity is a direct consequence of the UOT scaling equations. It separates two roles that are entangled in balanced transport: the routing term determines where retained mass is assigned, whereas UOT row mass determines which instances receive larger weight within each bag's routing signal.

Observation 1 (Functional role separation across surrogate loss components)**.** *The four components of $\mathcal{L}_{\mathrm{grnd}}$ act on distinct aspects of the transport plan.*

(i) ***Routing term $\mathcal{L}_{\mathrm{rout}}$: compatibility-weighted routing.*** *Define*

$$r_b := \sum_{k:o(k)=b} \exp(m_k), \qquad \hat{m}_i := \frac{\exp(m_i)}{M r_{o(i)}}.$$

Then the implemented routing component is

$$\mathcal{L}_{\mathrm{rout}} = \sum_{i=1}^{N} \hat{m}_i \, \mathrm{KL}(\bar{T}_i \| \bar{\pi}_i).$$

Thus low-compatibility instances (small m_i) receive smaller relative weight inside their bag's routing update. Under balanced OT, $m_i = \mu_i$; with uniform source prior this reduces to $\hat{m}_i = 1/(M N_{o(i)})$, where $N_{o(i)}$ is the number of instances in the source bag of row i, so the compatibility-dependent weighting disappears.

(ii) ***Source marginal KL: row-mass deviation regularization.*** *The per-entry gradient is*

$$\frac{\partial}{\partial \pi_{ij}} \mathrm{GKL}(\pi \mathbf{1} \| \boldsymbol{\mu}) = \log \frac{m_i}{\mu_i}.$$

This is negative when $m_i < \mu_i$ and positive when $m_i > \mu_i$, so gradient descent in plan space restores m_i toward the prior. Hence $\mathcal{R}_s$ penalizes row-mass deviations from the prior.

(iii) ***Target marginal KL and entropy: column balancing and plan spreading.*** *The regularization terms $\mathcal{R}_t$ and $\mathcal{R}_e$ respectively push the column mass toward the prior ν_j and promote plan spreading. Their gradients are standard generalized-KL derivatives and do not enter the downstream anti-collapse analysis.*

Remark 1 (Routing-gradient interpretation)**.** *Under the frozen-dual view used in the downstream anti-collapse analysis, we treat the within-bag weight $\hat{m}_i$ and the target-side Sinkhorn scaling variables v_j as fixed (so that $\bar{\pi}_i$ varies only through $K_{i\cdot}$) when differentiating with respect to $\bar{h}_i$. For the implemented routing term, this gives the local gradient formula*

$$\tilde{G}_i^{\mathrm{rout}} := \nabla_{\bar{h}_i} \mathcal{L}_{\mathrm{rout}} = \frac{\hat{m}_i}{\varepsilon} (\bar{y}_i^P - \bar{y}_i^T),$$

where

$$\bar{y}_i^P := \sum_{j=1}^{M} \bar{\pi}_{ij}\, \bar{p}_j, \qquad \bar{y}_i^T := \sum_{j=1}^{M} \bar{T}_{ij}\, \bar{p}_j.$$

That is, the display is an implicit approximation: it omits the derivative paths through $\hat{m}_i$ and through the dependence of the dual/scaling variables on the transport solution. Under this frozen-dual approximation, the formula makes the role separation in Observation 1(i) explicit at the gradient level: the row mass enters only through the scalar weight $\hat{m}_i$, so lower-compatibility instances contribute less to the within-bag normalized routing update.

Remark 2 (Source of nuisance suppression)**.** *Under a uniform source prior, lower effective compatibility yields lower retained mass (Proposition 2); mass falls below the prior precisely when $\kappa_i < \mu_i$. The source marginal KL penalizes deviations from that prior. Balanced OT fixes $m_i = \mu_i$ (Corollary 1). Removing $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ while keeping mass-weighted routing (align-only in Table 2) lowers instance accuracy on all three benchmarks, supporting the inclusion of these regularizers.*

Lemma 4 (Entropic UOT solution-map regularity via the implicit function theorem)**.** *Fix $\varepsilon > 0$, $\rho_{\mathrm{src}}, \rho_{\mathrm{tgt}} > 0$ (with $\varphi_{\mathrm{tgt}} := \rho_{\mathrm{tgt}}/(\rho_{\mathrm{tgt}}+\varepsilon) \in (0,1)$), and strictly positive source and target marginals $\boldsymbol{\mu} \in \mathbb{R}^N_{>0}$ and $\boldsymbol{\nu} \in \mathbb{R}^M_{>0}$. Let $K(C)_{ij} := \exp(-C_{ij}/\varepsilon)$, write $r_i := \log u_i$ and $s_j := \log v_j$, and define the log-domain fixed-point equations*

$$F_i^{\mathrm{src}}(r,s,C) := r_i - \varphi_{\mathrm{src}} \log \mu_i + \varphi_{\mathrm{src}} \log\Big(\sum_{j=1}^{M} K(C)_{ij} e^{s_j}\Big),$$

$$F_j^{\mathrm{tgt}}(r,s,C) := s_j - \varphi_{\mathrm{tgt}} \log \nu_j + \varphi_{\mathrm{tgt}} \log\Big(\sum_{i=1}^{N} K(C)_{ij} e^{r_i}\Big).$$

On every compact cost set $\mathcal{C}_0 \subset [0,2]^{N\times M}$, the interior entropic-UOT scaling solution of $F(r,s,C) = 0$ defines a C^2 map

$$C \mapsto (r(C), s(C)),$$

and therefore a C^2 plan map

$$\pi(C)_{ij} = e^{r_i(C)} K(C)_{ij} e^{s_j(C)}.$$

Moreover, on $\mathcal{C}_0$ there exist finite constants J_{uot} and H_{uot} such that

$$\|D\pi(C)\|_{\mathrm{op}} \le J_{\mathrm{uot}}, \qquad \|D^2\pi(C)\|_{\mathrm{op}} \le H_{\mathrm{uot}} \qquad \forall C \in \mathcal{C}_0.$$

Proof. On $\mathcal{C}_0$, every kernel entry satisfies $K(C)_{ij} \in [e^{-2/\varepsilon}, 1]$. Since $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are strictly positive and $\varphi_{\mathrm{src}}, \varphi_{\mathrm{tgt}} \in (0,1)$, the log-domain equations above are smooth in (r,s,C) throughout the positive interior. For a fixed interior solution (r,s,C), define the sensitivity matrices

$$A_{ij} := \frac{K(C)_{ij} e^{s_j}}{\sum_{\ell=1}^{M} K(C)_{i\ell} e^{s_\ell}}, \qquad B_{ji} := \frac{K(C)_{ij} e^{r_i}}{\sum_{\ell=1}^{N} K(C)_{\ell j} e^{r_\ell}}.$$

Then each row of $A \in \mathbb{R}^{N\times M}$ and $B \in \mathbb{R}^{M\times N}$ sums to one, so both are row-stochastic. The Jacobian with respect to (r, s) has the block form

$$D_{(r,s)}F = \begin{bmatrix} I_N & \varphi_{\mathrm{src}}A \\ \varphi_{\mathrm{tgt}}B & I_M \end{bmatrix}.$$

Taking the Schur complement with respect to I_N gives

$$S = I_M - \varphi_{\mathrm{src}}\varphi_{\mathrm{tgt}}BA.$$

Since A and B are row-stochastic, BA is also row-stochastic, hence $\|BA\|_\infty = 1$ and $\rho(BA) \le 1$. Because $\varphi_{\mathrm{src}}\varphi_{\mathrm{tgt}} < 1$, we have

$$\rho(\varphi_{\mathrm{src}}\varphi_{\mathrm{tgt}}BA) \le \varphi_{\mathrm{src}}\varphi_{\mathrm{tgt}} < 1,$$

so S is invertible by the Neumann series

$$S^{-1} = \sum_{k=0}^{\infty}(\varphi_{\mathrm{src}}\varphi_{\mathrm{tgt}}BA)^k, \qquad \|S^{-1}\|_\infty \le \frac{1}{1-\varphi_{\mathrm{src}}\varphi_{\mathrm{tgt}}}.$$

Therefore $D_{(r,s)}F$ is invertible at every interior solution. The implicit function theorem now gives, for each fixed $C \in \mathcal{C}_0$, a local C^2 solution branch $C \mapsto (r(C), s(C))$. Since the interior entropic-UOT scaling solution is unique, these local branches agree on overlaps and define a global C^2 map on $\mathcal{C}_0$. Finally, compactness of $\mathcal{C}_0$ and a finite-subcover argument give uniform bounds on the first and second derivatives of $(r(C), s(C))$. Composing with $\pi_{ij} = e^{r_i}K(C)_{ij}e^{s_j}$, whose dependence on (r, s, C) is smooth, yields the stated C^2 regularity and the uniform bounds J_{uot} and H_{uot} for the plan map. □

B.3 L-smoothness on compact trajectory neighborhoods

This subsection records the smoothness statement enabled by the entmax-weighted proxy.

Assumption 1 (C^2 network regularity)**.** *The encoder f_θ, scorer g_ϕ, and classifier q_ψ are C^2 on a compact trajectory neighborhood $\mathcal{K}_{\mathrm{traj}}$, and their Jacobians and Hessians are uniformly bounded there.*

Assumption 2 (Trajectory-level Lipschitz continuity of the entmax Jacobian)**.** *On $\mathcal{K}_{\mathrm{traj}}$,*

$$\|J_{\mathrm{ent}}(\Theta) - J_{\mathrm{ent}}(\Theta')\|_{\mathrm{op}} \le L_{\mathrm{ent}}\|\Theta - \Theta'\|_2.$$

This is imposed as a trajectory-level regularity condition; it holds automatically on compact subsets away from entmax support-boundary degeneracies, and on all of $\mathcal{K}_{\mathrm{traj}}$ when $1 < \rho_e \le 3/2$: on the support, $\partial\lambda_n^{2-\rho_e}/\partial z_n$ is proportional to $\lambda_n^{3-2\rho_e}$, which can diverge as $\lambda_n \to 0$ only when $\rho_e > 3/2$.

Theorem 1 (L-smoothness on compact trajectory neighborhoods)**.** *Assume Assumptions 1 and 2 and the hypotheses of Lemma 4, and suppose that $\|p_b(\Theta)\|_2 \ge \nu_p > 0$ and $\|h_{b,n}(\Theta)\|_2 \ge \nu_h > 0$ on $\mathcal{K}_{\mathrm{traj}}$. Let $F = \mathcal{L}_{\mathrm{bag}} + \lambda_{\mathrm{grnd}}\mathcal{L}_{\mathrm{grnd}}$. Then*

$$\|\nabla F(\Theta) - \nabla F(\Theta')\|_2 \le L\|\Theta - \Theta'\|_2 \qquad \forall\, \Theta, \Theta' \in \mathcal{K}_{\mathrm{traj}},$$

with $L = L_{\mathrm{bag}} + \lambda_{\mathrm{grnd}}L_{\mathrm{grnd}}$, where one may take

$$\begin{aligned} L_{\mathrm{grnd}} &\le \big(C_{\bar P}^2 + C_{\bar H}^2\big)L_{ph} && \text{(A) Gauss–Newton} \\ &\quad + G_{\mathrm{grnd}}\left(\frac{C_p^{(2)}}{\nu_p} + \frac{2C_p^2}{\nu_p^2} + \frac{H_f}{\nu_h} + \frac{2G_f^2}{\nu_h^2}\right) && \text{(B) Newton correction,} \\ L_{ph} &\le \big(J_{\mathrm{uot}}^2 L_{\pi\pi} + G_\pi H_{\mathrm{uot}}\big)\|D_X C\|^2 + G_\pi J_{\mathrm{uot}}\,\|D_X^2 C\| && \text{(C) UOT chain rule.} \end{aligned}$$

Here $X := (\bar P, \bar H)$, with $\bar P := (\bar p_b)_{b=1}^M$ and $\bar H := (\bar h_i)_{i=1}^N$, collects all normalized proxies and instance embeddings, and $\Psi(X)$ denotes $\mathcal{L}_{\mathrm{grnd}}$ as a function of X through the cost $C(X)$ and the plan $\pi(C)$. The constants are as follows.

- *$C_{\bar P}$ and $C_{\bar H}$ bound the stacked Jacobians $D_\Theta\bar P$ and $D_\Theta\bar H$. By Lemma 2 and $\|D\bar h_i\|_{\mathrm{op}} \le \|Dh_i\|_{\mathrm{op}}/\|h_i\|_2$, one may take $C_{\bar P}^2 \le MC_{\bar p}^2$ and $C_{\bar H}^2 \le NG_f^2/\nu_h^2$.*

- L_{ph} *bounds* $\|\nabla_X^2 \Psi\|_{\mathrm{op}}$.
- G_{grnd} *bounds* $\sum_b \|\nabla_{\bar{p}_b} \Psi\|_2 + \sum_i \|\nabla_{\bar{h}_i} \Psi\|_2$.
- L_{bag} *bounds* $\|\nabla_\Theta^2 \mathcal{L}_{\mathrm{bag}}\|_{\mathrm{op}}$.
- H_f *bounds* $\|\nabla_\Theta^2 h_{b,n}\|_{\mathrm{op}}$.
- $L_{\pi\pi}$ *and* G_π *bound the Hessian and the gradient of* $\mathcal{L}_{\mathrm{grnd}}$ *as a function of the plan* π.
- $\|D_X C\|$ *and* $\|D_X^2 C\|$ *bound the first and second derivatives of the cost map* $X \mapsto C$.

All of these constants are finite on the compact neighborhood.

Proof. Step 1: bounded Hessian implies L-smoothness. All maps are C^2 except the entmax map, whose Jacobian is Lipschitz (Assumption 2). Hence ∇F is Lipschitz on the compact convex set $\mathcal{K}_{\mathrm{traj}}$ and differentiable almost everywhere, and it suffices to bound $\|\nabla_\Theta^2 F\|_{\mathrm{op}}$ uniformly wherever it exists.

Step 2: exact Hessian decomposition. Wherever the second derivatives exist, the chain rule for $\Theta \mapsto X(\Theta) \mapsto \Psi(X)$ gives the identity

$$\nabla_\Theta^2 \mathcal{L}_{\mathrm{grnd}} = H_A + H_B, \qquad H_A := (D_\Theta X)^\top \big(\nabla_X^2 \Psi\big)(D_\Theta X),$$
$$H_B := \sum_{b=1}^{M} \big\langle \nabla_{\bar{p}_b} \Psi, \nabla_\Theta^2 \bar{p}_b \big\rangle + \sum_{i=1}^{N} \big\langle \nabla_{\bar{h}_i} \Psi, \nabla_\Theta^2 \bar{h}_i \big\rangle,$$

where $\langle v, \nabla_\Theta^2 x \rangle$ denotes the Hessian of the scalar map $\Theta \mapsto v^\top x(\Theta)$ with v held fixed. Moreover, $\Psi = \Phi \circ \pi \circ C$, where $\Phi(\pi)$ denotes $\mathcal{L}_{\mathrm{grnd}}$ as a function of the plan. Hence the inner Hessian decomposes exactly as

$$\nabla_X^2 \Psi = H_C := (D_X C)^\top \Big[(D\pi)^\top \big(\nabla_\pi^2 \Phi\big)(D\pi) + \big\langle \nabla_\pi \Phi, D^2 \pi \big\rangle \Big] (D_X C) + \big\langle (D\pi)^\top \nabla_\pi \Phi,\, D_X^2 C \big\rangle.$$

Step 3: bound each term. Since $\|D_\Theta X\|_{\mathrm{op}}^2 \le \|D_\Theta \bar{P}\|_{\mathrm{op}}^2 + \|D_\Theta \bar{H}\|_{\mathrm{op}}^2$,

$$\|H_A\|_{\mathrm{op}} \le \big(C_{\bar{P}}^2 + C_{\bar{H}}^2\big) L_{ph}.$$

By Lemma 3, $\|\nabla_\Theta^2 \bar{p}_b\|_{\mathrm{op}} \le C_p^{(2)}/\nu_p + 2C_p^2/\nu_p^2$. The same normalization argument applied to $\bar{h}_i = h_i/\|h_i\|_2$ gives $\|\nabla_\Theta^2 \bar{h}_i\|_{\mathrm{op}} \le H_f/\nu_h + 2G_f^2/\nu_h^2$. Therefore

$$\|H_B\|_{\mathrm{op}} \le G_{\mathrm{grnd}} \left(\frac{C_p^{(2)}}{\nu_p} + \frac{2C_p^2}{\nu_p^2} + \frac{H_f}{\nu_h} + \frac{2G_f^2}{\nu_h^2} \right).$$

Finally, Lemma 4 gives $\|D\pi\|_{\mathrm{op}} \le J_{\mathrm{uot}}$ and $\|D^2\pi\|_{\mathrm{op}} \le H_{\mathrm{uot}}$ on the compact cost set, so

$$\|H_C\|_{\mathrm{op}} \le \big(J_{\mathrm{uot}}^2 L_{\pi\pi} + G_\pi H_{\mathrm{uot}}\big) \|D_X C\|^2 + G_\pi J_{\mathrm{uot}} \|D_X^2 C\|,$$

which is the bound (C) on L_{ph}.

Step 4: collect constants. Combining $\|\nabla_\Theta^2 \mathcal{L}_{\mathrm{bag}}\|_{\mathrm{op}} \le L_{\mathrm{bag}}$ with $\|\nabla_\Theta^2 \mathcal{L}_{\mathrm{grnd}}\|_{\mathrm{op}} \le \|H_A\|_{\mathrm{op}} + \|H_B\|_{\mathrm{op}} \le L_{\mathrm{grnd}}$ bounds $\|\nabla_\Theta^2 F\|_{\mathrm{op}}$ by $L = L_{\mathrm{bag}} + \lambda_{\mathrm{grnd}} L_{\mathrm{grnd}}$ wherever it exists. Step 1 then gives the claim. □

C Anti-Collapse Theory

This section collects the local anti-collapse identities underlying the proxy-separation analysis. It first records the frozen-dual baseline and then develops the exact coupled weighted-proxy refinement.

