ANCHOR DIVERGENCE FOR SEMANTIC GEOMETRY IN CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper concerns how semantic context determines geometry in learned vector representations. Similarity is typically measured using cosine similarity, which provides a single fixed geometry. Semantic similarity, however, is inherently context dependent: two images may be similar because they depict the same object, share a visual style, or are relevant to the same clinical finding. We show that contrastive representations naturally encompass a family of geometries that can be specialized to particular semantic structure. The key idea is to use an interplay between contrastive learning, exponential families, and information geometry to establish a correspondence between probability distributions over "anchors" and Bregman geometries on the representation space. We use this correspondence to define *Anchor Divergences*, a method for specifying context-specific semantic geometries on fixed representations. Under this correspondence, modeling the anchor distribution models the geometry itself. Experiments on retrieval show that anchor divergences provide an effective and efficient way to specify context-specific semantic similarity.

1 INTRODUCTION

Modern AI systems encode semantic structure in dense vector representations. This paper asks how similarity between such representations should be measured. We focus on contrastive models such as CLIP and its successors (Radford et al., 2021), which are trained to score matched query–key pairs above mismatched ones using cosine-based softmax objectives. It is therefore natural to use cosine similarity as the geometry of the learned space. But this creates a basic mismatch: cosine provides a single fixed notion of similarity, while semantic similarity is inherently context dependent.

For concreteness, we focus on image similarity in vision–language models, and study how semantic similarity should be defined between pairs of images. Figure 1 illustrates the fundamental limitation of cosine similarity in this setting. A query image may share different semantic attributes with different database images, while cosine implicitly fixes how these sources of variation contribute to similarity. The relevant notion, however, depends on context: we may care about the same dog breed, background, or outfit. Understanding how to specify such context-specific similarity is important both foundationally—the question of how dense representations encode semantics is the heart of (mechanistic) interpretability—and practically. For example, when retrieving chest X-rays similar to a query image, we may want similarity to reflect specific pathologies rather than unrelated patient characteristics or acquisition artifacts; see Figure 2. A fixed cosine geometry has no reason to privilege the semantic aspect relevant to the task.

The main contribution of this paper is to show that the contrastive structure naturally defines a *family* of semantic geometries on the image representation space, each induced by a probability distribution over semantically relevant "anchors." This correspondence leads to *anchor divergences*, a method for specifying context-specific semantic geometries on fixed representations; see Figure 1.

Informally, the starting idea is to specify the semantics of interest using a set of text anchors. For example, in the chest X-ray setting, the anchors might consist of clinical descriptions of pathologies—the variation in the text includes only the aspects we care about. Relative to any such anchor set, the contrastive model uses the same softmax structure that appears during training to map each image to a distribution over the anchors, based on their image–text similarities. We can then compare

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Anchor divergences induce context-specific neighborhoods for the same query.** Cosine similarity gives a single fixed ranking. In contrast, changing the semantic anchors changes which aspect of the query defines its neighborhood. For example, breed anchors retrieve other Huskies while allowing background and clothing to vary; background and clothing anchors behave analogously. The underlying image representations remain fixed throughout.

two images through the distributions they induce over the anchors. In this way, the chosen anchors determine which semantic differences are emphasized by the comparison.

In principle, we could define context-specific similarity directly using a finite anchor set. However, this ties the geometry to a particular collection of anchors: it may reflect idiosyncrasies of that finite set, and the cost of representing the geometry grows with the number of anchors. The main technical insight is to move from identifying the geometry with a finite set of text anchors to identifying it with a probability distribution from which the anchors are drawn i.i.d. Using ideas from information geometry, we show that each such distribution corresponds to a Bregman divergence defined on the original image embeddings. This shifts the object specifying semantic geometry from a finite anchor set to an underlying semantic distribution, which can in turn be modeled efficiently.

The development of the paper is:

1. First, we show that the KL divergence between softmax distributions over a finite set of semantic anchors has a natural limit as a Bregman divergence on image representations. This reveals a direct role of information geometry in organizing semantic similarity in contrastive representations, and leads to *anchor divergences*—a principled method for specifying context-specific semantic geometries.