C.1 Frozen-dual baseline

Setup. Fix a class c and let

$$\mathcal{Q}_c = \mathcal{P}_c := \{j : c(j) = c\}, \qquad K_c := |\mathcal{Q}_c|.$$

We say that class c is in a *symmetric same-class collapse state* if there exist

$$\bar{p}^* \in \mathbb{S}^{d-1}, \qquad v^* \in \mathbb{R}_{>0}, \qquad \bar{p}_j = \bar{p}^*, \qquad v_j = v^*, \qquad \forall j \in \mathcal{Q}_c.$$

We work with the cosine-type cost

$$C_{ij} = 1 - \bar{h}_i^\top \bar{p}_j,$$

the exponentially normalized routing loss

$$r_b := \sum_{k:o(k)=b} \exp(m_k), \qquad \hat{m}_i := \frac{\exp(m_i)}{M r_{o(i)}}, \qquad \mathcal{L}_{\mathrm{rout}} := \sum_{i=1}^{N} \hat{m}_i \, \mathrm{KL}(\bar{T}_i \| \bar{\pi}_i),$$

and the row-normalized transport plan $\bar{\pi}_{ij} = \pi_{ij}/m_i$. For an instance i of class $c_i = c$, the normalized self-anchored target has the closed form

$$\bar{T}_{ij} = \tau_0 \, \mathbf{1}\{j = o(i)\} + \tau_1 \, \mathbf{1}\{j \in \mathcal{Q}_c, \, j \neq o(i)\},$$
$$\tau_0 := \frac{1}{1 + (K_c - 1)\gamma}, \qquad \tau_1 := \frac{\gamma}{1 + (K_c - 1)\gamma}.$$

Hence $\gamma = 0$ gives the self-only target and $\gamma = 1$ gives the uniform same-class target. Moreover,

$$\tau_0 - \tau_1 = \Delta_\gamma.$$

Under the frozen-dual baseline below, the dual variables u_i, v_j are held fixed and the exponentially normalized coefficients $\hat{m}_i$ are treated as local weights when differentiating with respect to the proxy embeddings.

For each class-c proxy $j \in \mathcal{Q}_c$, define

$$\mathcal{C}_c := \{i : c_i = c\}, \qquad \mathcal{C}_{c,j} := \{i \in \mathcal{C}_c : o(i) = j\},$$
$$\bar{\bar{g}}_c^{\mathrm{frozen}} := \frac{1}{K_c} \sum_{j \in \mathcal{Q}_c} \nabla_{\bar{p}_j}^{\mathrm{frozen}} \mathcal{L}_{\mathrm{rout}},$$
$$\bar{\bar{g}}_c^{\mathrm{exact}} := \frac{1}{K_c} \sum_{j \in \mathcal{Q}_c} \nabla_{\bar{p}_j}^{\mathrm{exact}} \mathcal{L}_{\mathrm{rout}},$$
$$M_{c,j} := \sum_{i \in \mathcal{C}_{c,j}} \hat{m}_i,$$
$$\bar{M}_c := \frac{1}{K_c} \sum_{j \in \mathcal{P}_c} M_{c,j}.$$

Frozen-dual gradient formula.

Lemma 5 (Frozen-dual gradient of $\mathcal{L}_{\mathrm{rout}}$)**.** *Under the cosine-type cost and the frozen-dual view, at the symmetric same-class collapse state,*

$$\nabla_{\bar{p}_j}^{\mathrm{frozen}} \mathcal{L}_{\mathrm{rout}} = \frac{1}{\varepsilon} \Pi_{\bar{p}_j} \sum_{i=1}^{N} \hat{m}_i (\bar{\pi}_{ij} - \bar{T}_{ij}) \bar{h}_i.$$

Proof. Since under the frozen-dual view the coefficients $\hat{m}_i$ are treated as fixed with respect to the proxy variables—equivalently, both $\exp(m_i)$ and $r_{o(i)}$ are frozen in the local differentiation—

$$\nabla_{\bar{p}_j}^{\mathrm{frozen}} \mathcal{L}_{\mathrm{rout}} = \sum_i \hat{m}_i \nabla_{\bar{p}_j}^{\mathrm{frozen}} \mathrm{KL}(\bar{T}_i \| \bar{\pi}_i).$$

For each row i, using $\bar{\pi}_{ij} = K_{ij} v_j / \kappa_i$ with $\kappa_i = \sum_r K_{ir} v_r$ under the frozen-dual view,

$$\nabla_{\bar{p}_j}^{\mathrm{frozen}} \mathrm{KL}(\bar{T}_i \| \bar{\pi}_i) = \frac{1}{\varepsilon} (\bar{\pi}_{ij} - \bar{T}_{ij}) \Pi_{\bar{p}_j} \bar{h}_i.$$

Multiplying by $\hat{m}_i$ and summing over i gives the stated formula. □

Baseline identity.

Proposition 3 (Frozen-dual baseline for self-anchored exponentially normalized routing)**.** *Under the setup above, define*

$$\Delta_\gamma := \frac{1-\gamma}{1+(K_c-1)\gamma}, \qquad \tilde{\beta}_c := \frac{\Delta_\gamma}{\varepsilon}, \qquad \tilde{w}_j := \Pi_{\bar{p}^*}\left(\sum_{i\in\mathcal{C}_{c,j}} \hat{m}_i \bar{h}_i - \frac{1}{K_c}\sum_{i\in\mathcal{C}_c} \hat{m}_i \bar{h}_i\right).$$

Then, for every $j \in \mathcal{Q}_c$,

$$\Pi_{\bar{p}^*}\left(\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}^{\mathrm{frozen}}_c\right) = -\tilde{\beta}_c\,\tilde{w}_j.$$

Moreover, for any C^2 retraction $R_{\bar{p}^*}$ with $R_{\bar{p}^*}(0) = \bar{p}^*$, $DR_{\bar{p}^*}(0) = I$, and $\bar{p}_j^+ := R_{\bar{p}^*}(\eta\tilde{d}_j)$ where $\tilde{d}_j := -\Pi_{\bar{p}^*}\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}}$,

$$\bar{p}_j^+ - \bar{p}_k^+ = \eta\tilde{\beta}_c(\tilde{w}_j - \tilde{w}_k) + O(\eta^2) \qquad \forall j,k \in \mathcal{Q}_c.$$

When $\gamma < 1$ and $\tilde{w}_j \neq \tilde{w}_k$, a sufficiently small step separates proxies j and k to first order. Equivalently, the frozen-dual field strength is proportional to the self-anchor gap Δ_γ. When $\gamma = 1$, $\Delta_\gamma = 0$ and the symmetry-breaking term vanishes exactly.

Proof. At the symmetric same-class collapse state, $\pi_{ij} = u_i K_i^* v^*$ is independent of $j \in \mathcal{P}_c$, hence $\bar{\pi}_{ij} = \bar{\pi}_{ik}$ for all $j,k \in \mathcal{P}_c$. By Lemma 5,

$$\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}} = \frac{1}{\varepsilon}\Pi_{\bar{p}^*}\sum_{i=1}^N \hat{m}_i(\bar{\pi}_{ij} - \bar{T}_{ij})\bar{h}_i.$$

Averaging over $j \in \mathcal{Q}_c$ gives

$$\bar{\tilde{g}}^{\mathrm{frozen}}_c = \frac{1}{\varepsilon}\Pi_{\bar{p}^*}\sum_{i=1}^N \hat{m}_i\left(\bar{\pi}_{ij} - \frac{1}{K_c}\sum_{k\in\mathcal{Q}_c}\bar{T}_{ik}\right)\bar{h}_i.$$

Here $\bar{\pi}_{ij}$ is independent of the choice of $j \in \mathcal{P}_c$ at the symmetric collapse state, so any representative same-class index j yields the same value in the display. Subtracting yields

$$\Pi_{\bar{p}^*}\left(\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}^{\mathrm{frozen}}_c\right) = -\frac{1}{\varepsilon}\Pi_{\bar{p}^*}\sum_{i=1}^N \hat{m}_i\left(\bar{T}_{ij} - \frac{1}{K_c}\sum_{k\in\mathcal{Q}_c}\bar{T}_{ik}\right)\bar{h}_i,$$

because the $\bar{\pi}$ terms cancel across same-class columns.

For $i \notin \mathcal{C}_c$, all same-class target entries vanish and the contribution is zero. For $i \in \mathcal{C}_c$, since $\tau_0 - \tau_1 = \Delta_\gamma$,

$$\hat{m}_i\left(\bar{T}_{ij} - \frac{1}{K_c}\sum_{k\in\mathcal{Q}_c}\bar{T}_{ik}\right) = \Delta_\gamma\hat{m}_i\left(\mathbf{1}\{i\in\mathcal{C}_{c,j}\} - \frac{1}{K_c}\mathbf{1}\{i\in\mathcal{C}_c\}\right).$$

Substituting gives

$$\Pi_{\bar{p}^*}\left(\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}^{\mathrm{frozen}}_c\right) = -\frac{\Delta_\gamma}{\varepsilon}\tilde{w}_j = -\tilde{\beta}_c\,\tilde{w}_j.$$

The one-step de-collapse statement follows from a direct C^2 retraction expansion. Since $R_{\bar{p}^*}(0) = \bar{p}^*$ and $DR_{\bar{p}^*}(0) = I$,

$$\bar{p}_j^+ = \bar{p}^* + \eta\tilde{d}_j + O(\eta^2), \qquad \bar{p}_k^+ = \bar{p}^* + \eta\tilde{d}_k + O(\eta^2).$$

Therefore

$$\bar{p}_j^+ - \bar{p}_k^+ = \eta(\tilde{d}_j - \tilde{d}_k) + O(\eta^2).$$

Using

$$\Pi_{\bar{p}^*}\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}} = \bar{\tilde{g}}^{\mathrm{frozen}}_c - \tilde{\beta}_c\tilde{w}_j, \qquad \Pi_{\bar{p}^*}\nabla^{\mathrm{frozen}}_{\bar{p}_k}\mathcal{L}_{\mathrm{rout}} = \bar{\tilde{g}}^{\mathrm{frozen}}_c - \tilde{\beta}_c\tilde{w}_k,$$

we obtain

$$\tilde{d}_j - \tilde{d}_k = -\Pi_{\bar{p}^*}\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{rout}} + \Pi_{\bar{p}^*}\nabla^{\mathrm{frozen}}_{\bar{p}_k}\mathcal{L}_{\mathrm{rout}} = \tilde{\beta}_c(\tilde{w}_j - \tilde{w}_k),$$

which proves the stated first-order separation formula. □

Remark 3 (Interpretation and relation to the exact coupled field)**.** *The parameter γ trades same-class sharing against anti-collapse strength through the self-anchor gap Δ_γ: as $\gamma \uparrow 1$, the local symmetry-breaking coefficient shrinks to zero, and at $\gamma = 1$ the frozen-dual field vanishes exactly. Proposition 3 therefore provides the baseline identity. The exact coupled result in Theorem 2 adds the target-dual correction term $\varphi_{\mathrm{tgt}}(M_{c,j} - \bar{M}_c)\Pi_{\bar{p}^*}\tilde{\mu}^*$, so the exact–frozen discrepancy is driven by same-class retained-mass imbalance. It vanishes when $M_{c,j} = \bar{M}_c$ for all $j \in \mathcal{Q}_c$. For the implemented weights, $\sum_{i:o(i)=j} \hat{m}_i = 1/M$ for every bag (Eq. (5)), so $M_{c,j} = \bar{M}_c = 1/M$ and this correction vanishes identically: at collapse, the exact coupled field coincides with the frozen-dual field $-\tilde{\beta}_c\tilde{w}_j$. The correction is kept in Theorem 2 because the identity holds for general row weights. For such general weights, because the exact field takes the form $-\tilde{\beta}_c[\tilde{w}_j - \text{correction}]$, the net anti-collapse effect depends on the relative alignment between $\tilde{w}_j$ and the correction term: an aligned correction partially cancels the frozen contribution, while an opposing correction amplifies it. The anti-collapse field therefore remains nonzero as long as the sufficient condition of Theorem 2 holds, regardless of whether the correction amplifies or partially cancels the frozen-dual term.*

Corollary 2 (Implemented surrogate preserves the anti-collapse identity)**.** *At a symmetric same-class collapse state,*

$$\Pi_{\bar{p}^*}\left(\nabla^{\mathrm{frozen}}_{\bar{p}_j}\mathcal{L}_{\mathrm{grnd}} - \bar{\tilde{g}}^{\mathrm{frozen}}_c\right) = -\tilde{\beta}_c\,\tilde{w}_j \qquad \forall j \in \mathcal{Q}_c.$$

Here $\bar{\tilde{g}}^{\mathrm{frozen}}_c$ denotes the class mean of the frozen-dual gradients of $\mathcal{L}_{\mathrm{grnd}}$.

Proof. At collapse, the same-class columns are symmetric, so the class-relative centered gradients of $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ vanish under the frozen-dual view. The remaining routing contribution is exactly Proposition 3. □

The exact coupled analogue for the full surrogate is stated in Corollary 3 below.

C.2 Exact coupled analysis under the weighted proxy pullback

This subsection establishes the exact coupled anti-collapse identity for the exponentially normalized routing loss $\mathcal{L}_{\mathrm{rout}}$ at a symmetric same-class collapse state. The theorem is formulated at the proxy-variable level, and its pullback to parameter space uses the entmax-weighted proxy regularity from Appendix B.1.

Assumption 3 (Weighted-proxy pullback regularity)**.** *In a neighborhood of the symmetric same-class collapse state considered below, define the entmax weights and raw proxy by*

$$\lambda_b(\Theta) = \mathrm{Entmax}_{\rho_e}(\mathbf{s}_b(\Theta)/\tau_e), \qquad p_b(\Theta) = \sum_{n=1}^{N_b} \lambda_{b,n}(\Theta)\, h_{b,n}(\Theta).$$

Assume $\|p_b(\Theta)\|_2 > 0$ throughout this neighborhood. Then, by Proposition 1, both p_b and the normalized proxy

$$\bar{p}_b(\Theta) = \frac{p_b(\Theta)}{\|p_b(\Theta)\|_2}$$

are C^1 in Θ.

The proof proceeds at the proxy-variable level. Assumption 3 is used only to justify the pullback of the proxy-level field through the weighted proxy map. In this subsection we write

$$y_j := \bar{p}_j, \qquad y^* := \bar{p}^*,$$

keeping $\bar{h}_i$ for the normalized instance embedding.

Weight-derivative term at collapse.

Lemma 6 (Weight-derivative term vanishes in class-relative centering)**.** *At the symmetric same-class collapse state, the class-relative centering of*

$$\sum_i \mathrm{KL}(\bar{T}_i\|\bar{\pi}_i)\,\nabla^{\mathrm{exact}}_{y_j}\hat{m}_i$$

vanishes:

$$\Pi_{y^*}\left(\sum_i \mathrm{KL}(\bar{T}_i\|\bar{\pi}_i)\nabla_{y_j}^{\mathrm{exact}}\hat{m}_i - \frac{1}{K_c}\sum_{\ell\in\mathcal{P}_c}\sum_i \mathrm{KL}(\bar{T}_i\|\bar{\pi}_i)\nabla_{y_\ell}^{\mathrm{exact}}\hat{m}_i\right) = 0.$$

Proof. At collapse $\mathrm{KL}(\bar{T}_i\|\bar{\pi}_i)$ is independent of the particular same-class proxy index $j \in \mathcal{P}_c$. Moreover, for each row i, the exponentially normalized coefficient $\hat{m}_i = \exp(m_i)/(Mr_{o(i)})$ depends only on $\exp(m_i)$ and on the bag sum $r_{o(i)}$, both of whose same-class proxy derivatives are symmetric at collapse. Therefore the centered quantity

$$\sum_i \mathrm{KL}(\bar{T}_i\|\bar{\pi}_i)\left(\nabla_{y_j}^{\mathrm{exact}}\hat{m}_i - \frac{1}{K_c}\sum_{\ell\in\mathcal{P}_c}\nabla_{y_\ell}^{\mathrm{exact}}\hat{m}_i\right)$$

vanishes by the same row-wise symmetry argument. □

Log-domain decomposition.

Proposition 4 (Exact log-domain decomposition)**.** *At the symmetric same-class collapse state,*

$$\Pi_{y^*}\left(\nabla_{y_j}^{\mathrm{exact}}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}_c^{\mathrm{exact}}\right) = \Pi_{y^*}\left(\nabla_{y_j}^{\mathrm{frozen}}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}_c^{\mathrm{frozen}}\right) + \tilde{\Delta}_j,$$

where

$$\tilde{\Delta}_j := \Pi_{y^*}\sum_{i=1}^{N}\hat{m}_i\sum_{k=1}^{M}(\bar{\pi}_{ik} - \bar{T}_{ik})\left(\nabla_{y_j}\log v_k - \frac{1}{K_c}\sum_{\ell\in\mathcal{P}_c}\nabla_{y_\ell}\log v_k\right).$$

Proof. Since $\nabla_{y_j}^{\mathrm{exact}}\mathcal{L}_{\mathrm{rout}} = \sum_i \nabla_{y_j}^{\mathrm{exact}}(\hat{m}_i\mathrm{KL}_i)$, by Lemma 6 the $\sum_i \mathrm{KL}_i\nabla\hat{m}_i$ term vanishes after class-relative centering. Thus

$$\Pi_{y^*}\left(\nabla_{y_j}^{\mathrm{exact}}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}_c^{\mathrm{exact}}\right) = \Pi_{y^*}\sum_i \hat{m}_i\left(\nabla_{y_j}^{\mathrm{exact}}\mathrm{KL}_i - \frac{1}{K_c}\sum_{\ell\in\mathcal{P}_c}\nabla_{y_\ell}^{\mathrm{exact}}\mathrm{KL}_i\right).$$

Using

$$\bar{\pi}_{ik} = \frac{K_{ik}v_k}{\kappa_i}, \qquad \kappa_i = \sum_r K_{ir}v_r, \qquad \log\bar{\pi}_{ik} = \log K_{ik} + \log v_k - \log\kappa_i,$$

and differentiating,

$$\nabla_{y_j}\log\bar{\pi}_{ik} = \nabla_{y_j}\log K_{ik} + \nabla_{y_j}\log v_k - \sum_r \bar{\pi}_{ir}\big(\nabla_{y_j}\log K_{ir} + \nabla_{y_j}\log v_r\big).$$

Since $\sum_k \bar{T}_{ik} = 1$,

$$\nabla_{y_j}^{\mathrm{exact}}\mathrm{KL}_i = -\sum_k \bar{T}_{ik}\nabla_{y_j}\log\bar{\pi}_{ik} = \sum_k(\bar{\pi}_{ik} - \bar{T}_{ik})\big(\nabla_{y_j}\log K_{ik} + \nabla_{y_j}\log v_k\big).$$

Under the cosine-type cost $C_{ij} = 1 - \bar{h}_i^\top y_j$, $K_{ij} = \exp(-C_{ij}/\varepsilon)$, only the $k = j$ kernel depends directly on y_j:

$$\nabla_{y_j}\log K_{ij} = \frac{1}{\varepsilon}\Pi_{y_j}\bar{h}_i.$$

Therefore

$$\nabla_{y_j}^{\mathrm{exact}}\mathrm{KL}_i = (\bar{\pi}_{ij} - \bar{T}_{ij})\frac{1}{\varepsilon}\Pi_{y_j}\bar{h}_i + \sum_k(\bar{\pi}_{ik} - \bar{T}_{ik})\nabla_{y_j}\log v_k.$$

After weighting by $\hat{m}_i$, summing over i, and taking the class-relative centered part, the first term is precisely

$$\Pi_{y^*}\left(\nabla_{y_j}^{\mathrm{frozen}}\mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}_c^{\mathrm{frozen}}\right),$$

while the second term gives $\tilde{\Delta}_j$. □

Intermediate identities at collapse. The following identities describe the response of the UOT fixed point at the symmetric same-class collapse state.

Write the log-domain fixed-point equations from Lemma 4 as

$$\log u_i = \varphi_{\mathrm{src}} \log \mu_i - \varphi_{\mathrm{src}} \log \kappa_i, \qquad \kappa_i := \sum_{r=1}^{M} K_{ir} v_r,$$

$$\log v_k = \varphi_{\mathrm{tgt}} \log \nu_k - \varphi_{\mathrm{tgt}} \log \kappa'_k, \qquad \kappa'_k := \sum_{i=1}^{N} K_{ik} u_i.$$

For each same-class proxy index $j \in \mathcal{P}_c$, define the class-relative source-dual sensitivity by

$$(\Delta F_{f,j})_i := \nabla_{y_j} \log u_i - \frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \nabla_{y_\ell} \log u_i.$$

Proposition 5 (Source-side cancellation)**.** *At the symmetric same-class collapse state, the class-relative source-dual sensitivity vanishes:*

$$(\Delta F_{f,j})_i = 0 \qquad \forall i, \quad j \in \mathcal{P}_c.$$

Proof. At collapse all same-class proxies coincide at y^*, so for every $j \in \mathcal{P}_c$,

$$K_{ij} = K_i^*, \qquad \kappa_i = \sum_{r \notin \mathcal{P}_c} K_{ir} v_r + K_i^* \sum_{r \in \mathcal{P}_c} v_r.$$

Differentiating with respect to y_j:

$$\nabla_{y_j} \kappa_i = \frac{K_i^* v_j}{\varepsilon} \Pi_{y^*} \bar{h}_i + K_i^* \sum_{r \in \mathcal{P}_c} \nabla_{y_j} v_r + \sum_{r \notin \mathcal{P}_c} K_{ir} \nabla_{y_j} v_r.$$

The first term is independent of the chosen same-class index after class-relative centering because $v_j = v^*$ for all $j \in \mathcal{P}_c$ at collapse. The remaining sums are permutation-symmetric across same-class indices, so their class-relative centered parts vanish. Hence

$$\nabla_{y_j} \log \kappa_i - \frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \nabla_{y_\ell} \log \kappa_i = 0.$$

Since $\log u_i = \varphi_{\mathrm{src}} \log \mu_i - \varphi_{\mathrm{src}} \log \kappa_i$ and μ_i is constant, the same identity holds for $\log u_i$. □

Proposition 6 (Target-dual response)**.** *At the symmetric same-class collapse state,*

$$\Pi_{y^*} \left(\nabla_{y_j} \log v_k - \frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \nabla_{y_\ell} \log v_k \right) = \begin{cases} -\dfrac{\varphi_{\mathrm{tgt}}}{\varepsilon} \left(\delta_{kj} - \dfrac{1}{K_c} \right) \Pi_{y^*} \tilde{\mu}^*, & k \in \mathcal{P}_c, \\ 0, & k \notin \mathcal{P}_c, \end{cases}$$

where

$$\tilde{\mu}^* := \frac{\displaystyle\sum_{i=1}^{N} u_i K_i^* \bar{h}_i}{\displaystyle\sum_{i=1}^{N} u_i K_i^*} = \frac{\displaystyle\sum_{i=1}^{N} \pi_{ij}^* \bar{h}_i}{\displaystyle\sum_{i=1}^{N} \pi_{ij}^*} \qquad (j \in \mathcal{P}_c).$$

Proof. Fix $k \in \mathcal{P}_c$. Differentiating $\log v_k = \varphi_{\mathrm{tgt}} \log \nu_k - \varphi_{\mathrm{tgt}} \log \kappa'_k$:

$$\nabla_{y_j} \log v_k = -\frac{\varphi_{\mathrm{tgt}}}{\kappa'_k} \sum_{i=1}^{N} \Big(u_i \nabla_{y_j} K_{ik} + K_{ik} \nabla_{y_j} u_i \Big).$$

At collapse $K_{ik} = K_i^*$ for every $k \in \mathcal{P}_c$, and only the $k = j$ kernel depends directly on y_j:

$$\nabla_{y_j} K_{ik} = \frac{K_i^*}{\varepsilon} \delta_{kj} \Pi_{y^*} \bar{h}_i.$$

Because $\nabla_{y_j} u_i = u_i \nabla_{y_j} \log u_i$, Proposition 5 implies that the class-relative centered contribution of the $K_{ik} \nabla_{y_j} u_i$ term vanishes. Thus

$$\Pi_{y^*} \left(\nabla_{y_j} \log v_k - \frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \nabla_{y_\ell} \log v_k \right) = -\frac{\varphi_{\mathrm{tgt}}}{\varepsilon \kappa'_k} \left(\delta_{kj} - \frac{1}{K_c} \right) \Pi_{y^*} \sum_{i=1}^{N} u_i K_i^* \bar{h}_i.$$

Since $\kappa'_k = \sum_i u_i K_i^*$ is independent of $k \in \mathcal{P}_c$ at collapse, this is exactly the stated formula.

If $k \notin \mathcal{P}_c$, then K_{ik} does not depend directly on any same-class proxy y_j. The only remaining dependence is through u_i, whose class-relative centered derivative vanishes by Proposition 5. □

Main exact theorem.

Theorem 2 (Exact coupled anti-collapse under self-anchored routing)**.** *At the symmetric same-class collapse state, under Assumption 3, retain the frozen-dual quantities Δ_γ, $\tilde{\beta}_c$, $\tilde{w}_j$, $M_{c,j}$, and $\bar{M}_c$. Then*

$$\Pi_{\bar{p}^*} \left(\nabla^{\mathrm{exact}}_{\bar{p}_j} \mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}^{\mathrm{exact}}_c \right) = -\tilde{\beta}_c \left[\tilde{w}_j - \varphi_{\mathrm{tgt}} (M_{c,j} - \bar{M}_c) \Pi_{\bar{p}^*} \tilde{\mu}^* \right].$$

Consequently, if $\gamma < 1$ and

$$\tilde{w}_j - \varphi_{\mathrm{tgt}} (M_{c,j} - \bar{M}_c) \Pi_{\bar{p}^*} \tilde{\mu}^* \neq 0,$$

then the class-relative exact coupled gradient is nonzero. In particular, if $M_{c,j} = \bar{M}_c$, then $\tilde{w}_j \neq 0$ is sufficient. When $\gamma = 1$, $\Delta_\gamma = 0$, and the self-anchor-gap-controlled symmetry-breaking term vanishes exactly.

Proof. By Proposition 4,

$$\Pi_{y^*} \left(\nabla^{\mathrm{exact}}_{y_j} \mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}^{\mathrm{exact}}_c \right) = -\tilde{\beta}_c \tilde{w}_j + \tilde{\Delta}_j.$$

It remains to evaluate $\tilde{\Delta}_j$. By Proposition 6, only same-class columns contribute:

$$\tilde{\Delta}_j = -\frac{\varphi_{\mathrm{tgt}}}{\varepsilon} \tilde{R}_j \Pi_{y^*} \tilde{\mu}^*,$$

where

$$\tilde{R}_j = \sum_i \hat{m}_i \sum_{k \in \mathcal{P}_c} (\bar{\pi}_{ik} - \bar{T}_{ik}) \left(\delta_{kj} - \frac{1}{K_c} \right).$$

The extra class-centering term vanishes:

$$\frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \sum_i \hat{m}_i \sum_{k \in \mathcal{P}_c} (\bar{\pi}_{ik} - \bar{T}_{ik}) \left(\delta_{k\ell} - \frac{1}{K_c} \right) = 0.$$

At symmetric same-class collapse, define

$$M_c := \sum_{i \in \mathcal{C}_c} \hat{m}_i = K_c \bar{M}_c, \qquad \tilde{r}_j := \sum_i \hat{m}_i (\bar{\pi}_{ij} - \bar{T}_{ij}).$$

Since $\hat{m}_i \bar{\pi}_{ij} = \exp(m_i)\, u_i K_i^* v^* / (M\, m_i\, r_{o(i)})$ is independent of $j \in \mathcal{P}_c$, so is $\sum_i \hat{m}_i \bar{\pi}_{ij}$. Therefore the class-relative variation of $\tilde{r}_j$ is entirely target-driven:

$$\tilde{R}_j = \tilde{r}_j - \frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \tilde{r}_\ell.$$

Using $\sum_i \hat{m}_i \bar{T}_{ij} = \tau_0 M_{c,j} + \tau_1 (M_c - M_{c,j})$,

$$\tilde{R}_j = -\Delta_\gamma (M_{c,j} - \bar{M}_c).$$

Therefore

$$\tilde{\Delta}_j = \frac{\varphi_{\mathrm{tgt}} \Delta_\gamma}{\varepsilon} (M_{c,j} - \bar{M}_c) \Pi_{y^*} \tilde{\mu}^* = \varphi_{\mathrm{tgt}} \tilde{\beta}_c (M_{c,j} - \bar{M}_c) \Pi_{y^*} \tilde{\mu}^*.$$

Combining:

$$\Pi_{y^*} \left(\nabla^{\mathrm{exact}}_{y_j} \mathcal{L}_{\mathrm{rout}} - \bar{\tilde{g}}^{\mathrm{exact}}_c \right) = -\tilde{\beta}_c \left[\tilde{w}_j - \varphi_{\mathrm{tgt}} (M_{c,j} - \bar{M}_c) \Pi_{y^*} \tilde{\mu}^* \right]. \qquad \square$$

Remark 4 (Pullback through the weighted proxy)**.** *The theorem is stated at the proxy-variable level. Under Assumption 3, the differential of Proposition 1 shows that the exact coupled proxy-level field pulls back through two weighted-proxy routes: the scorer path through the entmax Jacobian and the encoder path through the directly weighted feature differentials. The scorer route modulates the sparse proxy weights, while the encoder route perturbs the weighted feature sum directly through $Dh_{b,n}$.*

Surrogate extension.

Lemma 7 (Non-routing terms vanish in class-relative centering at collapse)**.** *At a symmetric same-class collapse state, the class-relative centered contributions of $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ vanish.*

Proof. At collapse the symmetric same-class columns satisfy $\pi_{ij} = u_i K_i^* v^*$ for all $j \in \mathcal{P}_c$. Hence the row masses, target-side marginal factors, and entropy factors are symmetric across $j \in \mathcal{P}_c$, so their class-relative centered gradients vanish. □

Corollary 3 (Corrected exact class-relative field for the full surrogate)**.** *Assume that the same-class target priors are symmetric, $\nu_j = \nu_c^*$ for all $j \in \mathcal{P}_c$. At the symmetric same-class collapse state, under Assumption 3 and symmetric same-class target priors,*

$$\Pi_{\bar{p}^*}\left(\nabla^{\mathrm{exact}}_{\bar{p}_j}\mathcal{L}_{\mathrm{grnd}} - \bar{\tilde{g}}_c^{\mathrm{exact}}\right) = -\tilde{\beta}_c\left[\tilde{w}_j - \varphi_{\mathrm{tgt}}(M_{c,j} - \bar{M}_c)\Pi_{\bar{p}^*}\tilde{\mu}^*\right].$$

Here $\bar{\tilde{g}}_c^{\mathrm{exact}}$ denotes the class mean of the exact gradients of $\mathcal{L}_{\mathrm{grnd}}$.

Proof. By Lemma 7, $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ vanish in class-relative centering at collapse. The result follows from Theorem 2. □

D Stability and Separation

This section collects the local curvature analysis, its stability under imbalance, and the trajectory-level non-persistence theorem for the implemented exponentially normalized routing surrogate.

D.1 Gauss–Newton curvature of $\mathcal{L}_{\mathrm{rout}}$ at collapse

We use the notation of Appendix C.1. To keep the curvature formulas compact, we additionally write

$$\mathcal{P}_c := \mathcal{Q}_c, \qquad y_j := \bar{p}_j, \qquad y^* := \bar{p}^*.$$

In this subsection only $q_i^{(c)} := \sum_{j\in\mathcal{P}_c}\pi_{ij}^*$ is new; under equal bag sizes, the factor $1/N_b$ becomes $1/N$, and the exponentially normalized routing weights satisfy $\hat{m}_i^* = 1/N$.

Fix a class c and assume the equal-bag-size symmetric collapse state:

$$y_j = y^*, \qquad v_j = v^*, \qquad \nu_j = \bar{\nu}_c \quad \text{for all } j \in \mathcal{P}_c.$$

Define the centered tangent subspace

$$V_c(y^*) := \left\{ w = (w_j)_{j\in\mathcal{P}_c} : w_j \in T_{y^*}\mathbb{S}^{d-1},\ \sum_{j\in\mathcal{P}_c} w_j = 0 \right\},$$

and set

$$\hat{h}_i := \Pi_{y^*}\bar{h}_i, \qquad q_i^{(c)} := \sum_{j\in\mathcal{P}_c}\pi_{ij}^*.$$

Here N is the fixed number of instances in the equal-bag-size setting, and $m_i^* = \sum_{j=1}^M \pi_{ij}^*$ denotes the retained mass at collapse. In the equal-bag-size symmetric collapse state considered below, we additionally assume that $m_i^* = m^* > 0$ is uniform across rows. Consequently,

$$\hat{m}_i^* = \frac{\exp(m^*)}{MN_b\exp(m^*)} = \frac{1}{N}.$$

Gauss–Newton curvature formula.

Proposition 7 (Frozen-dual centered-subspace Gauss–Newton form for $\mathcal{L}_{\mathrm{rout}}$)**.** *Under the assumptions above, for every $w \in V_c(y^*)$,*

$$Q_{\mathrm{GN}}^{\mathrm{fr}}(\mathcal{L}_{\mathrm{rout}})(w) = \frac{1}{Nm^*\varepsilon^2 K_c}\sum_{i=1}^N q_i^{(c)}\sum_{j\in\mathcal{P}_c}\left(\hat{h}_i^\top w_j\right)^2.$$

Equivalently, defining

$$\Sigma_c^{\mathrm{fr}} := \sum_{i=1}^{N} q_i^{(c)} \hat{h}_i \hat{h}_i^\top,$$

we have

$$Q_{\mathrm{GN}}^{\mathrm{fr}}(\mathcal{L}_{\mathrm{rout}})(w) = \frac{1}{N m^* \varepsilon^2 K_c} \sum_{j \in \mathcal{P}_c} w_j^\top \Sigma_c^{\mathrm{fr}} w_j.$$

Proof. Step 1 (First-order cancellation). At the symmetric same-class collapse state, the frozen-dual surrogate treats $\hat{m}_i$ as fixed local weights. Hence $\delta \hat{m}_i = 0$, and therefore

$$\delta^2 \mathcal{L}_{\mathrm{rout}}^{\mathrm{GN}} = \sum_{i=1}^{N} \hat{m}_i^* \, \delta^2(\mathrm{KL}_i)^{\mathrm{GN}} = \frac{1}{N} \sum_{i=1}^{N} \delta^2(\mathrm{KL}_i)^{\mathrm{GN}}.$$

Step 2 (Raw-logit perturbation). With u_i, v_j fixed, $\xi_{ij}^{\mathrm{fr}} := \log u_i + \log v_j - (1 - \bar{h}_i^\top y_j)/\varepsilon$, so differentiating at $y_j = y^*$ in direction $w_j \in T_{y^*}\mathbb{S}^{d-1}$ gives

$$\delta \xi_{ij}^{\mathrm{fr}} = \frac{(\Pi_{y^*} \bar{h}_i)^\top w_j}{\varepsilon} = \frac{\hat{h}_i^\top w_j}{\varepsilon}.$$

Step 3 (Summation). For $\mathrm{KL}(\bar{T}_i \| \bar{\pi}_i)$, using $\bar{\pi}_{ij} = \pi_{ij}/m_i$ under the frozen-dual view, the Hessian of KL_i with respect to the logits $\xi_{i\cdot}^{\mathrm{fr}}$ is

$$\nabla^2_{\xi_{i\cdot}} \mathrm{KL}_i = \mathrm{diag}(\bar{\pi}_{i\cdot}) - \bar{\pi}_{i\cdot} \bar{\pi}_{i\cdot}^\top \qquad \text{(Gauss–Newton part)}.$$

At collapse, $\bar{\pi}_{ij}^* = \pi_{ij}^*/m_i^* = q_i^{(c)}/(K_c m^*)$ for all $j \in \mathcal{P}_c$. Since $\delta \xi_{ij}^{\mathrm{fr}} = 0$ for $j \notin \mathcal{P}_c$ and $\sum_{j \in \mathcal{P}_c} w_j = 0$, the rank-one part contributes $\big(\sum_j \bar{\pi}_{ij}^* \delta \xi_{ij}^{\mathrm{fr}}\big)^2 = \big(\frac{q_i^{(c)}}{K_c m^* \varepsilon} \hat{h}_i^\top \sum_{j \in \mathcal{P}_c} w_j\big)^2 = 0$. Therefore

$$\delta^2(\mathrm{KL}_i)^{\mathrm{GN}} = \sum_{j \in \mathcal{P}_c} \bar{\pi}_{ij}^* (\delta \xi_{ij}^{\mathrm{fr}})^2 = \frac{q_i^{(c)}}{K_c m^* \varepsilon^2} \sum_{j \in \mathcal{P}_c} (\hat{h}_i^\top w_j)^2.$$

Summing over i and dividing by N:

$$Q_{\mathrm{GN}}^{\mathrm{fr}}(\mathcal{L}_{\mathrm{rout}})(w) = \frac{1}{N m^* K_c \varepsilon^2} \sum_{i=1}^{N} q_i^{(c)} \sum_{j \in \mathcal{P}_c} (\hat{h}_i^\top w_j)^2,$$

which gives the stated covariance form. □

Remark 5 (Exponential mass factor and curvature scale)**.** *At collapse, $m_i^* = m^*$ gives $\hat{m}_i^* = 1/N$ because the numerator $\exp(m^*)$ and the full denominator $M N_b \exp(m^*) = N \exp(m^*)$ cancel. Equivalently, within each bag one has $r_b = N_b \exp(m^*)$. However, the row-normalized routing term retains $\bar{\pi}_{ij}^* = q_i^{(c)}/(K_c m^*)$, so the curvature scale acquires an additional $1/m^*$ factor. Only $\exp(m^*)$ cancels; the retained mass m^* itself does not. The covariance matrix $\Sigma_c^{\mathrm{fr}} = \sum_i q_i^{(c)} \hat{h}_i \hat{h}_i^\top$ remains unchanged.*

Remark 6 (Geodesic-curvature correction)**.** *Along the sphere geodesic with initial velocity $w_j \in T_{y^*}\mathbb{S}^{d-1}$, one has $\ddot{y}_j(0) = -\|w_j\|_2^2 y^*$. The exact frozen-dual second variation contains the additional correction*

$$-\frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \mathcal{P}_c} (\bar{\pi}_{ij}^* - \bar{T}_{ij}) \frac{(\bar{h}_i^\top y^*) \|w_j\|_2^2}{\varepsilon}.$$

This generally does not vanish at collapse when $\bar{\pi}_{ij}^* \neq \bar{T}_{ij}$ for some (i, j), so Proposition 7 is an exact formula for the Gauss–Newton part only, not a full Hessian identity.