2. A key consequence of the Anchor Divergence construction is that the anchor distribution determines the induced Bregman geometry. Drawing on classical connections between exponential families and Bregman geometry (Efron, 1978; Amari, 1982; Wainwright & Jordan, 2008; Amari & Cichocki, 2010), we use the observed anchors to fit a parametric model of the underlying population anchor distribution, whose fitted distribution induces an explicit Bregman divergence. This has two practical benefits: once the model is fit, the cost of representing and evaluating the geometry need not scale with the number of observed anchors, and the parametric structure can provide useful approximations when only a limited number of anchors are available.

3. As a further consequence of modeling the anchor distribution, we show how broad semantic structure can be adapted to a specific target using an empirical-Bayes-type construction. The idea is to first learn a broad anchor distribution—and hence a broad semantic geometry—from a large collection of anchors that are loosely related to the target, e.g., all text about diseases, and then refine it using a small set of targeted anchors. Empirically, we find that this can reduce task-specific anchor requirements by a full order of magnitude.

4. Finally, experiments test the consequences of the framework. We show that the anchor distribution controls which semantic factors govern similarity; parametric models are useful when anchors are limited, while the non-parametric softmax improves as anchor coverage grows; and broad semantic structure can be adapted using few target anchors. Across these settings, the resulting geometries improve over task-agnostic cosine similarity and task-specific heuristic approaches.

2 ANCHOR DIVERGENCE

We begin with a finite set of text anchors $\Gamma = \{\gamma_1, \ldots, \gamma_N\}$ chosen to describe the semantics of interest. For example, in a pathology-specific geometry, the anchors may be clinical descriptions of different findings. Given an image embedding λ, the contrastive softmax induces a distribution over these anchors, $\mathbb{P}_\Gamma(\gamma_k \mid \lambda) \propto \exp(\lambda^\top \gamma_k/\tau)$, where $\tau > 0$ is the temperature. A natural way to compare two image embeddings is then by the KL divergence between the distributions they induce:

$$d_{\mathrm{SM},\Gamma}(\lambda_1 \| \lambda_2) := \mathrm{KL}(\mathbb{P}_\Gamma(\cdot \mid \lambda_1) \,\|\, \mathbb{P}_\Gamma(\cdot \mid \lambda_2)) \,.$$

We will refer to this as the *softmax anchor divergence*.

Specifying a divergence via a finite set of anchors is appealingly simple. However, we may worry that the procedure may be overly sensitive to the specific choice of anchors. For example, if swapping out a single clinical note substantively altered the geometry of the pathology-specific anchor divergence, it would suggest that we have not captured a meaningful notion of pathology-specific geometry. This motivates viewing the observed anchors as samples from an underlying distribution over semantically relevant anchors. In this view, the distribution, rather than any particular finite realization, specifies the semantic context. We now see that such a distribution induces a well-defined population geometry on the image embedding space.

To state the result, we recall two standard definitions.

Definition 1. *For differentiable, strictly convex f, the* Bregman divergence *is*

$$B_f(\eta_1 \| \eta_2) := f(\eta_1) - f(\eta_2) - \nabla f(\eta_2)^\top (\eta_1 - \eta_2).$$

Definition 2. *For a distribution q on $\mathbb{R}^d$, its* cumulant generating function *(CGF) is*

$$\kappa_q(\eta) := \log \int \exp(\eta^\top \gamma)\, q(d\gamma).$$

The next result identifies the corresponding population geometry.

Theorem 1 (Softmax anchor divergence converges to a population Bregman divergence)**.** *Let $\Gamma_N = \{\gamma_1, \ldots, \gamma_N\}$ be sampled i.i.d. from an anchor distribution q on $\mathbb{S}^{d-1}$. Then, as $N \to \infty$,*

$$\sup_{\lambda_1, \lambda_2 \in \mathbb{S}^{d-1}} \left| d_{\mathrm{SM},\Gamma_N}(\lambda_1 \| \lambda_2) - B_{\kappa_q}\left(\frac{\lambda_2}{\tau} \middle\| \frac{\lambda_1}{\tau}\right) \right| \xrightarrow{\text{a.s.}} 0. \tag{1}$$

The proof is deferred to Appendix A. The key observation is that the contrastive softmax is an exponential tilt of the empirical measure over anchors. The resulting natural exponential family links KL divergence to the Bregman geometry generated by κ_q.

Thus, the simple idea of comparing images through their distributions over semantic anchors yields a population Bregman geometry determined by the anchor distribution. This motivates our central definition.

Definition 3. *For a distribution q on $\mathbb{R}^d$, the* anchor divergence *is*

$$d_q(\lambda_1 \| \lambda_2) := B_{\kappa_q}\left(\frac{\lambda_2}{\tau} \middle\| \frac{\lambda_1}{\tau}\right). \tag{2}$$

Under this definition, the anchor distribution q specifies the semantic context and thereby determines the geometry on the image embedding space. The softmax anchor divergence is the special case induced by the empirical distribution. This distributional view will allow us to model and adapt the semantic geometry in the following sections.

Table 1: **Modeling the anchor distribution induces different semantic geometries.** Each column specifies an anchor-distribution model and its CGF. The empirical model places uniform mass on the observed anchors; for the Gaussian model, $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance; for the Gaussian mixture, $\{(\hat{\pi}_m, \hat{\mu}_m, \hat{\Sigma}_m)\}_{m=1}^{M}$ are fit by expectation–maximization.

Anchor model	Empirical	Gaussian	Gaussian Mixture
CGF $\kappa(\eta)$	$\log \frac{1}{\lvert\Gamma\rvert} \sum_{\gamma \in \Gamma} \exp(\eta^\top \gamma)$	$\hat{\mu}^\top \eta + \frac{1}{2} \eta^\top \hat{\Sigma} \eta$	$\log \sum_{m=1}^{M} \hat{\pi}_m \exp\left(\hat{\mu}_m^\top \eta + \frac{1}{2} \eta^\top \hat{\Sigma}_m \eta\right)$

3 Modeling Structured Semantic Geometries

We have arrived at anchor divergences as Bregman divergences induced by an anchor *distribution* q that represents the semantics of interest. In practice, however, we observe only a finite set of anchors. We therefore turn to modeling the underlying distribution from these observations, and hence the induced semantic geometry.

The simplest approach is to use the empirical measure on the anchor set. It is straightforward to verify that the anchor divergence associated with this empirical distribution is exactly the softmax anchor divergence defined earlier. As established by Theorem 1 this will converge to the target population geometry as we get more and more anchors. With limited anchors, however, the finite anchor set may poorly represent the underlying semantics.

Since q determines the induced geometry, we can impose structure on the geometry by modeling q. The idea is to specify a parametric model q_θ for the anchor distribution. For example, we may take q_θ to be a multivariate Gaussian distribution with θ its mean and covariance. We then fit $\hat{\theta}$ to the specified anchors, say by maximum likelihood. The fitted distribution $q_{\hat{\theta}}$ then induces the corresponding anchor divergence through Definition 3. Thus, different anchor-distribution assumptions induce different structured geometries in image embedding space.

This reflects the usual tradeoff of parametric modeling. Simpler models impose stronger assumptions, which can help when anchors only partially specify the target semantics, but may miss the target geometry if those assumptions are too restrictive, even with many anchors. Moreover, once fit, the divergence is evaluated through the fitted distribution rather than all N observed anchors, so its evaluation cost need not scale with the anchor-set size; we examine this empirically in Appendix F.4.

Table 1 illustrates this perspective for three anchor models. The empirical distribution recovers softmax anchor divergence. The Gaussian model is the simplest parametric alternative: fitting its mean and covariance yields a quadratic CGF, and hence an elliptical geometry. The resulting geometry emphasizes directions along which the modeled anchor distribution varies. A Gaussian mixture allows richer structure: its CGF is a log-sum-exp of quadratic potentials, allowing the induced geometry to reflect multiple semantic modes rather than a single global covariance structure.