D.2 Curvature stability under within-class imbalance

This subsection proves that the centered-subspace curvature of the exponentially normalized routing loss remains stable under bounded relative within-class target-prior imbalance. We retain the collapse-state setup of Appendix D.1 and additionally quantify the within-class imbalance by $\delta_c := \max_{j \in \mathcal{P}_c} |\delta_{c,j}|$.

Setup and imbalance parametrization. Retain the collapse-state setup of Appendix D.1; in particular, $\mathcal{P}_c$, K_c, $\hat{h}_i$, $q_i^{(c)}$, m_i^*, Σ_c^{fr}, and $V_c(y^*)$ are as defined there. We additionally quantify the target-side within-class imbalance by

$$\nu_j = \bar{\nu}_c(1+\delta_{c,j}), \qquad j \in \mathcal{P}_c,$$

where

$$\delta_{c,j} := \frac{\nu_j - \bar{\nu}_c}{\bar{\nu}_c}, \qquad \delta_c := \max_{j\in\mathcal{P}_c} |\delta_{c,j}|, \qquad \sum_{j\in\mathcal{P}_c} \delta_{c,j} = 0.$$

We assume $\delta_c < 1$.

First-order dual perturbation under imbalance.

Lemma 8 (First-order dual perturbation under imbalance)**.** *At the symmetric same-class collapse state, the target dual satisfies*

$$v_j = v^*(1+\varphi_{\mathrm{tgt}}\delta_{c,j}) + O(\delta_c^2), \qquad j \in \mathcal{P}_c,$$

while the source dual is unchanged to first order:

$$u_i^{\mathrm{imbal}} = u_i^{\mathrm{bal}} + O(\delta_c^2).$$

Consequently,

$$\pi_{ij}^{\mathrm{imbal}} = \frac{q_i^{(c)}}{K_c}(1+\varphi_{\mathrm{tgt}}\delta_{c,j}) + O(\delta_c^2), \qquad j \in \mathcal{P}_c,$$

and

$$m_i^{\mathrm{imbal}} = m^* + O(\delta_c^2), \qquad m_{\mathrm{tot}}^{\mathrm{imbal}} = Nm^* + O(\delta_c^2).$$

Proof. The target fixed-point equation gives, within the same-class block,

$$v_j \propto \nu_j^{\varphi_{\mathrm{tgt}}} = \bar{\nu}_c^{\varphi_{\mathrm{tgt}}}(1+\delta_{c,j})^{\varphi_{\mathrm{tgt}}} = v^*(1+\varphi_{\mathrm{tgt}}\delta_{c,j}) + O(\delta_c^2).$$

At collapse, $K_{ij} = K_i^*$ is independent of $j \in \mathcal{P}_c$. Hence

$$\kappa_i^{\mathrm{imbal}} = \sum_{j\in\mathcal{P}_c} K_i^* v_j = K_i^* v^* K_c \left(1 + \frac{\varphi_{\mathrm{tgt}}}{K_c}\sum_{j\in\mathcal{P}_c}\delta_{c,j}\right) + O(\delta_c^2) = K_i^* v^* K_c + O(\delta_c^2).$$

Thus the source dual is unchanged to first order, and

$$\pi_{ij}^{\mathrm{imbal}} = u_i^{\mathrm{imbal}} K_i^* v_j = \pi_{ij}^*(1+\varphi_{\mathrm{tgt}}\delta_{c,j}) + O(\delta_c^2).$$

Since $\pi_{ij}^* = q_i^{(c)}/K_c$ for $j \in \mathcal{P}_c$, this gives the stated perturbation of π_{ij}. Finally,

$$m_i^{\mathrm{imbal}} = \sum_j \pi_{ij}^{\mathrm{imbal}} = m^* + \sum_{j\in\mathcal{P}_c} \pi_{ij}^*\varphi_{\mathrm{tgt}}\delta_{c,j} + O(\delta_c^2) = m^* + O(\delta_c^2),$$

because $\sum_{j\in\mathcal{P}_c}\delta_{c,j} = 0$. Summing over i gives $m_{\mathrm{tot}}^{\mathrm{imbal}} = Nm^* + O(\delta_c^2)$. □

Remark 7 (Why row mass is unchanged to first order)**.** *The first-order same-class contribution to the row mass is*

$$\sum_{j\in\mathcal{P}_c} \pi_{ij}^*\varphi_{\mathrm{tgt}}\delta_{c,j} = \frac{q_i^{(c)}}{K_c}\varphi_{\mathrm{tgt}}\sum_{j\in\mathcal{P}_c}\delta_{c,j} = 0.$$

Thus imbalance perturbs the row-normalized routing pattern $\bar{\pi}_{ij}$ to first order, but not the retained mass scale m_i or the total mass m_{tot}.

Gauss–Newton curvature stability.

Theorem 3 (Gauss–Newton curvature stability under bounded imbalance)**.** *Retain the assumptions of Proposition 7, except replace symmetric same-class target priors by bounded relative target-prior imbalance $\delta_c < 1$. Then, for every $w \in V_c(y^*)$, the frozen-dual Gauss–Newton curvature satisfies*

$$Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) = \frac{1}{Nm^*\varepsilon^2 K_c} \sum_{j \in \mathcal{P}_c} w_j^\top \Sigma_c^{\mathrm{fr}} w_j + R_{\mathrm{GN}}^{\mathrm{imbal}}(w),$$

where

$$\left|R_{\mathrm{GN}}^{\mathrm{imbal}}(w)\right| \le \frac{\varphi_{\mathrm{tgt}} \delta_c \lambda_{\max}(\Sigma_c^{\mathrm{fr}})}{Nm^*\varepsilon^2 K_c} \|w\|_2^2 + O(\delta_c^2 \|w\|_2^2).$$

Proof. We track the Gauss–Newton component of the frozen-dual row-normalized routing curvature. In this calculation, the exponentially normalized coefficients are evaluated at the symmetric collapse state, where $\hat{m}_i^* = 1/N$, while the row-normalized routing probabilities are perturbed by the target-side imbalance. Thus normalization-related mass-variation terms and geodesic-curvature corrections are not included in $Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}$.

By Lemma 8,

$$\bar{\pi}_{ij}^{\mathrm{imbal}} = \frac{\pi_{ij}^{\mathrm{imbal}}}{m_i^{\mathrm{imbal}}} = \frac{q_i^{(c)}}{K_c m^*}(1 + \varphi_{\mathrm{tgt}} \delta_{c,j}) + O(\delta_c^2), \qquad j \in \mathcal{P}_c.$$

The frozen-dual raw-logit perturbation is unchanged:

$$\delta\xi_{ij}^{\mathrm{fr}} = \frac{\hat{h}_i^\top w_j}{\varepsilon}.$$

Therefore the perturbed row-wise Gauss–Newton second variation is

$$\delta^2(\mathrm{KL}_i^{\mathrm{imbal}})^{\mathrm{GN}} = \sum_{j \in \mathcal{P}_c} \bar{\pi}_{ij}^{\mathrm{imbal}} (\delta\xi_{ij}^{\mathrm{fr}})^2$$

$$= \frac{q_i^{(c)}}{K_c m^*} \sum_{j \in \mathcal{P}_c} (1 + \varphi_{\mathrm{tgt}} \delta_{c,j}) \frac{(\hat{h}_i^\top w_j)^2}{\varepsilon^2} + O(\delta_c^2 \|w\|_2^2).$$

The rank-one part of the logit Hessian now contributes $\left(\sum_j \bar{\pi}_{ij}^{\mathrm{imbal}} \delta\xi_{ij}^{\mathrm{fr}}\right)^2 = \left(\frac{q_i^{(c)} \varphi_{\mathrm{tgt}}}{K_c m^* \varepsilon} \sum_{j \in \mathcal{P}_c} \delta_{c,j} \hat{h}_i^\top w_j\right)^2 + O(\delta_c^3 \|w\|_2^2) = O(\delta_c^2 \|w\|_2^2)$, since $\sum_{j \in \mathcal{P}_c} w_j = 0$; it is absorbed into the remainder. Multiplying by $1/N$, summing over i, and collecting terms,

$$\begin{aligned} Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) &= \frac{1}{Nm^* K_c \varepsilon^2} \sum_{i=1}^N q_i^{(c)} \sum_{j \in \mathcal{P}_c} (1 + \varphi_{\mathrm{tgt}} \delta_{c,j})(\hat{h}_i^\top w_j)^2 + O(\delta_c^2 \|w\|_2^2) \\ &= \frac{1}{Nm^* K_c \varepsilon^2} \sum_{i=1}^N q_i^{(c)} \sum_{j \in \mathcal{P}_c} (\hat{h}_i^\top w_j)^2 \\ &\quad + \frac{\varphi_{\mathrm{tgt}}}{Nm^* K_c \varepsilon^2} \sum_{i=1}^N q_i^{(c)} \sum_{j \in \mathcal{P}_c} \delta_{c,j} (\hat{h}_i^\top w_j)^2 + O(\delta_c^2 \|w\|_2^2). \end{aligned}$$

The first term is

$$\frac{1}{Nm^* \varepsilon^2 K_c} \sum_{j \in \mathcal{P}_c} w_j^\top \Sigma_c^{\mathrm{fr}} w_j.$$

For the imbalance remainder, using $|\delta_{c,j}| \le \delta_c$,

$$\left| \frac{\varphi_{\mathrm{tgt}}}{Nm^* K_c \varepsilon^2} \sum_i q_i^{(c)} \sum_{j \in \mathcal{P}_c} \delta_{c,j} (\hat{h}_i^\top w_j)^2 \right| \le \frac{\varphi_{\mathrm{tgt}} \delta_c}{Nm^* K_c \varepsilon^2} \sum_{j \in \mathcal{P}_c} w_j^\top \Sigma_c^{\mathrm{fr}} w_j.$$

Since

$$w_j^\top \Sigma_c^{\mathrm{fr}} w_j \le \lambda_{\max}(\Sigma_c^{\mathrm{fr}}) \|w_j\|_2^2,$$

we obtain

$$\left|R_{\mathrm{GN}}^{\mathrm{imbal}}(w)\right| \le \frac{\varphi_{\mathrm{tgt}} \delta_c \lambda_{\max}(\Sigma_c^{\mathrm{fr}})}{N m^* K_c \varepsilon^2} \|w\|_2^2 + O(\delta_c^2 \|w\|_2^2).$$

□

Corollary 4 (Positive definiteness of the Gauss–Newton curvature)**.** *Under the hypotheses of Theorem 3, suppose*

$$\delta_c < \frac{1}{\varphi_{\mathrm{tgt}} \kappa(\Sigma_c^{\mathrm{fr}})}, \qquad \kappa(\Sigma_c^{\mathrm{fr}}) := \frac{\lambda_{\max}(\Sigma_c^{\mathrm{fr}})}{\lambda_{\min}(\Sigma_c^{\mathrm{fr}})}.$$

Here $\lambda_{\min}$ and $\lambda_{\max}$ denote the extreme eigenvalues of Σ_c^{fr} restricted to $T_{y^*}\mathbb{S}^{d-1}$ (note $\Sigma_c^{\mathrm{fr}} y^* = 0$), and we assume $\lambda_{\min}(\Sigma_c^{\mathrm{fr}}) > 0$. Then, for sufficiently small δ_c, every $w \in V_c(y^*)$ satisfies

$$Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) \ge \frac{\mu_y^{\mathrm{GN,imbal}}}{N m^* \varepsilon^2 K_c} \|w\|_2^2,$$

where

$$\mu_y^{\mathrm{GN,imbal}} := \lambda_{\min}(\Sigma_c^{\mathrm{fr}}) \left(1 - \varphi_{\mathrm{tgt}} \delta_c \kappa(\Sigma_c^{\mathrm{fr}})\right) - O(\delta_c^2) > 0.$$

Hence the frozen-dual Gauss–Newton curvature remains positive definite on the class-centered tangent subspace under bounded relative imbalance.

Proof. By Theorem 3,

$$Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) \ge \frac{1}{N m^* \varepsilon^2 K_c} \sum_{j \in \mathcal{P}_c} w_j^\top \Sigma_c^{\mathrm{fr}} w_j - \frac{\varphi_{\mathrm{tgt}} \delta_c \lambda_{\max}(\Sigma_c^{\mathrm{fr}})}{N m^* \varepsilon^2 K_c} \|w\|_2^2 - O(\delta_c^2 \|w\|_2^2).$$

Using

$$w_j^\top \Sigma_c^{\mathrm{fr}} w_j \ge \lambda_{\min}(\Sigma_c^{\mathrm{fr}}) \|w_j\|_2^2 \qquad (w_j \in T_{y^*}\mathbb{S}^{d-1}),$$

we obtain

$$Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) \ge \frac{\lambda_{\min}(\Sigma_c^{\mathrm{fr}}) - \varphi_{\mathrm{tgt}} \delta_c \lambda_{\max}(\Sigma_c^{\mathrm{fr}}) - O(\delta_c^2)}{N m^* \varepsilon^2 K_c} \|w\|_2^2.$$

The stated condition makes the leading coefficient positive, and the $O(\delta_c^2)$ term is dominated for sufficiently small δ_c. □

Full-Hessian extension. The Gauss–Newton curvature above is the sign-definite part of the frozen-dual second variation. The full frozen-dual Hessian also contains geodesic-curvature and normalization-related mass-variation corrections. These terms are not sign-definite in general, so a full-Hessian positivity statement requires an additional small-remainder condition.

Assumption 4 (Full-Hessian remainder control)**.** *At the symmetric same-class collapse state, the non-Gauss–Newton second-order terms of the frozen-dual routing loss satisfy, for every $w \in V_c(y^*)$,*

$$|R_{\mathrm{geo}}(w) + R_{\mathrm{quot}}(w)| \le \rho_0 \frac{\lambda_{\min}(\Sigma_c^{\mathrm{fr}})}{N m^* \varepsilon^2 K_c} \|w\|_2^2$$

for some $0 \le \rho_0 < 1$. Here R_{geo} denotes the sphere geodesic-curvature correction, and R_{quot} denotes the second-order correction arising from the weighted numerator $\sum_i \hat{m}_i \mathrm{KL}_i$.

Theorem 4 (Full-Hessian curvature stability under imbalance)**.** *Retain the hypotheses of Theorem 3 and Assumption 4. Then, for every $w \in V_c(y^*)$,*

$$\left\langle w, \mathrm{Hess}_y^{\mathrm{frozen}} \mathcal{L}_{\mathrm{rout}}(y^*) w \right\rangle \ge \frac{(1 - \rho_0)\lambda_{\min}(\Sigma_c^{\mathrm{fr}}) - \varphi_{\mathrm{tgt}} \delta_c \lambda_{\max}(\Sigma_c^{\mathrm{fr}}) - O(\delta_c^2)}{N m^* \varepsilon^2 K_c} \|w\|_2^2.$$

Consequently, the frozen-dual full Hessian is positive definite on $V_c(y^*)$ whenever

$$\rho_0 + \varphi_{\mathrm{tgt}} \delta_c \kappa(\Sigma_c^{\mathrm{fr}}) + O(\delta_c^2) < 1.$$

Proof. Decompose the full frozen-dual second variation as

$$\Big\langle w, \mathrm{Hess}_y^{\mathrm{frozen}}\, \mathcal{L}_{\mathrm{rout}}(y^*)w \Big\rangle = Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) + R_{\mathrm{geo}}(w) + R_{\mathrm{quot}}(w).$$

By Theorem 3,

$$Q_{\mathrm{GN}}^{\mathrm{fr,imbal}}(\mathcal{L}_{\mathrm{rout}})(w) \geq \frac{\lambda_{\min}(\Sigma_c^{\mathrm{fr}}) - \varphi_{\mathrm{tgt}}\delta_c\lambda_{\max}(\Sigma_c^{\mathrm{fr}}) - O(\delta_c^2)}{Nm^*\varepsilon^2 K_c}\|w\|_2^2.$$

By Assumption 4,

$$R_{\mathrm{geo}}(w) + R_{\mathrm{quot}}(w) \geq -\rho_0 \frac{\lambda_{\min}(\Sigma_c^{\mathrm{fr}})}{Nm^*\varepsilon^2 K_c}\|w\|_2^2.$$

Combining the two inequalities gives the stated lower bound. The positivity condition follows after factoring out $\lambda_{\min}(\Sigma_c^{\mathrm{fr}})$. □

Remark 8 (Sufficient conditions for small full-Hessian remainder)**.** *Assumption 4 is satisfied, for example, when the target mismatch and mass-variation second-order effects are small relative to the covariance curvature. A typical sufficient regime is*

$$\max_i \sum_{j\in\mathcal{P}_c} |\bar{\pi}_{ij}^* - \bar{T}_{ij}| \quad \textit{small},$$

together with bounded target-side second-order variation on $V_c(y^*)$. In this regime, the geodesic and target-side corrections are lower-order compared with the positive Gauss–Newton covariance term.

Structural remarks.

Remark 9 (Dependence on the $1/(Nm^*)$ prefactor)**.** *Under equal bag sizes, the routing objective contributes a global factor $1/(Nm^*)$ to the curvature scale. Accordingly, the imbalance tolerance depends on Σ_c^{fr}, φ_{tgt}, and ε, while the overall curvature magnitude scales as $1/(Nm^*\varepsilon^2 K_c)$.*

Remark 10 (Weighted-proxy pullback)**.** *The curvature statements above are formulated at the proxy-variable level. By the differential of Proposition 1, the proxy-level curvature pulls back through two weighted-proxy routes: the scorer path through the entmax Jacobian and the encoder path through the directly weighted feature differentials. No fixed-support stratum is required.*

D.3 Trajectory-level non-persistence of same-class collapse under $\exp(m_i)$-weighted routing

This subsection shows that symmetric same-class proxy collapse at non-degenerate collapse states (the set $\mathcal{A}_c$ of Assumption 5) cannot persist along the gradient trajectory under the implemented exponentially normalized routing surrogate. Three statements are established with explicit constants and for sufficiently small fixed step sizes: (i) sufficiently deep near-collapse states around $\mathcal{A}_c$ are escaped in one step, so collapse states in $\mathcal{A}_c$ are not fixed points; (ii) no trajectory converges to a collapse state whose non-degeneracy level exceeds σ_z; and (iii) along the whole trajectory, the same-class diversity $D_c(t)$ can fall below a fixed threshold around $\mathcal{A}_c$ only finitely many times. Degenerate collapse states, at which the anti-collapse field vanishes, are discussed in Remark 11. The argument uses the exact class-relative field at collapse from Appendices C.1 and C.2.