Bregman divergence in a projected anchor space. We also consider mildly generalizing from Bregman divergences on the contrastive embedding space to Bregman divergences on a subspace of the embedding space. This is motivated by evidence that semantic structure is often encoded in low-dimensional subspaces (Oldfield et al., 2023; Bhalla et al., 2024; Engels et al., 2025; Park et al., 2025). We learn a projection, optionally followed by normalization, onto a semantically relevant subspace from the anchor set and define the Anchor Divergence on the projected space. In Appendix B, we show that this corresponds to a latent-variable model of the anchor distribution and yields a procedure for estimating the projector. We refer to the resulting divergences as *latent Gaussian* and *latent spherical* anchor divergences. Operationally, both project onto the top-k eigenvectors of the empirical anchor covariance; the latent Gaussian divergence then uses Euclidean distance, while the latent spherical divergence uses cosine distance. [1]

[1] We note the latent spherical case yields the same projected cosine structure proposed in (Herzog & Wang, 2026)

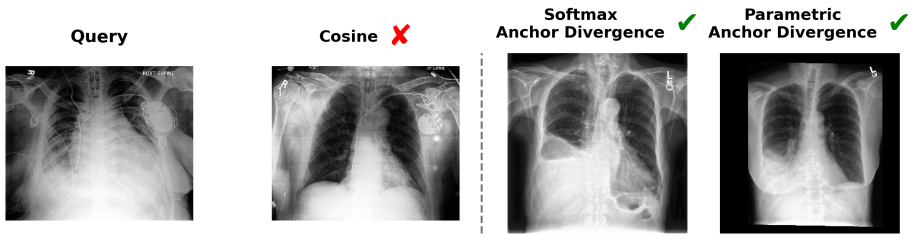

Figure 2: **Anchor divergences can emphasize relevant semantics over nuisance variation.** The query X-ray shows the target pathology, pleural effusion, as well as a support device. Cosine similarity retrieves an image sharing device-related structure but not the target pathology, whereas both softmax and parametric anchor divergences retrieve images with pleural effusion.

The empirical and parametric constructions provide different ways of representing the anchor distribution, and hence different approximations to the same target semantic geometry. Figure 2 provides a qualitative illustration of the usefulness of the anchor divergence framework and its parametric approximations: both emphasize the pathology of interest over nuisance device-related structure.

4 Adapting Semantic Geometries from Broad Anchors

Obtaining many target-specific anchors can be cumbersome, while broader related anchors may be easier to obtain. The distributional view enables an empirical-Bayes approach: broad anchors provide prior information about the parameters of the anchor distribution, which is then updated using scarce target anchors. For example, in pathology-specific retrieval, we may have abundant general clinical text but only a few precise descriptions of the target pathology. The goal is to specialize broad semantic structure using a small number of pathology-specific anchors.

Let Γ_{broad} and Γ_{target} denote the broad and target anchor sets, respectively. Let $\{q_\theta : \theta \in \Theta\}$ be a parametric anchor model with prior $p(\theta \mid \Omega)$. We adapt the anchor distribution in three steps:

1. fit q_θ to Γ_{broad} and use the fit to set the prior hyperparameters Ω;
2. update this prior using Γ_{target} via a model-specific posterior or posterior-like update;
3. summarize the update by $\hat{\theta}$, such as a posterior mean or MAP estimate.

The adapted distribution $q_{\hat{\theta}}$ induces the corresponding semantic geometry through its anchor divergence, giving the following definition.

Definition 4. *Given* Γ_{broad} *and* Γ_{target}*, let* $q_{\hat{\theta}}$ *denote the adapted anchor distribution obtained by using* Γ_{broad} *to define prior information and updating it with* Γ_{target}*. The* posterior anchor divergence *is the anchor divergence induced by* $q_{\hat{\theta}}$*.*

All methods target the same object: a Bregman geometry induced by an anchor distribution. They differ only in how that distribution is estimated. Softmax anchor divergence uses the empirical distribution over target anchors; parametric anchor divergence fits a structured model to those anchors; posterior anchor divergence further regularizes this structured estimate using broad, related anchors as prior information. When $|\Gamma_{\text{target}}|$ is small, broad anchors stabilize the geometry; as $|\Gamma_{\text{target}}|$ grows, target anchors increasingly determine it. Model-specific updates are in Appendix D.

5 Experiments

Our experiments are designed to test the geometric and statistical predictions of the framework. We organize the experiments around the following questions:

1. Can Anchor Divergences effectively specialize geometry to different semantic factors?
2. With limited anchors, can parametric models usefully approximate the target geometry?
3. Does posterior anchor modeling enable semantic adaptation using few target anchors?

Table 2: **Pathology-specific anchors specialize clinical retrieval.** Condition-wise AUPRC differences between softmax anchor divergence and each reference geometry; positive values favor anchor divergence. Results are averaged over 10 independently generated anchor sets and 10 CheXzero checkpoints, with 95% intervals accounting for bootstrap and anchor-generation variability. Values > 0.005 are bold, values in $[-0.005, 0.005]$ are gray, and values < -0.005 are bold red.

Pathology	Softmax − Cosine	Softmax − ITI
Enlarged Cardiomediastinum	-0.0002 ± 0.0003	-0.0045 ± 0.0010
Cardiomegaly	$\mathbf{0.0279 \pm 0.0034}$	$\mathbf{0.0069 \pm 0.0091}$
Lung Opacity	$\mathbf{0.0090 \pm 0.0032}$	$\mathbf{0.0137 \pm 0.0035}$
Lung Lesion	0.0001 ± 0.0005	$\mathbf{0.0051 \pm 0.0010}$
Edema	$\mathbf{0.0316 \pm 0.0044}$	$\mathbf{0.0286 \pm 0.0072}$
Consolidation	$\mathbf{0.0062 \pm 0.0014}$	$\mathbf{0.0073 \pm 0.0023}$
Pneumonia	$\mathbf{0.0081 \pm 0.0013}$	$\mathbf{0.0117 \pm 0.0018}$
Atelectasis	$\mathbf{0.0095 \pm 0.0027}$	$\mathbf{0.0113 \pm 0.0036}$
Pneumothorax	$\mathbf{0.0080 \pm 0.0013}$	$\mathbf{0.0128 \pm 0.0026}$
Pleural Effusion	$\mathbf{0.0465 \pm 0.0060}$	$\mathbf{0.0327 \pm 0.0098}$
Pleural Other	-0.0009 ± 0.0003	0.0025 ± 0.0005
Fracture	0.0003 ± 0.0002	0.0030 ± 0.0005
Support Devices	$\mathbf{0.0208 \pm 0.0056}$	$\mathbf{0.0366 \pm 0.0105}$
Average	$\mathbf{0.0128 \pm 0.0011}$	$\mathbf{0.0129 \pm 0.0014}$

5.1 Implementation details

All natural-image experiments (Russakovsky et al., 2015; Hendrycks et al., 2021; Sangkloy et al., 2016; Wah et al., 2011) use MetaCLIP-2 (Chuang et al., 2025). Class-name anchors use the prompt template "a photo of a {class}", and style anchors use "a {style} rendition of an object". For the clinical experiments, we use CheXpert (Irvin et al., 2019) with CheXzero (Tiu et al., 2022). Image and text embeddings are normalized to unit ℓ_2 norm, and we fix the inverse temperature $1/\tau$ at the model-provided logit scale.

For asymmetric divergences, we rank keys using the symmetrized score

$$d_{\mathrm{sym}}(\lambda_q, \lambda_k) = \frac{1}{2}\left[d(\lambda_q \| \lambda_k) + d(\lambda_k \| \lambda_q)\right].$$

We report Area Under the Precision–Recall Curve (AUPRC), which is suitable when relevant keys are sparse and prevalence varies across semantic targets. Relevance is defined by the semantic labels of each experiment.

We compare against two reference geometries. Cosine similarity provides the standard task-agnostic geometry. Image-text inversion (ITI) provides a task-specific heuristic baseline using a predict-then-retrieve procedure: it first matches the query image to the closest anchor and then ranks keys by cosine similarity to this anchor. This follows the spirit of image inversion methods in Baldrati et al. (2023); Mistretta et al. (2025). See Appendix E for full implementation details.