Setup. We retain all notation from Section 4 and Appendix C.1, and write $y_j := \bar{p}_j$. As in Appendices C.1–C.2, the analysis is carried out at the proxy level. The normalized instance embeddings $\{\bar{h}_i\}$ are held fixed, and all proxies $y = (y_j)_{j=1}^M \in (\mathbb{S}^{d-1})^M$ are updated by fixed-step Riemannian gradient descent on the exact grounding loss $f(y) := \mathcal{L}_{\mathrm{grnd}}(y)$, whose UOT plan is the coupled solution of Lemma 4:

$$y_j(t+1) = \frac{y_j(t) - \eta\, g_j(t)}{\|y_j(t) - \eta\, g_j(t)\|_2}, \qquad g_j(t) := \mathrm{grad}_j\, f(y(t)) = \Pi_{y_j(t)}\nabla_{y_j} f(y(t)), \tag{7}$$

for $j = 1, \ldots, M$ (projection retraction). Fix a class c with $K_c \geq 2$ and $\gamma < 1$, with symmetric same-class target priors. Recall the same-class proxy diversity measure

$$D_c(t) := \frac{2}{K_c(K_c-1)} \sum_{\substack{j<k \\ j,k\in\mathcal{P}_c}} \big(1 - y_j(t)^\top y_k(t)\big) = \frac{1}{K_c(K_c-1)} \sum_{\substack{j<k \\ j,k\in\mathcal{P}_c}} \|y_j(t) - y_k(t)\|_2^2.$$

Define the class mean $\bar{g}_c(y) := \frac{1}{K_c}\sum_{j\in\mathcal{P}_c} g_j(y)$ and the class-relative field

$$\tilde{F}_{c,j}(y) := \Pi_{y_j}\big(g_j(y) - \bar{g}_c(y)\big).$$

Let $\mathcal{C}_c^{\mathrm{col}} := \{y : \ y_j = y_k \ \forall j,k \in \mathcal{P}_c\}$ be the (compact) set of symmetric same-class collapse states of class c, and let $G := \max_y \max_j \|g_j(y)\|_2$, which is finite by compactness of $(\mathbb{S}^{d-1})^M$.

Assumption 5 (Non-degenerate collapse states)**.** *For a symmetric same-class collapse state $y \in \mathcal{C}_c^{\mathrm{col}}$ with common direction y_c (and arbitrary other-class proxies), define its non-degeneracy level $\sigma(y) \geq 0$ by*

$$\sigma(y)^2 := \frac{1}{K_c}\sum_{j\in\mathcal{P}_c}\left\|\Pi_{y_c}\left(\tilde{\mu}_{c,j}^{(\exp)}(y) - \frac{1}{K_c}\sum_{\ell\in\mathcal{P}_c}\tilde{\mu}_{c,\ell}^{(\exp)}(y)\right)\right\|_2^2,$$

$$\tilde{\mu}_{c,j}^{(\exp)}(y) := \frac{\sum_{i\in\mathcal{C}_{c,j}}\exp(m_i)\bar{h}_i}{\sum_{i\in\mathcal{C}_{c,j}}\exp(m_i)},$$

where m_i is evaluated at y. Fix $\sigma_z > 0$. We assume that the set of σ_z-non-degenerate collapse states

$$\mathcal{A}_c := \{\, y \in \mathcal{C}_c^{\mathrm{col}} : \ \sigma(y) \geq \sigma_z \,\}$$

is nonempty.

Remark 11. *Assumption 5 formalizes the heterogeneous-bag premise at the collapse states under consideration: distinct same-class bags have distinguishable exponentially weighted instance compositions in the proxy tangent space. The class-level reference is the average of the bag-level means, $\bar{\tilde{\mu}}_c^{(\exp)} := \frac{1}{K_c}\sum_{\ell\in\mathcal{P}_c}\tilde{\mu}_{c,\ell}^{(\exp)}$. Because $\sum_{i:o(i)=b}\hat{m}_i = 1/M$ for every bag (Eq. (5)), this is exactly the mean that appears in $\tilde{w}_j$, and $\sigma(y)^2 = \frac{M^2}{K_c}\sum_j\|\tilde{w}_j\|_2^2$ (Lemma 10). Since m_i depends continuously on the proxies (Lemma 4), $\sigma(\cdot)$ is continuous and $\mathcal{A}_c$ is compact.*

A collapse state is degenerate, $\sigma(y) = 0$, exactly when every bag-mean deviation $\tilde{\mu}_{c,j}^{(\exp)} - \frac{1}{K_c}\sum_\ell\tilde{\mu}_{c,\ell}^{(\exp)}$ is parallel to y_c (or zero). Since σ is continuous on the compact set $\mathcal{C}_c^{\mathrm{col}}$, $\mathcal{A}_c \neq \emptyset$ if and only if $\sigma_z \leq \max_y\sigma(y)$, i.e., if and only if some non-degenerate collapse state exists and σ_z is chosen no larger than its level. This excludes, for instance, the case in which all same-class bag means coincide, where $\tilde{w}_j \equiv 0$ and no anti-collapse field exists.

For $K_c \geq 3$, degeneracy cannot occur at a state where two of the deviations are linearly independent. For $K_c = 2$, the two deviations are opposite, so degeneracy reduces to the vanishing of one tangent vector field $y_c \mapsto \Pi_{y_c}(\tilde{\mu}_{c,1}^{(\exp)} - \bar{\tilde{\mu}}_c^{(\exp)})$ on $\mathbb{S}^{d-1}$ (for fixed other-class proxies). When d is odd, the hairy-ball theorem guarantees at least one zero. When d is even, zeros are not topologically forced in general. They are still forced by a degree argument whenever the class-c embeddings span a proper subspace of $\mathbb{R}^d$, or the deviation varies little with y_c. Hence uniform non-degeneracy cannot be assumed without additional structure.

At a degenerate state, $\tilde{w}_j = 0$ for all j, so by Lemma 10 all same-class gradients coincide. A gradient step therefore maps the collapse state to another collapse state: collapse is preserved for that step, and the present argument provides no anti-collapse guarantee. Stationary collapse can also occur: if all same-class bags are identical, then $\tilde{w}_j \equiv 0$ on $\mathcal{C}_c^{\mathrm{col}}$, so all same-class gradients coincide there, and any minimizer of f over the compact set $\mathcal{C}_c^{\mathrm{col}}$ is a critical point of f, hence a fixed point of (7)*. Restricting to $\mathcal{A}_c$ is therefore necessary for the one-step escape in Theorem 5(i) rather than an artifact of the proof. As $\sigma_z \downarrow 0$, $\mathcal{A}_c$ exhausts all non-degenerate collapse states.*

Assumption 6 (Descent inequality on the proxy manifold)**.** *There is $L_R < \infty$ such that, for all $y \in (\mathbb{S}^{d-1})^M$ and all tangent v,*

$$f(R_y(v)) \leq f(y) + \langle \operatorname{grad} f(y), v\rangle + \frac{L_R}{2}\|v\|_2^2,$$

where R is the projection retraction in (7)*.*

This holds because f is C^2 (Lemma 9) and $(\mathbb{S}^{d-1})^M$ is a compact submanifold. Moreover, f is bounded below by some $f_\star > -\infty$: $\mathcal{L}_{\mathrm{rout}}$, $\mathcal{R}_s$, and $\mathcal{R}_t$ are nonnegative, and each entry of $\mathcal{R}_e$ is at least -1.

Preparatory lemmas.

Lemma 9 (Gradient field Lipschitz continuity)**.** *The loss f is C^2 on a neighborhood of $(\mathbb{S}^{d-1})^M$, and there exists $L_g < \infty$ such that, in ambient coordinates,*

$$\|g(y) - g(y')\|_2 \le L_g \|y - y'\|_2 \qquad \forall\, y, y' \in (\mathbb{S}^{d-1})^M,$$

where $g(y) = (g_j(y))_{j=1}^M$ is the stacked Riemannian gradient.

Proof. The cost $C_{ij} = 1 - \bar{h}_i^\top y_j$ is linear in y, and by Lemma 4 the entropic UOT plan and dual variables are C^2 functions of the cost. On $(\mathbb{S}^{d-1})^M$ we have $C_{ij} \in [0, 2]$, so $K_{ij} \in [e^{-2/\varepsilon}, 1]$. The plan entries, and hence m_i and $\bar{\pi}_{ij}$, are therefore bounded away from zero, and $\mathcal{L}_{\mathrm{rout}}$, $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ are C^∞ in the plan there. Consequently f is C^2 on a neighborhood of $(\mathbb{S}^{d-1})^M$, and $g_j(y) = \Pi_{y_j} \nabla_{y_j} f(y)$ is C^1. A C^1 map is locally Lipschitz, and a locally Lipschitz map is Lipschitz on a compact set. □

Lemma 10 (Exact-collapse field bounds)**.** *At every $y^* \in \mathcal{C}_c^{\mathrm{col}}$ with common direction y_c^*, and for all $j \in \mathcal{P}_c$,*

$$\tilde{F}_{c,j}(y^*) = -\tilde{\beta}_c \tilde{w}_j, \qquad \|\tilde{w}_j\|_2 \le \frac{2}{M}, \qquad \tilde{\beta}_c = \frac{1-\gamma}{\varepsilon(1 + (K_c - 1)\gamma)}.$$

If moreover $y^* \in \mathcal{A}_c$, then $\frac{1}{K_c} \sum_{\ell \in \mathcal{P}_c} \|\tilde{w}_\ell\|_2^2 \ge \sigma_z^2 / M^2$, and consequently

$$\sum_{j \in \mathcal{P}_c} \|\tilde{F}_{c,j}(y^*)\|_2^2 \;\ge\; b_c := \frac{K_c\, \tilde{\beta}_c^2 \sigma_z^2}{M^2} > 0, \qquad \sum_{j \in \mathcal{P}_c} \|g_j(y^*)\|_2^2 \;\ge\; b_c.$$

Proof. By the per-bag normalization of $\hat{m}_i$ (Eq. (5)), $\sum_{i:o(i)=j} \hat{m}_i = 1/M$ for every bag, so $M_{c,j} = \bar{M}_c$. The target-dual correction in Corollary 3 therefore vanishes, and $\tilde{F}_{c,j}(y^*) = -\tilde{\beta}_c \tilde{w}_j$. This identity holds at every collapse state; only the lower bound below uses $y^* \in \mathcal{A}_c$. Moreover, $\hat{m}_i \propto \exp(m_i)$ within each bag, so $\sum_{i \in \mathcal{C}_{c,j}} \hat{m}_i \bar{h}_i = \tilde{\mu}_{c,j}^{(\exp)} / M$. This gives

$$\tilde{w}_j = \frac{1}{M} \Pi_{y_c^*} \left(\tilde{\mu}_{c,j}^{(\exp)} - \bar{\tilde{\mu}}_c^{(\exp)} \right).$$

The lower bound on $\frac{1}{K_c} \sum_\ell \|\tilde{w}_\ell\|_2^2$ is Assumption 5, since $\frac{1}{K_c} \sum_\ell \|\tilde{w}_\ell\|_2^2 = \sigma(y^*)^2 / M^2$. For the upper bound, orthogonal projection is contractive, and $\tilde{\mu}_{c,j}^{(\exp)}$ and $\bar{\tilde{\mu}}_c^{(\exp)}$ are convex combinations of unit vectors, so each has norm at most 1.

Summing $\|\tilde{F}_{c,j}\|_2^2 = \tilde{\beta}_c^2 \|\tilde{w}_j\|_2^2$ over the K_c proxies gives the first consequence. For the second, at collapse all $g_j(y^*)$ lie in $T_{y_c^*} \mathbb{S}^{d-1}$, so $\tilde{F}_{c,j}(y^*) = g_j(y^*) - \bar{g}_c(y^*)$. A sum of squared deviations from the mean is at most the sum of squares, so $\sum_j \|g_j\|_2^2 \ge \sum_j \|g_j - \bar{g}_c\|_2^2 \ge b_c$. □

Lemma 11 (Diversity controls the distance to collapse)**.** *Let y satisfy $D_c(y) \le 1$, and let $\bar{y} := \frac{1}{K_c} \sum_{j \in \mathcal{P}_c} y_j$. Define $y^* \in \mathcal{C}_c^{\mathrm{col}}$ to agree with y except that $y_j^* := \bar{y} / \|\bar{y}\|_2$ for $j \in \mathcal{P}_c$. Then*

$$\|y - y^*\|_2 \le 2\sqrt{(K_c - 1)\, D_c(y)}.$$

Proof. From the second expression for D_c and $\sum_{j<k} \|y_j - y_k\|_2^2 = K_c \sum_j \|y_j - \bar{y}\|_2^2$, we get $\sum_j \|y_j - \bar{y}\|_2^2 = (K_c - 1) D_c(y)$. Hence $\|\bar{y}\|_2^2 = 1 - \frac{1}{K_c} \sum_j \|y_j - \bar{y}\|_2^2 = 1 - \frac{K_c - 1}{K_c} D_c(y) > 0$. Since $1 - \|\bar{y}\|_2 \le 1 - \|\bar{y}\|_2^2$,

$$\begin{aligned} \|y - y^*\|_2 &\le \Big(\sum_j \|y_j - \bar{y}\|_2^2 \Big)^{1/2} + \sqrt{K_c} \left(1 - \|\bar{y}\|_2\right) \\ &\le \sqrt{(K_c - 1) D_c} + \tfrac{K_c - 1}{\sqrt{K_c}} D_c \le 2\sqrt{(K_c - 1) D_c}, \end{aligned}$$

where the last step uses $(K_c - 1) D_c \le K_c$. □

Main theorem.

Theorem 5 (Trajectory-level non-persistence of same-class collapse under $\exp(m_i)$-weighted routing)**.** *Suppose $\gamma < 1$, $K_c \geq 2$, the target priors are symmetric within class c, and Assumption 5 holds. For y with $D_c(y) \leq 1$, let $y^*(y) \in \mathcal{C}_c^{\mathrm{col}}$ denote the collapse state of Lemma 11; conditions on $y^*(y(t))$ below are imposed only together with $D_c(t) \leq 1$. Let b_c, G, and L_g be as above, and define*

$$a_c := \frac{3\, b_c}{32(K_c-1)}, \qquad r_c := \frac{b_c}{32(K_c-1)}, \qquad \eta_0^2 := \min\Big\{\frac{1}{G^2},\ \frac{1}{4(K_c-1)L_g^2},\ \frac{b_c}{(K_c-1)G^4}\Big\}.$$

Then the iterates of (7) *satisfy the following.*

(i) ***One-step escape.*** *For every $0 < \eta \leq \eta_0$ and every t satisfying*

$$D_c(t) \leq r_c\eta^2 \quad \textit{and} \quad y^*(y(t)) \in \mathcal{A}_c,$$

one has $D_c(t+1) - D_c(t) \geq a_c\eta^2$. (Here $r_c\eta^2 \leq 1/16$, so $y^*(y(t))$ is well defined.) In particular, a symmetric collapse state in $\mathcal{A}_c$ is escaped in one step: if $y(t) \in \mathcal{A}_c$, then $D_c(t+1) \geq a_c\eta^2 > 0$.

(ii) ***No convergence to non-degenerate collapse.*** *For $0 < \eta \leq \eta_0$, no trajectory converges to a symmetric same-class collapse state $\hat{y}$ with $\sigma(\hat{y}) > \sigma_z$.*

(iii) ***Trajectory-level diversity.*** *Suppose in addition that Assumption 6 holds, and let*

$$\vartheta_c := \min\Big\{1,\ \frac{b_c}{16(K_c-1)L_g^2}\Big\}, \qquad T_c := \frac{8\,\big(f(y(0)) - f_\star\big)}{\eta\, b_c}.$$

Then, for every $0 < \eta \leq 1/L_R$,

$$\#\{t \geq 0 :\ D_c(t) \leq \vartheta_c \textit{ and } y^*(y(t)) \in \mathcal{A}_c\} \ \leq\ T_c,$$

that is, around non-degenerate collapse states, $D_c(t) > \vartheta_c$ for all but at most T_c iterations.

Proof. We establish an exact one-step expression, compare it with the collapse configuration, and then collect constants. Throughout, $\Delta y_{jk} := y_j - y_k$ and $\Delta g_{jk} := g_j - g_k$ for $j, k \in \mathcal{P}_c$, evaluated at iteration t.

Step 1: Exact one-step expression. Let $z_j := y_j - \eta g_j$ and $A_j := \|z_j\|_2^2 = 1 + \eta^2\|g_j\|_2^2$, using $g_j \perp y_j$. Then $y_j(t+1) = z_j/\sqrt{A_j}$. From $\|z_j - z_k\|_2^2 = A_j + A_k - 2z_j^\top z_k$,

$$1 - y_j(t+1)^\top y_k(t+1) = \frac{\|z_j - z_k\|_2^2 - (\sqrt{A_j} - \sqrt{A_k})^2}{2\sqrt{A_jA_k}}.$$

This quantity is nonnegative. Moreover, $(\sqrt{A_j} - \sqrt{A_k})^2 = (A_j - A_k)^2/(\sqrt{A_j} + \sqrt{A_k})^2 \leq \eta^4G^4/4$ and $\sqrt{A_jA_k} \leq 1 + \eta^2G^2$. Summing over the $K_c(K_c-1)/2$ pairs gives

$$D_c(t+1) \ \geq\ \frac{\sum_{j<k}\|\Delta y_{jk} - \eta\,\Delta g_{jk}\|_2^2 - K_c(K_c-1)\eta^4G^4/8}{K_c(K_c-1)\,(1+\eta^2G^2)}. \tag{8}$$

No higher-order remainder is involved.

Step 2: Comparison with the collapse configuration. Assume $D_c(t) \leq 1$ and $y^*(y(t)) \in \mathcal{A}_c$, and let $y^* = y^*(y(t))$ be the collapse state of Lemma 11, so that Lemma 10 applies at y^*. Write $\Delta g_{jk}^* := g_j(y^*) - g_k(y^*)$. Since all $g_j(y^*)$ lie in the same tangent space, the class mean cancels, and Lemma 10 gives $\Delta g_{jk}^* = \tilde{F}_{c,j}(y^*) - \tilde{F}_{c,k}(y^*)$ with $\sum_j \tilde{F}_{c,j}(y^*) = 0$. Hence

$$\sum_{j<k}\|\Delta g_{jk}^*\|_2^2 = K_c\sum_j\|\tilde{F}_{c,j}(y^*)\|_2^2 \ \geq\ K_c\, b_c.$$

By Lemmas 9 and 11,

$$\|\Delta g_{jk} - \Delta g_{jk}^*\|_2 \leq \sqrt{2}\,\|g(y) - g(y^*)\|_2 \leq \sqrt{2}\,L_g\|y - y^*\|_2,$$
$$\|\Delta g_{jk} - \Delta g_{jk}^*\|_2^2 \leq 8(K_c-1)L_g^2\, D_c(t).$$

Apply $\|u+v\|_2^2 \geq \frac{1}{2}\|u\|_2^2 - \|v\|_2^2$ with $u = -\eta\Delta g_{jk}^*$ and $v = \Delta y_{jk} - \eta(\Delta g_{jk} - \Delta g_{jk}^*)$. Summing over pairs and using $\sum_{j<k}\|\Delta y_{jk}\|_2^2 = K_c(K_c-1)D_c(t)$,

$$\sum_{j<k}\|\Delta y_{jk} - \eta\Delta g_{jk}\|_2^2 \;\geq\; \frac{\eta^2 K_c b_c}{2} - 2K_c(K_c-1)\,D_c(t)\big(1+4(K_c-1)\eta^2 L_g^2\big).$$

The first-order (in η) mismatch between the current and collapsed configurations is thus kept explicitly as the $D_c(t)$ term rather than absorbed into a higher-order remainder.