5.2 Anchor distributions select the semantic coordinate

We first isolate the effect of the anchor distribution with the embeddings fixed. Figure 1 gives a qualitative illustration. We test this systematically: if the anchor distribution determines the semantic geometry, then changing it should change which semantic coordinate is preserved by retrieval.

Clinical coordinate control. For each CheXpert finding, we independently generate 10 sets of report-style clinical text anchors from a fixed LLM prompt; see Appendix E.6. We evaluate retrieval using the corresponding pathology label and average over anchor generations and the 10 released CheXzero checkpoints. Table 2 reports AUPRC differences relative to cosine similarity and ITI, with 95% intervals accounting for validation-query and anchor-generation variability. Anchor divergence consistently matches or improves upon the reference geometries across findings, with largest gains for pleural effusion, cardiomegaly, edema, and support devices. Thus, pathology-specific anchor divergence emphasizes the intended clinical coordinate rather than generic radiographic similarity.

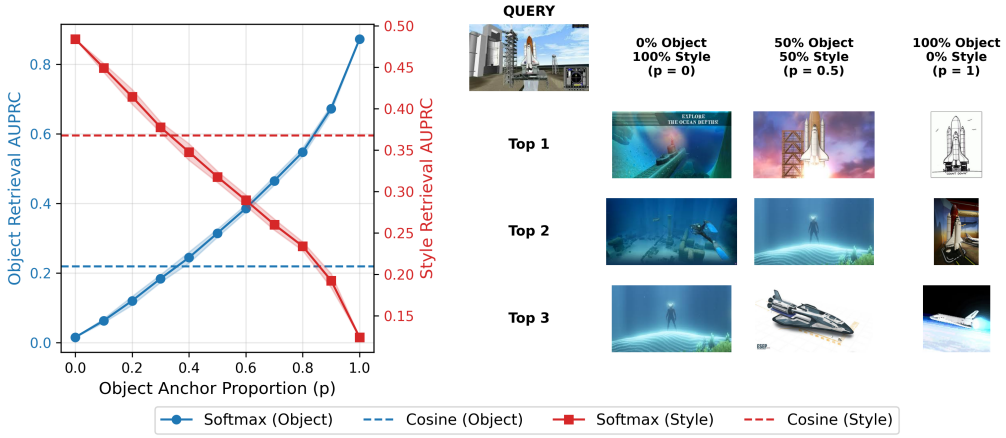

Figure 3: **Anchor distributions select the semantic coordinate. Left:** Mixing object and style anchor distributions on ImageNet-R shifts retrieval from preserving style to preserving object identity as p increases. **Right:** For a video-game-style space shuttle query, retrieved neighbors similarly shift from style-matched to object-matched images.

Object–style control. We next test whether varying the anchor distribution can interpolate between two semantic coordinates. ImageNet-R contains ImageNet object classes rendered in diverse styles, allowing retrieval to be evaluated by either object identity or style. Let q_{obj} and q_{style} denote the corresponding empirical anchor distributions, and mix them as $q_p = pq_{\text{obj}} + (1-p)q_{\text{style}}$, with $p \in [0, 1]$. We evaluate retrieval as p varies, averaging over 10 random realizations of the anchor set. Figure 3 shows that increasing object-anchor mass increases object retrieval while decreasing style retrieval.

Together, these experiments show that the same fixed image embeddings support different semantic geometries solely through changes to the anchor distribution.

5.3 Parametric modeling is useful with limited anchors

We next test whether parametric models of the anchor distribution provide useful approximations to the target geometry when only limited anchors are available. Softmax uses the empirical anchor distribution directly, whereas parametric models impose additional structure. We therefore expect parametric models to be most useful when anchors are limited, while the empirical construction should improve as anchor coverage grows.

We first vary the number of ImageNet class anchors and evaluate retrieval on ImageNet. Figure 4 left shows the intuitive sample-efficiency pattern: structured geometries recover useful semantic neighborhoods at moderate anchor budgets, while softmax improves with anchor coverage and eventually becomes strongest. Similar trends hold across the other datasets; see Appendix F.