Step 3: Constants for (i). Let $B := b_c/(2(K_c-1))$, and suppose $D_c(t) \leq r_c\eta^2$ with $r_c = B/16$ and $0 < \eta \leq \eta_0$.

- The condition $\eta^2 \leq 1/(4(K_c-1)L_g^2)$ bounds the $D_c(t)$ term in Step 2, after division by $K_c(K_c-1)$, by $4r_c\eta^2 = B\eta^2/4$.
- The condition $\eta^2 \leq b_c/((K_c-1)G^4) = 2B/G^4$ gives $\eta^4G^4/8 \leq B\eta^2/4$.
- The condition $\eta^2G^2 \leq 1$ gives $1+\eta^2G^2 \leq 2$.

Combining these with (8),

$$D_c(t+1) \;\geq\; \frac{B\eta^2 - B\eta^2/4 - B\eta^2/4}{2} = \frac{B\eta^2}{4},$$

and therefore $D_c(t+1) - D_c(t) \geq B\eta^2/4 - B\eta^2/16 = \frac{3B}{16}\eta^2 = a_c\eta^2$. The requirement $D_c(t) \leq 1$ of Lemma 11 holds automatically, because $b_c \leq K_cG^2$ (Lemma 10) gives $r_c\eta_0^2 \leq K_c/(32(K_c-1)) \leq 1/16$.

Step 4: Proof of (ii). Suppose $y(t) \to \hat{y} \in \mathcal{C}_c^{\mathrm{col}}$ with $\sigma(\hat{y}) > \sigma_z$. Then $D_c(t) \to 0$. The map $y \mapsto y^*(y)$ is continuous on $\{D_c < 1\}$ with $y^*(\hat{y}) = \hat{y}$, so $y^*(y(t)) \to \hat{y}$. Since σ is continuous, $\sigma(y^*(y(t))) \geq \sigma_z$ for all large t. So for some T, $D_c(t) \leq r_c\eta^2$ and $y^*(y(t)) \in \mathcal{A}_c$ for all $t \geq T$, and by (i) $D_c(t) \geq D_c(T) + (t-T)a_c\eta^2 \to \infty$, which contradicts $D_c \leq 2$.

Step 5: Proof of (iii). Suppose $D_c(t) \leq \vartheta_c$ and $y^*(y(t)) \in \mathcal{A}_c$. By Lemma 11, $\|y(t) - y^*\|_2 \leq 2\sqrt{(K_c-1)\vartheta_c} \leq \sqrt{b_c}/(2L_g)$. By Lemmas 9 and 10,

$$\|g(y(t))\|_2 \;\geq\; \|g(y^*)\|_2 - L_g\|y(t)-y^*\|_2 \;\geq\; \sqrt{b_c} - \sqrt{b_c}/2 = \sqrt{b_c}/2.$$

By Assumption 6 with $v = -\eta g(y(t))$ and $\eta \leq 1/L_R$, every iteration satisfies $f(y(t+1)) \leq f(y(t)) - \frac{\eta}{2}\|g(y(t))\|_2^2$. So every iteration counted in (iii) decreases f by at least $\eta b_c/8$, and no iteration increases it. Since $f \geq f_\star$, there are at most $8(f(y(0)) - f_\star)/(\eta b_c) = T_c$ such iterations. □

Remark 12 (Robustness to bag-size and prior-mass imbalance)**.** *Because $\hat{m}_i$ is normalized within each bag, $M_{c,j} = \sum_{i\in\mathcal{C}_{c,j}}\hat{m}_i = 1/M$ regardless of the bag sizes N_j and the source prior masses μ_i. Hence $\tilde{w}_j$, b_c, and the constants a_c and r_c in Theorem 5 depend on the bags only through their exponentially weighted means $\tilde{\mu}_{c,j}^{(\exp)}$. No additional imbalance term arises. In contrast, for row weights without per-bag normalization, a mass imbalance $M_{c,j} \neq \bar{M}_c$ enters both $\tilde{w}_j$ and the target-dual correction of Theorem 2.*

Remark 13 (Why a time-uniform lower bound on $D_c(t)$ is not claimed)**.** *Theorem 5 is a non-persistence result, not a time-uniform separation bound $D_c(t) \geq \tilde{D}_{\min}$ for all t. Such a bound can fail, because by* (8) *the one-step change contains the cross term $-2\eta\,\Delta y_{jk}^\top\Delta g_{jk}$. This term is first order in η and negative whenever two proxies are displaced opposite to their separating direction. Such proxies first move toward each other and then separate. In a small numerical example with $K_c = 2$, D_c decreases from 5.0×10^{-3} to 3.6×10^{-6} and then grows to 0.31. Items (i)–(iii) are the trajectory-level statements that remain valid. They concern non-degenerate collapse states ($\mathcal{A}_c$): discrete iterates may transiently approach such states from outside the $O(\eta)$ neighborhood $\{D_c \leq r_c\eta^2\}$, but they cannot remain near them, cannot converge to a collapse state $\hat{y}$ with $\sigma(\hat{y}) > \sigma_z$, and can enter the fixed neighborhood $\{D_c \leq \vartheta_c\}$ around $\mathcal{A}_c$ only finitely often. No such statement is made for degenerate collapse states (Remark 11).*

Remark 14 (Relation to the exact-collapse identity)**.** *If $D_c(t) = 0$, then $\Delta y_{jk} = 0$, and the identity of Step 1 gives*

$$D_c(t+1) = \frac{\eta^2}{K_c - 1} \sum_{j \in \mathcal{P}_c} \|\tilde{F}_{c,j}(y^*)\|_2^2 + O(\eta^4) = \frac{\eta^2 \tilde{\beta}_c^2}{K_c - 1} \sum_{j \in \mathcal{P}_c} \|\tilde{w}_j\|_2^2 + O(\eta^4).$$

This recovers the one-step de-collapse formula of Proposition 3 at the level of diversity. The constants a_c and r_c in Theorem 5 are conservative: the factors $3/32$ and $1/32$ come from the elementary inequality $\|u+v\|^2 \geq \frac{1}{2}\|u\|^2 - \|v\|^2$ and are not optimized.

Implemented gradients. *The implementation stops gradients through the Sinkhorn scaling updates (Appendix F.1). At a symmetric collapse state, the class-relative part of this stop-gradient field also equals $-\tilde{\beta}_c \tilde{w}_j$: the routing term gives Proposition 3, and the derivatives of $\hat{m}_i$, $\mathcal{R}_s$, $\mathcal{R}_t$, and $\mathcal{R}_e$ through the kernel are identical across $j \in \mathcal{P}_c$. Items (i)–(ii) of Theorem 5 use only this value at collapse and the Lipschitz continuity of the field, so they hold verbatim for the implemented field (with G and L_g computed for it). Item (iii) additionally uses the descent inequality and applies to the exact gradient.*

Remark 15 (Dependence of the constants on problem parameters)**.** *The constants in Theorem 5 depend on the problem through:*

- *(i) $\tilde{\beta}_c = (1-\gamma)/(\varepsilon(1 + (K_c - 1)\gamma))$: larger for smaller γ and smaller ε;*
- *(ii) σ_z/M: the (root-mean-square) non-degeneracy level of the collapse states under consideration, i.e., the tangential spread of the bag-level exponentially weighted means, reflecting both instance heterogeneity and UOT mass retention;*
- *(iii) $b_c = K_c \tilde{\beta}_c^2 \sigma_z^2 / M^2$: the escape increment a_c and the deep-collapse radius r_c scale linearly with b_c, and the diversity threshold ϑ_c scales as b_c/L_g^2;*
- *(iv) G and L_g: the gradient magnitude and gradient-field Lipschitz constant of the implemented routing surrogate, which limit the admissible step size η_0.*

In particular, $a_c, r_c, \vartheta_c \to 0$ and $T_c \to \infty$ as $\gamma \to 1$, consistent with Proposition 3.

D.4 Parameter-level anti-collapse under the weighted proxy pullback

This subsection explains when the proxy-level anti-collapse field survives a full parameter-gradient step through the weighted proxy map.

Setup. Fix a class c and write

$$Y_c(\Theta) := (y_j(\Theta))_{j \in \mathcal{P}_c}, \qquad y_j(\Theta) := \bar{p}_j(\Theta).$$

At a symmetric same-class collapse state $Y_c^* := (y^*, \ldots, y^*)$, define the class-relative tangent subspace

$$\mathcal{T}_c^{\mathrm{rel}}(y^*) := \Big\{ v = (v_j)_{j \in \mathcal{P}_c} : v_j \in T_{y^*}\mathbb{S}^{d-1},\ \sum_{j \in \mathcal{P}_c} v_j = 0 \Big\},$$

and let P_c denote the orthogonal projector onto $\mathcal{T}_c^{\mathrm{rel}}(y^*)$. Let

$$J_Y(\Theta) := D_\Theta Y_c(\Theta)$$

be the weighted-proxy Jacobian restricted to the class-c proxy block, and, absorbing λ_{grnd} into $\mathcal{L}_{\mathrm{rout}}$ (and hence into $\tilde{\beta}_c$) as in Appendix E.1, decompose the full parameter gradient as

$$\nabla_\Theta F(\Theta) = J_Y(\Theta)^\top \nabla_Y \mathcal{L}_{\mathrm{rout}}(\Theta) + r_\Theta(\Theta),$$

where

$$r_\Theta(\Theta) := \nabla_\Theta F(\Theta) - J_Y(\Theta)^\top \nabla_Y \mathcal{L}_{\mathrm{rout}}(\Theta)$$

collects all remaining parameter-gradient contributions. In particular, r_Θ absorbs the bag-loss gradient, the non-routing grounding terms, and any routing-gradient components not mediated by the class-c proxy block Y_c; it is a residual term, not necessarily an orthogonal one. For one gradient-descent step $\Theta^+ = \Theta - \eta \nabla_\Theta F(\Theta)$, define the induced proxy displacement

$$\delta Y := Y_c(\Theta^+) - Y_c(\Theta).$$

Assumption 7 (Proxy pullback non-degeneracy and relative-subspace invariance)**.** *In a neighborhood of a non-degenerate collapse state (one in $\mathcal{A}_c$ of Assumption 5; at degenerate states $P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}=0$ and the last condition below cannot hold), the weighted proxy map is C^2, and there exists $\mu_p>0$ such that*

$$\langle v, J_Y(\Theta)J_Y(\Theta)^\top v\rangle \geq \mu_p\|v\|_2^2 \qquad \forall v\in\mathcal{T}_c^{\mathrm{rel}}(y^*).$$

Assume in addition that the pullback operator preserves the class-relative tangent subspace:

$$J_Y(\Theta)J_Y(\Theta)^\top v\in\mathcal{T}_c^{\mathrm{rel}}(y^*) \qquad \forall v\in\mathcal{T}_c^{\mathrm{rel}}(y^*).$$

Assume additionally that there exists $\sigma_c>0$ such that

$$\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}(\Theta)\|_2\geq\tilde{\beta}_c\sigma_c$$

throughout the same neighborhood.

Theorem 6 (Parameter-level anti-collapse under weighted proxy pullback)**.** *Assume Assumption 7, together with the additional C^2 regularity of $h_{b,n}$ and $s_{b,n}$ and the Lipschitz-Jacobian assumption on entmax used in Lemma 3. Suppose also that the residual term obeys*

$$\|P_cJ_Y(\Theta)r_\Theta(\Theta)\|_2\leq\frac{1}{2}\mu_p\tilde{\beta}_c\sigma_c$$

throughout the same neighborhood. Then there exist constants $C_{\mathrm{pb}}<\infty$ and $\eta_0>0$ such that, for every $0<\eta\leq\eta_0$,

$$\langle-\delta Y, P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}(\Theta)\rangle\geq\frac{1}{2}\eta\mu_p\tilde{\beta}_c^2\sigma_c^2-C_{\mathrm{pb}}\eta^2.$$

In particular, for sufficiently small η, the projected proxy displacement has positive first-order inner product with the proxy-level anti-collapse field. Hence, to first order in η, a parameter-gradient step moves the class-c proxies along the proxy-level anti-collapse direction $-P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}$, which equals $(\tilde{\beta}_c\tilde{w}_j)_{j\in\mathcal{P}_c}$ at collapse.

Proof. Under the additional C^2 and Lipschitz-Jacobian assumptions above, Lemma 3 gives the second-order control needed for the weighted proxy map. Taylor expansion therefore gives

$$\delta Y=-\eta J_YJ_Y^\top\nabla_Y\mathcal{L}_{\mathrm{rout}}(\Theta)-\eta J_Yr_\Theta(\Theta)+R_\eta,\qquad\|R_\eta\|_2\leq C_{\mathrm{pb}}\eta^2.$$

Taking the inner product with $P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}(\Theta)$ yields

$$\langle-\delta Y,P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\rangle=\eta\langle\nabla_Y\mathcal{L}_{\mathrm{rout}},J_YJ_Y^\top P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\rangle+\eta\langle P_c\nabla_Y\mathcal{L}_{\mathrm{rout}},P_cJ_Yr_\Theta\rangle+O(\eta^2).$$

By the relative-subspace invariance part of Assumption 7, the vector $J_YJ_Y^\top P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}$ lies in $\mathcal{T}_c^{\mathrm{rel}}(y^*)$. Hence

$$\langle\nabla_Y\mathcal{L}_{\mathrm{rout}},J_YJ_Y^\top P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\rangle=\langle P_c\nabla_Y\mathcal{L}_{\mathrm{rout}},J_YJ_Y^\top P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\rangle.$$

The pullback non-degeneracy assumption therefore gives

$$\langle P_c\nabla_Y\mathcal{L}_{\mathrm{rout}},J_YJ_Y^\top P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\rangle\geq\mu_p\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2^2\geq\mu_p\tilde{\beta}_c^2\sigma_c^2.$$

Using the residual bound together with Cauchy–Schwarz gives

$$\langle-\delta Y,P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\rangle\geq\eta\mu_p\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2^2-\eta\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2\,\|P_cJ_Yr_\Theta\|_2-C_{\mathrm{pb}}\eta^2.$$

Since $\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}(\Theta)\|_2\geq\tilde{\beta}_c\sigma_c$ by Assumption 7 and $\|P_cJ_Y(\Theta)r_\Theta(\Theta)\|_2\leq\frac{1}{2}\mu_p\tilde{\beta}_c\sigma_c$, we obtain

$$\begin{aligned}&\eta\mu_p\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2^2-\eta\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2\,\|P_cJ_Yr_\Theta\|_2\\&\qquad\geq\eta\mu_p\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2\left(\|P_c\nabla_Y\mathcal{L}_{\mathrm{rout}}\|_2-\frac{\tilde{\beta}_c\sigma_c}{2}\right)\geq\frac{1}{2}\eta\mu_p\tilde{\beta}_c^2\sigma_c^2.\end{aligned}$$

Combining these estimates with the Taylor remainder proves the claim. □

Remark 16 (Connection to proxy-level separation)**.** *Theorem 6 complements the proxy-level result of Theorem 5. The proxy-level theorem identifies the separating class-relative direction, while the pullback non-degeneracy assumption ensures that the weighted proxy Jacobian does not annihilate that direction when the dynamics are viewed in parameter space.*

E Optimization Theory

E.1 Full-batch descent on compact trajectory neighborhoods

Setup. Let $\Theta = (\theta, \phi, \psi)$ denote all trainable parameters. For simplicity, we absorb λ_{grnd} into $\mathcal{L}_{\mathrm{grnd}}$ and define the full objective

$$F(\Theta) = \mathcal{L}_{\mathrm{bag}}(\Theta) + \mathcal{L}_{\mathrm{grnd}}(\Theta).$$

The analysis in this subsection is carried out in the deterministic full-batch setting.

Standing assumptions.

Assumption 8 (Compact-neighborhood smoothness)**.** *There exists a compact convex trajectory neighborhood $\mathcal{K}_{\mathrm{traj}}$ containing the iterates of interest such that the hypotheses of Theorem 1 hold there. In particular, F is C^1 on $\mathcal{K}_{\mathrm{traj}}$ and has an L-Lipschitz gradient on $\mathcal{K}_{\mathrm{traj}}$.*

Assumption 9 (Objective lower boundedness)**.** *There exists $F_\star > -\infty$ such that*

$$F(\Theta) \geq F_\star \qquad \forall\, \Theta \in \mathcal{K}_{\mathrm{traj}}.$$

Assumption 10 (Lipschitz continuity of the grounding loss in the proxy pool)**.** *On $\mathcal{K}_{\mathrm{traj}}$, the grounding loss $\mathcal{L}_{\mathrm{grnd}}$, viewed as a function of the normalized proxy pool with the instance embeddings held fixed, admits an L_p-Lipschitz extension to proxy collections whose entries lie in the closed unit ball. In particular, for any two normalized proxy collections $\{\bar{p}_b\}_{b=1}^M$ and $\{\bar{p}'_b\}_{b=1}^M$,*

$$|\mathcal{L}_{\mathrm{grnd}}(\{\bar{p}_b\}) - \mathcal{L}_{\mathrm{grnd}}(\{\bar{p}'_b\})| \leq L_p \sum_{b=1}^{M} \|\bar{p}_b - \bar{p}'_b\|_2 .$$

Theorem 7 (Full-batch descent on compact trajectory neighborhoods)**.** *Under Assumptions 8 and 9, consider full-batch gradient descent*

$$\Theta_{t+1} = \Theta_t - \eta \nabla F(\Theta_t)$$

with the entire trajectory contained in $\mathcal{K}_{\mathrm{traj}}$. If $0 < \eta \leq 1/L$, then for every $T \geq 1$,

$$\min_{0 \leq t < T} \|\nabla F(\Theta_t)\|_2^2 \leq \frac{2\big(F(\Theta_0) - F_\star\big)}{\eta T}.$$

Proof. Since F has L-Lipschitz gradient on $\mathcal{K}_{\mathrm{traj}}$, the standard descent lemma gives

$$F(\Theta_{t+1}) \leq F(\Theta_t) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla F(\Theta_t)\|_2^2.$$

For $0 < \eta \leq 1/L$, this simplifies to

$$F(\Theta_{t+1}) \leq F(\Theta_t) - \frac{\eta}{2} \|\nabla F(\Theta_t)\|_2^2.$$

Summing from $t = 0$ to $T - 1$ yields

$$\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(\Theta_t)\|_2^2 \leq F(\Theta_0) - F(\Theta_T) \leq F(\Theta_0) - F_\star.$$

Dividing by T gives the claimed bound. □

Remark 17 (Scope of the full-batch descent result)**.** *The weighted proxy yields a compact-neighborhood smoothness theory. No fixed-support strata, proxy-jump lemmas, or switch-budget corrections are needed in this formulation.*

E.2 Probabilistic mini-batch anti-collapse

E.2.1 Setup

We analyze the mini-batch setting where, at each training step t, a mini-batch $\mathcal{B}_t$ of B bags is sampled without replacement from the full dataset of N_{total} bags; in this appendix, B denotes the mini-batch size (M in the main text). Let C denote the number of classes and N_c the number of bags of class c, with class frequency $p_c = N_c/N_{\text{total}}$ and $\sum_{c=1}^{C} p_c = 1$. The number of class-c bags included in $\mathcal{B}_t$ is denoted by $K_{c,t}$.

For the coefficient calculations only, assume equal bag sizes $N_b = n$ and common retained mass at the locally symmetric state, so that $\hat{m}_i^* = 1/(Bn)$ and $N_{\mathcal{B}} = Bn$ is deterministic. The mini-batch anti-collapse coefficient induced by the self-anchor gap is

$$\beta_c^{\text{mini}} \;=\; \frac{1-\gamma}{N_{\mathcal{B}}\,\varepsilon\,(1+(K_{c,t}-1)\gamma)} \cdot \mathbf{1}\{K_{c,t} \geq 2\}, \tag{9}$$

The prefactor $N_{\mathcal{B}}^{-1}$ comes from the uniform routing weights under these assumptions. When $K_{c,t} < 2$, the class-relative term vanishes. The sampling probabilities below do not require equal bag sizes or common mass.

E.2.2 Per-step anti-collapse probability

Proposition 8 (Per-step anti-collapse probability)**.** *Under sampling without replacement, $K_{c,t}$ follows a* Hypergeometric$(N_{\text{total}}, N_c, B)$ *distribution. For $N_{\text{total}} \gg B$, it is well approximated by* Binomial(B, p_c)*. More precisely, if $q_c^{\text{hyper}} := P_{\text{hyper}}(K_{c,t} < 2)$ and $q_c^{\text{bin}} := P_{\text{bin}}(K_{c,t} < 2)$, then a standard finite-population approximation gives*

$$\left|q_c^{\text{hyper}} - q_c^{\text{bin}}\right| \leq C_{\text{hg}} \frac{B^2}{N_{\text{total}}}$$

for some absolute constant $C_{\text{hg}} > 0$. Thus the Binomial surrogate is accurate in the regime $N_{\text{total}} \gg B$, giving

$$P(K_{c,t} \geq 2) \;\approx\; 1 - (1-p_c)^{B-1}\big(1+(B-1)p_c\big), \tag{10}$$

and complementary collapse probability

$$q_c \;:=\; P(K_{c,t} < 2) \;\approx\; (1-p_c)^{B-1}\big(1+(B-1)p_c\big). \tag{11}$$

The joint probability that every class represented in training satisfies the anti-collapse condition at a given step obeys

$$1 - \sum_{c=1}^{C} q_c \;\leq\; P(K_{c,t} \geq 2,\ \forall c) \;\leq\; \prod_{c=1}^{C} P(K_{c,t} \geq 2) \;\approx\; \prod_{c=1}^{C} \Big[1 - (1-p_c)^{B-1}\big(1+(B-1)p_c\big)\Big], \tag{12}$$

where the lower bound is a union bound and the upper bound holds since the class counts of a sample drawn without replacement are negatively associated and the events $\{K_{c,t} \geq 2\}$ are increasing in disjoint coordinates [32]. Moreover, $P(K_{c,t} \geq 2)$ is strictly increasing in B and p_c, and the bottleneck class is the rarest one.

Proof. The Binomial approximation gives

$$P(K_{c,t} < 2) = P(K_{c,t} = 0) + P(K_{c,t} = 1) = (1-p_c)^B + Bp_c(1-p_c)^{B-1},$$

which simplifies to (11); subtracting from one gives (10). Monotonicity in B and p_c follows by direct differentiation of the closed form. □

E.2.3 Expected anti-collapse coefficient

Proposition 9 (Expected self-anchor-gap anti-collapse coefficient)**.** *Under the equal-size, common-mass assumptions above and the Binomial approximation,*

$$\mathbb{E}\big[\beta_c^{\text{mini}}\big] \;=\; \frac{1-\gamma}{N_{\mathcal{B}}\,\varepsilon} \sum_{k=2}^{B} \binom{B}{k} p_c^k (1-p_c)^{B-k} \cdot \frac{1}{1+(k-1)\gamma}. \tag{13}$$

In particular, $\mathbb{E}[\beta_c^{\mathrm{mini}}] = 0$ *for* $B = 1$*, and it decreases with* γ*. For* $\gamma = 0$*, the sum in* (13) *equals* $P(K_{c,t} \geq 2)$*, which increases with* B *and* p_c*; the prefactor* $1/N_{\mathcal{B}}$ *(with* $N_{\mathcal{B}} = BN_b$ *under equal bag sizes) decreases in* B*. For* $\gamma > 0$*, it is in general not monotone in* B *or* p_c*, since larger* B *or* p_c *make* $K_{c,t} \geq 2$ *more likely but also enlarge* $K_{c,t}$ *in the denominator of* (9)*.*

Proof. Substitute the Binomial PMF into the definition of β_c^{mini} in (9) and sum over $k \geq 2$. Each summand is decreasing in γ, since $\frac{d}{d\gamma}\frac{1-\gamma}{1+(k-1)\gamma} = \frac{-k}{(1+(k-1)\gamma)^2} < 0$. For $\gamma = 0$, the sum equals $P(K_{c,t} \geq 2)$, which increases with B and p_c by Proposition 8. □

E.2.4 Trajectory-level guarantee

Proposition 10 (Trajectory-level collapse rate bound)**.** *Assume mini-batches are sampled independently across training steps. Define the per-class collapse indicator* $X_{c,t} = \mathbf{1}\{K_{c,t} < 2\}$ *with* $\mathbb{E}[X_{c,t}] = q_c$*, and let* $\bar{Q}_c := T^{-1}\sum_{t=1}^{T} X_{c,t}$ *be the empirical collapse rate over* T *steps. Then Hoeffding's inequality and a union bound give*

$$P\big(\bar{Q}_c \leq q_c + \delta,\ \forall c \in \{1, \ldots, C\}\big)\ \geq\ 1 - C\exp(-2T\delta^2). \tag{14}$$

Proof. For each fixed c, the indicators $X_{c,t} \in [0,1]$ are Bernoulli random variables with mean q_c. The independence of the mini-batch draws across steps implies that $\{X_{c,t}\}_{t=1}^{T}$ are independent across t, so Hoeffding's inequality yields $P(\bar{Q}_c \geq q_c + \delta) \leq e^{-2T\delta^2}$. A union bound over the C classes gives (14). □

E.3 Mini-batch descent on compact trajectory neighborhoods

We extend the compact-neighborhood full-batch descent result to mini-batch updates with explicit bias and variance terms.

E.3.1 Setup and assumptions

At each step t, a mini-batch $\mathcal{B}_t$ of B bags is sampled without replacement from the full dataset of N_{total} bags. Let

$$\hat{g}_t := \hat{\nabla}F(\Theta_t; \mathcal{B}_t) = \hat{\nabla}\mathcal{L}_{\mathrm{bag}}(\Theta_t; \mathcal{B}_t) + \hat{\nabla}\mathcal{L}_{\mathrm{grnd}}(\Theta_t; \mathcal{B}_t)$$

denote the exact mini-batch gradient, including UOT solution derivatives, and consider the update

$$\Theta_{t+1} = \Theta_t - \eta\hat{g}_t.$$

Assumption 11 (MB1 (Controlled bias))**.** *On* $\mathcal{K}_{\mathrm{traj}}$*, the conditional bias of the stochastic gradient is uniformly bounded by*

$$\delta_B := \sup_{\Theta \in \mathcal{K}_{\mathrm{traj}}} \big\|\mathbb{E}_{\mathcal{B}_t}[\hat{g}_t \mid \Theta] - \nabla F(\Theta)\big\|_2 \leq L_p'\left(1 - \frac{B}{N_{\mathrm{total}}}\right)$$

for some finite constant L_p'*.*

Assumption 12 (MB2 (Bounded variance))**.** *On* $\mathcal{K}_{\mathrm{traj}}$*, the conditional variance of the stochastic gradient is uniformly bounded by*

$$\sigma_B^2 := \mathbb{E}_{\mathcal{B}_t}\Big[\|\hat{g}_t - \mathbb{E}[\hat{g}_t \mid \Theta_t]\|_2^2 \mid \Theta_t\Big] \leq \frac{C_\sigma^2}{B}$$

for some finite constant $C_\sigma < \infty$*.*

Remark 18 (Scope of the moment assumptions)**.** *The bias and variance bounds are independent assumptions; loss-value Lipschitz continuity alone does not establish the gradient-bias bound. The stop-gradient implementation would require an additional bias term.*

Theorem 8 (Mini-batch descent on compact trajectory neighborhoods)**.** *Assume Assumptions 8, 9, 11 and 12, and suppose the mini-batch trajectory remains in* $\mathcal{K}_{\mathrm{traj}}$*. If* $0 < \eta \leq 1/(4L)$*, then for every* $T \geq 1$*,*

$$\min_{0 \leq t < T} \mathbb{E}\big[\|\nabla F(\Theta_t)\|_2^2\big] \leq \frac{4\big(F(\Theta_0) - F_\star\big)}{\eta T} + 3\delta_B^2 + 2L\eta\sigma_B^2.$$

Proof. Let $\bar{g}_t := \mathbb{E}[\hat{g}_t \mid \Theta_t]$ and $b_t := \bar{g}_t - \nabla F(\Theta_t)$, so $\|b_t\|_2 \le \delta_B$. By L-smoothness,

$$F(\Theta_{t+1}) \le F(\Theta_t) - \eta\langle \nabla F(\Theta_t), \hat{g}_t\rangle + \frac{L\eta^2}{2}\|\hat{g}_t\|_2^2.$$

Taking conditional expectation given Θ_t yields

$$\mathbb{E}[F(\Theta_{t+1}) \mid \Theta_t] \le F(\Theta_t) - \eta\langle \nabla F(\Theta_t), \bar{g}_t\rangle + \frac{L\eta^2}{2}\mathbb{E}[\|\hat{g}_t\|_2^2 \mid \Theta_t].$$

For the inner-product term,

$$-\langle \nabla F, \bar{g}_t\rangle = -\|\nabla F\|_2^2 - \langle \nabla F, b_t\rangle \le -\frac{1}{2}\|\nabla F\|_2^2 + \frac{1}{2}\delta_B^2.$$

For the second moment,

$$\mathbb{E}[\|\hat{g}_t\|_2^2 \mid \Theta_t] = \|\bar{g}_t\|_2^2 + \mathbb{E}[\|\hat{g}_t - \bar{g}_t\|_2^2 \mid \Theta_t] \le 2\|\nabla F(\Theta_t)\|_2^2 + 2\delta_B^2 + \sigma_B^2.$$

Combining the two bounds gives

$$\mathbb{E}[F(\Theta_{t+1}) \mid \Theta_t] \le F(\Theta_t) - \left(\frac{\eta}{2} - L\eta^2\right)\|\nabla F(\Theta_t)\|_2^2 + \left(\frac{\eta}{2} + L\eta^2\right)\delta_B^2 + \frac{L\eta^2}{2}\sigma_B^2.$$

If $0 < \eta \le 1/(4L)$, then $\frac{\eta}{2} - L\eta^2 \ge \frac{\eta}{4}$ and $\frac{\eta}{2} + L\eta^2 \le \frac{3\eta}{4}$, so

$$\mathbb{E}[F(\Theta_{t+1}) \mid \Theta_t] \le F(\Theta_t) - \frac{\eta}{4}\|\nabla F(\Theta_t)\|_2^2 + \frac{3\eta}{4}\delta_B^2 + \frac{L\eta^2}{2}\sigma_B^2.$$

Taking total expectations, summing over $t = 0, \dots, T-1$, and using $F(\Theta_T) \ge F_\star$ gives

$$\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F(\Theta_t)\|_2^2\big] \le F(\Theta_0) - F_\star + \frac{3\eta T}{4}\delta_B^2 + \frac{L\eta^2 T}{2}\sigma_B^2.$$

Dividing by $\eta T/4$ yields the stated bound. □

Corollary 5 (Joint descent and anti-collapse controls under mini-batch training)**.** *Suppose the hypotheses of Theorem 8 hold. Then:*

(1) ***(Descent)*** *The bound of Theorem 8 holds in expectation over a T-step trajectory.*

(2) ***(Anti-collapse control)*** *Assume additionally that mini-batches are sampled independently across training steps. Then, with probability at least $1 - C\exp(-2T\delta^2)$ over the draw of mini-batches $\mathcal{B}_1, \dots, \mathcal{B}_T$* (Proposition 10)*, for every class c, the empirical collapse rate satisfies*

$$\bar{Q}_c \le q_c + \delta.$$

Under the equal-size, common-mass coefficient assumptions and $0 \le \gamma < 1$, on every step with $K_{c,t} \ge 2$, the mini-batch anti-collapse coefficient β_c^{mini} in (9) *is strictly positive, and its expectation is lower-controlled by Proposition 9. A nonzero separating field additionally requires the non-degeneracy condition of Appendix D.3.*

Proof. Part (1) is exactly Theorem 8, so it is an expectation statement. Part (2) follows from Proposition 10, together with the positivity of β_c^{mini} whenever $K_{c,t} \ge 2$ by (9), and the expectation lower control in Proposition 9. □

Remark 19 (Comparison with full-batch descent)**.** *As $B \to N_{\mathrm{total}}$, the bias term δ_B vanishes for the exact-gradient estimator (the stop-gradient field of the implementation retains its own bias) and the variance term shrinks, so the mini-batch bound approaches the deterministic full-batch descent result. For fixed B, the expected gradient norm obeys the standard stochastic-descent tradeoff between optimization error, bias floor, and variance.*

Remark 20 (Scope of the mini-batch result)**.** *Theorem 8 shares the same scope as Appendix E.1: it is a compact-neighborhood result and is not intended as a global stochastic convergence theorem outside the region where the stated regularity and moment assumptions hold.*

F Experiments and Optional Components

This appendix collects implementation details, secondary training regimes, and extended ablation studies that complement the main-text results. It records alternative settings and broader parameter sweeps omitted from the main comparison.

F.1 Detailed experimental settings

Table 5 summarizes the detailed dataset configurations, including the numbers of training and held-out bags, bag length, signal ratio, backbone, and batch size.

All baselines are evaluated under the same multiclass bag-classification protocol. The final prediction layer outputs C class logits and is trained with bag-level cross-entropy. For baselines whose original formulation targets binary classification, ranking, or survival prediction, we modify only the task-specific head and objective, leaving the MIL aggregation or transport module unchanged.

For the WSI comparison, TransMIL [7] and DTFD-MIL [8] use the same input features, data splits, training budget, checkpoint-selection criterion, and evaluation metrics as the other baselines. We preserve TransMIL's correlated transformer-based MIL aggregation and DTFD-MIL's two-tier feature-distillation procedure, modifying only the final prediction head and objective for the common multiclass setting.

We use $\rho_e = 1.5$ for the synthetic benchmarks and CLE, and $\rho_e = 1.1$ for the WSI experiments. Except for the dedicated sensitivity sweep in Appendix F.3, the relaxed-target parameter is fixed to $\gamma = 0.7$ in all experiments.

We used a small pilot validation sweep only to identify a stable operating region rather than to maximize validation or test performance. Unless otherwise stated, the resulting hyperparameter settings were fixed within each dataset family and reused across folds and random seeds.

For the synthetic benchmarks, the signal/noise split is class-balanced: class indices 0–4 are treated as signal classes and 5–9 as noise classes. Each bag has length 100 and signal ratio 0.05. MNIST uses 2,000 training bags and 900 held-out bags, while Fashion-MNIST and CIFAR10 use 2,000 training bags and 1,000 held-out bags. Across these synthetic settings, we use $\lambda_{\mathrm{grnd}} = 1$ and set $\rho_{\mathrm{src}} = \rho_{\mathrm{tgt}} = \varepsilon = 0.1$. All models are optimized with AdamW (learning rate 10^{-4}, weight decay 0.01) for 30 epochs under a cosine learning-rate schedule, and checkpoint selection is based on held-out bag accuracy. Signal instances are not reused across bags, whereas nuisance/noise negatives may repeat across different bags. The known held-out instance labels are used only for evaluation and never for training or model selection. All synthetic results are reported as mean $\pm$ std over 5 random seeds.

WSI experimental settings. For the whole-slide image benchmarks (TCGA Lung and BRACS), we follow the standard two-stage MIL pipeline: patch extraction, feature embedding, and bag-level aggregation.

Patch extraction. Non-overlapping 256×256 patches are extracted at $20\times$ magnification from the segmented tissue area of each slide. Tissue regions are identified via Otsu thresholding on a downsampled thumbnail, and patches whose tissue content falls below a minimum threshold are discarded.

Feature extraction. Each patch is encoded by a ResNet-50 pretrained on ImageNet [33, 34]. The encoder is kept frozen throughout all experiments; no slide-level fine-tuning of the backbone is performed.

Feature projection. A 6-layer MLP (input dimension $1{,}024$, output dimension 512, with ReLU activations between layers) projects the patch features into a lower-dimensional embedding space. This projection head is trained jointly with the MIL head under bag-level cross-entropy supervision, while the ResNet-50 backbone remains frozen.

Dataset splits and evaluation. For TCGA Lung, slides are partitioned following a fixed $65:10:25$ train : validation : test ratio and evaluated under 4-fold cross-validation. For BRACS, we follow the official dataset split with 5 random seeds. All WSI results are reported as mean $\pm$ std over the respective folds or seeds.

Optimization. All WSI models are trained with AdamW ($\mathrm{lr} = 10^{-4}$, weight decay $= 0.01$) for 100 epochs under a cosine learning-rate schedule. Model selection is based on validation AUC. For the WSI experiments, we use $\lambda_{\mathrm{grnd}} = 0.1$ and set $\rho_{\mathrm{src}} = \rho_{\mathrm{tgt}} = \varepsilon = 0.5$. Batch sizes are set to 4 for TCGA Lung and 20 for BRACS.

CLE experimental settings. For CLE, we use the data collection employed in Byun et al. [23] for model training and evaluation. The main-text bag metric is computed under 5-fold bag-level cross-validation. The frame metric is evaluated on a held-out quality-annotated test set disjoint from the training folds, using the encoder trained only from bag supervision and frozen at evaluation time without frame-level fine-tuning. Because the CLE backbone is an ImageNet-pretrained ViT-Base [34, 35], we run these experiments with automatic mixed precision (AMP) via `autocast` for both the forward and backward passes. All CLE models are optimized with AdamW (learning rate 3×10^{-5}, weight decay 0.01) for 50 epochs under a cosine learning-rate schedule, and checkpoint selection is based on bag accuracy. As in the synthetic benchmarks, we use $\lambda_{\mathrm{grnd}} = 1$ and set $\rho_{\mathrm{src}} = \rho_{\mathrm{tgt}} = \varepsilon = 0.1$.