We then consider the high-coverage regime using full dataset-matched anchor sets on ImageNet-R, CUB-200, and Sketchy. Figure 4 right shows that softmax anchor divergence is strongest in this setting, while parametric anchor divergences retain much of its gain over cosine similarity. Together with the ImageNet results, this illustrates the modeling tradeoff from Section 3: parametric structure can be useful with limited anchors, while the empirical anchor distribution becomes increasingly effective as anchor coverage grows.

5.4 Posterior modeling adapts broad geometry with few target anchors

Finally, we test whether broad semantic structure can be adapted toward a target geometry using only a few task-specific anchors. This setting exposes a limitation of softmax anchor divergence: as a nonparametric estimator, it places mass only on observed target anchors, so with few anchors it can yield a poor approximation of the target geometry. Posterior anchor models instead use broad, semantically related anchors as prior information and specialize this structure using target anchors.

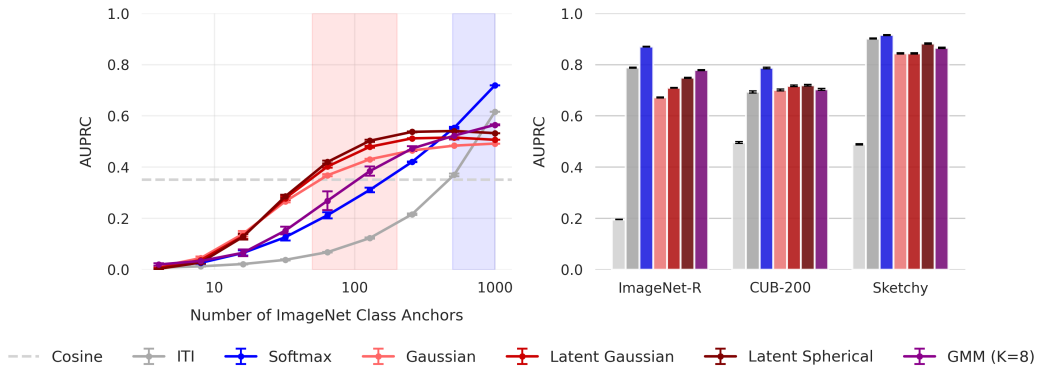

Figure 4: **Parametric modeling is useful with limited anchors. Left:** On ImageNet, parametric anchor divergences are strongest in the red-shaded limited-anchor regime, while softmax improves with coverage and becomes strongest in the blue high-coverage regime. **Right:** With full dataset-matched anchor sets on ImageNet-R, CUB-200, and Sketchy, softmax gives the strongest retrieval performance, while parametric anchor divergences retain much of its gain over cosine similarity.

We evaluate this on ImageNet, using WordNet anchors as broad semantic structure and ImageNet class names as target-specific anchors. We vary the number of target anchors and compare target-only softmax, latent spherical, and GMM anchor divergences with posterior latent spherical and posterior GMM models adapted from WordNet, alongside cosine similarity. Results are averaged over five random seeds; model-specific adaptation details are given in Appendix D.

Figure 5 shows that posterior anchor models are most useful when target anchors are scarce. Broad anchors provide prior semantic structure, while target anchors specialize the geometry toward the ImageNet class coordinate. As target-anchor coverage grows, the gap to target-only models narrows as the target anchors increasingly determine the geometry.

Summary. The experiments support the main consequences of the framework. Changing the anchor distribution changes which semantic factors govern similarity; parametric models provide useful approximations when anchors are limited, while the empirical softmax geometry improves with anchor coverage; and posterior anchor models adapt broad semantic structure using few task-specific anchors. Together, these results support anchor divergences as a flexible framework for specifying, modeling, and adapting context-dependent semantic geometry in contrastive representation spaces.