All experiments were conducted on NVIDIA RTX 3090 and RTX A5000 GPUs, each with 24 GB of memory. Runtime varies with the dataset, bag size, and model configuration. The complete experimental suite, including the hyperparameter sweeps reported in this appendix, required approximately 300 GPU-hours in total.

Computational cost of UOT grounding. We measure the added cost against the same attention-pooling backbone with only the UOT grounding branch disabled. On CIFAR10 at signal ratio 0.05 and batch size 20, one epoch takes 455.50 seconds with UOT and 333.12 seconds without UOT, corresponding to a 36.7% increase (approximately 37%). On TCGA Lung at batch size 4, where a slide contains 8,545 patches on average, one epoch over 172 mini-batches takes approximately 32 seconds with UOT and 22 seconds without UOT, corresponding to a 45.5% increase (approximately 45%). Peak GPU memory increases by only approximately 1–2% in both settings. Both timing measurements use FP32; AMP is disabled for the CIFAR10 and TCGA Lung runs, whereas the CLE experiments use AMP as described above. The UOT branch is used only during training; standard full-bag attention is used at inference, so there is no additional Sinkhorn cost at inference time.

For a mini-batch of M bags containing N total instances with embedding dimension d, constructing the instance–proxy cost matrix requires $O(NMd)$ computation. With L Sinkhorn iterations, transport scaling requires $O(LNM)$ computation, and the additional cost and transport matrices require $O(NM)$ memory. Gradients are stopped through the iterative Sinkhorn scaling updates, while gradients to the encoder and scorer are retained through the final cost-dependent transport expression. Thus, the iterative solver graph is not stored, and for fixed d, M, and L, the additional computation scales linearly with the total number of instances. We use fixed iteration budgets of $L = 200$ for CIFAR10 and $L = 400$ for TCGA Lung, with a convergence tolerance of 10^{-10}.

Table 5: Summary of the experimental settings. For CLE, representative per-fold statistics under 5-fold cross-validation are reported, and the signal ratio is not applicable (–).

Dataset	Train Bags	Held-out Bags	Bag Length	Signal Ratio	Backbone	Batch Size
MNIST	2,000	900	100	0.05	3layer MLP	20
Fashion-MNIST	2,000	1,000	100	0.05	3layer MLP	20
CIFAR-10	2,000	1,000	100	0.05	ImageNet Pretrained ResNet-18 [33, 34]	20
CLE	362	91	30–31	–	ImageNet Pretrained ViT-Base [35, 34]	10

F.2 Bag-level heterogeneity score

The main text uses a brief bag-level heterogeneity diagnostic to interpret the mechanism-isolation ablation across the synthetic benchmarks. Because all synthetic datasets fix the bag length at 100 and the signal ratio at 5%, the performance differences in Table 2 cannot be explained by signal prevalence alone. We therefore quantify domain-level bag heterogeneity using a simple variance-equivalent within-bag feature spread score in a frozen pretrained feature space. Let $h_{b,i}$ denote the frozen

encoder feature of instance i in bag b, and define the within-bag feature centroid

$$c_b^{\text{feat}} = \frac{1}{N_b} \sum_{i=1}^{N_b} h_{b,i}.$$

We define

$$D_b = \frac{1}{N_b - 1} \sum_{i=1}^{N_b} \|h_{b,i} - c_b^{\text{feat}}\|_2^2,$$

so larger D_b indicates broader feature spread within a bag, regardless of whether that spread arises from signal variation, weakly relevant instances, background variation, or signal–noise separation.

The rightmost panel of Figure 2 reports the empirical distribution of D_b across synthetic bags. MNIST exhibits lower and narrower heterogeneity values, FMNIST is intermediate, and CIFAR10 occupies a markedly higher and broader range. Together with Table 2, where removing UOT-based contribution control hurts instance accuracy most on CIFAR10, this suggests that the grounding mechanism matters most on the most heterogeneous benchmark. Since bag length and signal ratio are fixed across datasets, this difference is more plausibly associated with domain-level feature heterogeneity than with differences in signal prevalence.

F.3 Full sensitivity to the relaxed-target parameter γ

Table 6 provides the full sweep. Bag accuracy varies little with γ on all three datasets. For instance accuracy, every nonzero γ improves over the self-only target ($\gamma = 0$) on MNIST and CIFAR10, with CIFAR10 peaking at intermediate values ($\gamma = 0.6$–0.7) rather than at the fully uniform target, whereas FMNIST is largely flat. The response is not monotone in γ.

Table 6: Bag and instance accuracy across values of γ (mean over 5 seeds).

Dataset (ratio)	Metric	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR10 (r=0.05)	**Bag Acc**	0.8712	0.8694	0.8654	0.8712	0.8782	0.8746	0.8700	0.8844	0.8752	0.8756	0.8772
	Inst. Acc	0.7358	0.7581	0.7512	0.7510	0.7520	0.7532	0.7600	0.7596	0.7532	0.7544	0.7514
FMNIST (r=0.05)	**Bag Acc**	0.9404	0.9310	0.9368	0.9322	0.9390	0.9352	0.9334	0.9388	0.9332	0.9378	0.9340
	Inst. Acc	0.8317	0.8353	0.8374	0.8311	0.8268	0.8300	0.8293	0.8284	0.8240	0.8390	0.8261
MNIST (r=0.05)	**Bag Acc**	0.9873	0.9942	0.9931	0.9942	0.9933	0.9933	0.9933	0.9896	0.9944	0.9978	0.9944
	Inst. Acc	0.8963	0.9183	0.9190	0.9210	0.9205	0.9195	0.9190	0.9052	0.9271	0.9261	0.9210

F.4 Ablation on transport hyperparameters

To better understand the role of source-side and target-side marginal relaxation, we vary ρ_{src} and ρ_{tgt} separately. Table 7 reports bag and instance accuracy across varying ε, $\rho_{\text{src}}/\varepsilon$, and $\rho_{\text{tgt}}/\varepsilon$ for MNIST and CIFAR10, using the W&B sweep outputs. The results clarify that performance depends on the balance between source-side mass flexibility and target-side matching regularity.

Table 7: Bag and instance accuracy from a single seed across varying ε, $\rho_{\mathrm{src}}/\varepsilon$, and $\rho_{\mathrm{tgt}}/\varepsilon$ for MNIST ($r = 0.05$, left) and CIFAR10 ($r = 0.05$, right). Sweep runs use a different batch size from the main experiments.

			MNIST				CIFAR10			
Metric	ε	$\rho_{\mathrm{src}}/\varepsilon \backslash \rho_{\mathrm{tgt}}/\varepsilon$	**x0.5**	**x1**	**x2**	**x5**	**x0.5**	**x1**	**x2**	**x5**
Bag	**0.01**	**x0.5**	1.0000	1.0000	0.9989	0.9989	0.8890	0.8870	0.8910	0.8690
		x1	1.0000	1.0000	1.0000	0.9989	0.8680	0.9050	0.8950	0.8740
		x2	1.0000	1.0000	1.0000	0.9989	0.8910	0.8970	0.8790	0.8810
		x5	0.9200	0.4033	0.4467	0.8256	0.3060	0.3670	0.7620	0.8290
	0.03	**x0.5**	1.0000	1.0000	1.0000	0.9956	0.8940	0.8920	0.8730	0.8340
		x1	1.0000	1.0000	1.0000	0.9989	0.8640	0.8830	0.8710	0.8060
		x2	1.0000	1.0000	1.0000	1.0000	0.8830	0.8910	0.8980	0.8400
		x5	0.9978	0.9978	0.9978	0.9978	0.8320	0.8570	0.8420	0.8290
	0.05	**x0.5**	1.0000	1.0000	1.0000	0.9978	0.8800	0.8830	0.8470	0.7780
		x1	1.0000	1.0000	1.0000	1.0000	0.8890	0.8940	0.8530	0.8360
		x2	1.0000	1.0000	1.0000	1.0000	0.8680	0.8590	0.8840	0.8360
		x5	0.9956	0.9956	0.9956	0.9922	0.8480	0.8400	0.8550	0.6710
	0.1	**x0.5**	1.0000	1.0000	1.0000	1.0000	0.8620	0.8700	0.8450	0.8200
		x1	0.9956	0.9956	0.9956	0.9956	0.8810	0.8870	0.8860	0.8640
		x2	0.9944	0.9933	0.9933	0.9956	0.8520	0.7850	0.7360	0.7840
		x5	0.9922	0.9922	0.9856	0.9867	0.5980	0.5750	0.2510	0.5000
Instance	**0.01**	**x0.5**	0.9624	0.9380	0.9547	0.9542	0.7866	0.7830	0.7904	0.7542
		x1	0.9622	0.9598	0.9562	0.9524	0.7852	0.7888	0.7890	0.7778
		x2	0.9600	0.9493	0.9549	0.9516	0.7646	0.7800	0.7948	0.7698
		x5	0.9142	0.8269	0.8309	0.8440	0.4366	0.4172	0.6972	0.7162
	0.03	**x0.5**	0.9711	0.9682	0.9564	0.9669	0.7680	0.7876	0.7716	0.7142
		x1	0.9691	0.9678	0.9638	0.9644	0.7480	0.7720	0.7648	0.7234
		x2	0.9640	0.9636	0.9600	0.9593	0.7570	0.7576	0.7538	0.7258
		x5	0.9573	0.9573	0.9556	0.9549	0.7202	0.7096	0.7302	0.7094
	0.05	**x0.5**	0.9729	0.9684	0.9669	0.9667	0.7664	0.7470	0.7424	0.6824
		x1	0.9698	0.9698	0.9656	0.9680	0.7386	0.7338	0.7414	0.7166
		x2	0.9500	0.9422	0.8316	0.9653	0.7422	0.7046	0.7352	0.6974
		x5	0.9453	0.9404	0.9304	0.9204	0.7242	0.7210	0.6838	0.5338
	0.1	**x0.5**	0.9704	0.9682	0.9716	0.9720	0.7400	0.7292	0.7236	0.6840
		x1	0.9533	0.9453	0.9493	0.9387	0.7202	0.7384	0.7116	0.6962
		x2	0.9296	0.9116	0.9173	0.9167	0.7072	0.6300	0.5226	0.6352
		x5	0.9053	0.9033	0.9289	0.8973	0.3588	0.3432	0.2138	0.2766

F.5 Sensitivity to mini-batch size

Because the proxy pool is built from the bags in the current mini-batch, we evaluate the sensitivity to the mini-batch size M on CIFAR10 ($r = 0.05$) under the same fixed 30-epoch budget as Table 1. With 2,000 training bags, a larger M also reduces the number of optimizer updates and enlarges the transport problem per step, so this experiment measures practical sensitivity rather than the effect of proxy-pool size alone. As shown in Table 8, performance stays in a similar range for $M = 5$–20: the default $M = 20$ gives the highest mean bag accuracy and $M = 10$ the highest mean instance accuracy. The drop at $M = 40$ and 80 cannot be attributed to proxy-pool size alone, since the number of optimizer updates decreases at the same time.

Table 8: Sensitivity to the mini-batch size M on CIFAR10 ($r = 0.05$) under a fixed 30-epoch budget (mean $\pm$ std over 5 seeds).

M	Optimizer updates	Bag Acc	Inst Acc
5	12,000	88.20 ± 1.03	73.70 ± 2.09
10	6,000	87.16 ± 0.50	$\mathbf{76.46 \pm 0.64}$
20 (default)	3,000	$\mathbf{88.44 \pm 0.96}$	75.96 ± 0.66
40	1,500	78.34 ± 2.31	66.22 ± 2.68
80	750	72.30 ± 3.16	59.86 ± 1.98

F.6 Qualitative examples on BRACS

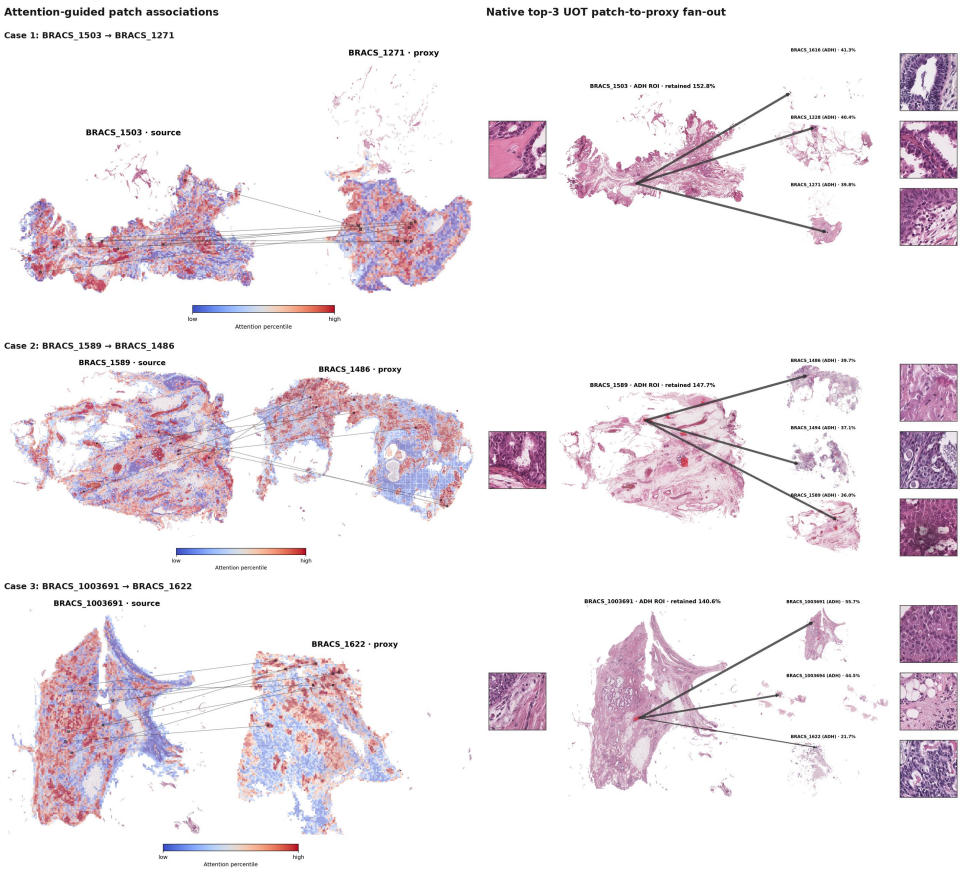


Figure 3: Qualitative examples on BRACS [31]. Left: attention heatmaps with post-hoc patch associations to a proxy slide. Right: UOT mass from a patch inside an annotated ADH ROI to its top-3 proxies, as a percentage of the patch's unit prior mass. Since BRACS provides only ROI-level annotations, these examples are illustrative: such patches retain more than their prior mass and route most of it to same-class proxies.

G Broader Impacts

Sparse Support Grounding aims to make weakly supervised analysis of medical images more reliable by emphasizing informative instances, which may assist pathology and intraoperative diagnosis workflows. Potential negative impacts include over-reliance on unvalidated predictions, incorrect instance grounding being mistaken for clinical evidence, performance degradation under cohort or site shift, and privacy risks associated with clinical data. The retained UOT mass is a model-dependent relative influence score, not a calibrated instance label, and the method should not be used for autonomous clinical diagnosis without prospective, multi-site validation and appropriate regulatory review.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

Justification: The abstract and introduction state the proposed Sparse Support Grounding framework, its support–routing–mass mechanism, and the evaluated settings. The claims are limited to heterogeneous MIL and are supported by experiments on synthetic MIL, WSI benchmarks, and CLE.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 discusses the identifiability limits of bag-level supervision, the approximate nature of the bag-specific proxy, the locality of mini-batch supports, and the need for broader multi-site validation. Appendix A.1 states the local, proxy-level scope of the theory, and Appendix F reports the additional computation of the UOT branch and the sensitivity to transport hyperparameters.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper states the assumptions and provides derivations and formal analyses in the appendices, including the UOT retained-mass derivation, balanced-OT comparison, anti-collapse analysis, and compact-neighborhood convergence result.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper describes the model components, training objective, datasets, baseline adaptations, splits, folds/seeds, evaluation metrics, and implementation details needed to reproduce the main results. The supplemental zip additionally provides code, configuration files, and running instructions.

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

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

   Answer: [Yes]

   Justification: We provide an anonymized supplemental code zip with implementation, configuration files, environment instructions, and scripts for reproducing the public-data experiments. Public datasets are accessed through their original sources, while the clinical CLE cohort is not redistributed due to privacy and IRB restrictions; the paper provides the protocol needed to verify the evaluation under approved access.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper and appendix specify dataset construction, train/validation/test or cross-validation splits, bag sizes, signal ratios, backbones, batch sizes, optimizer, learning rate, weight decay, scheduler, model-selection criterion, training epochs, and evaluation protocol. They also state that a small pilot validation sweep was used only to identify a stable operating region, after which hyperparameters were fixed within each dataset family and reused across folds and random seeds.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The main experimental tables report mean and standard deviation over multiple seeds or folds. Synthetic experiments are reported over 5 random seeds, WSI results over the corresponding folds or random seeds, and CLE results over 5-fold cross-validation.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The appendix reports the GPU types and memory used, per-epoch runtime with and without the UOT branch for representative settings, and the total compute required for the full experimental suite including hyperparameter sweeps.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The study uses public benchmarks and an IRB-approved, de-identified clinical CLE cohort, does not release identifiable clinical data, and presents the method as a research benchmark rather than a deployed clinical diagnostic system.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts for weakly supervised medical-imaging analysis and potential negative impacts including over-reliance on unvalidated clinical predictions, incorrect instance grounding, cohort or site bias, and privacy concerns. It also states that the method should not be used for autonomous clinical diagnosis without prospective validation. (Appendix G)

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models, scraped datasets, or dual-use pretrained models. The released zip contains research code only, and clinical data are de-identified and not redistributed.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the original sources for the datasets, baselines, and pretrained backbones used in the experiments. The supplemental code zip includes a README documenting the licenses or access terms of the public assets where available, and states that restricted clinical data are not redistributed.

13. **New assets**

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

    Answer: [Yes]

Justification: The paper releases a new anonymized code asset as a supplemental zip. The zip includes documentation, environment setup, configuration files, run scripts, and notes on reproducibility, limitations, and asset usage; no new dataset is released.

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

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

    Answer: [N/A]

    Justification: The study does not involve crowdsourcing, participant-facing tasks, new subject recruitment, or compensation. The clinical component uses an existing de-identified CLE cohort under an approved protocol.

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

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

    Answer: [Yes]

    Justification: The CLE evaluation follows an IRB-approved, de-identified clinical cohort and protocol. No identifiable patient data are released.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core method, experiments, evaluation, or scientific claims.