6 Discussion and Related Work

Semantic geometry in contrastive representations. Prior work has shown that learned representations often encode semantic structure geometrically, from word-embedding analogies and latent-variable accounts of embeddings to linear structure in neural representations (Mikolov et al., 2013; Arora et al., 2016; Saxe et al., 2019; Park et al., 2023; 2025). In vision–language models, related work has studied how CLIP-like representations organize concepts through directions, parts, subspaces, and interpretable coordinates (Radford et al., 2021; Chuang et al., 2025; Oldfield et al., 2023; Bhalla et al., 2024; Gandelsman et al., 2024; Merullo et al., 2022). Our question is complementary: given that a representation may encode many semantic factors simultaneously, which of them should determine similarity in a particular context? Anchor divergences answer this by representing the semantic context through an anchor distribution, which induces a corresponding Bregman geometry on the fixed representation space.

Image–image comparison in CLIP-like models. CLIP-style image embeddings are commonly compared using cosine similarity across applications (Abbasi et al., 2025; Zheng et al., 2025; Denner et al., 2025; Tiu et al., 2022). Several works have proposed alternatives to raw image–image cosine comparison, including cache-based adaptation, cross-modal aggregation, text-based inversion, and methods addressing image–text or intra-modal mismatch (Zhang et al., 2022; Udandarao et al., 2023;

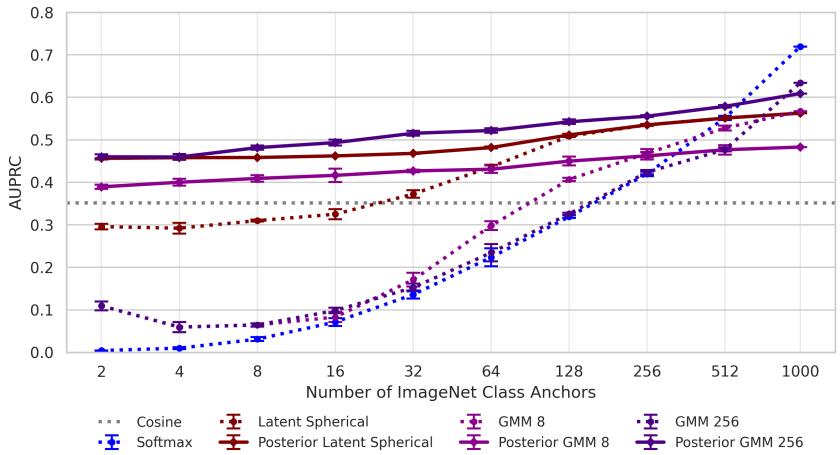

Figure 5: **Posterior anchor modeling adapts broad geometries using few target anchors.** Dotted curves use only target-specific ImageNet anchors, while solid curves use posterior anchor models adapted from WordNet with the same target anchors. Posterior models are strongest at small and moderate target-anchor budgets, with the advantage narrowing as coverage grows.

Yi et al., 2024; Mistretta et al., 2025; Baldrati et al., 2023; Herzog & Wang, 2026). These works motivate alternatives to the fixed cosine geometry, but do not generally ask how semantic context itself should determine the geometry on a fixed representation space.

Relation to TIP-X. TIP-X uses text as a bridge for image–image comparison: each image is mapped to a softmax affinity profile over downstream class-text prompts, and images are compared by the KL divergence between these profiles (Udandarao et al., 2023). This is the same finite-anchor construction as our softmax anchor divergence when the anchors are the downstream class-text prompts. Our contribution begins from this finite construction and studies the geometry determined by an underlying distribution over anchors. We show that an anchor distribution induces a population Bregman geometry and use this perspective to model and adapt semantic geometry.

Information geometry and metric learning. Our construction builds on classical connections between exponential families, KL divergence, and Bregman geometry (Efron, 1978; Amari, 1982; Banerjee et al., 2005; Wainwright & Jordan, 2008; Amari & Cichocki, 2010). It is also related to metric learning, which learns task-specific distances from supervised constraints on the representation space (Xing et al., 2002; Weinberger & Saul, 2009; Schroff et al., 2015; Movshovitz-Attias et al., 2017; Musgrave et al., 2020; Roth et al., 2020). Anchor divergences instead specify the semantic context through a distribution over semantic text anchors, without requiring labeled image pairs or representation fine-tuning. Because this anchor distribution determines the induced geometry, parametric modeling imposes corresponding geometric structure, while Bayesian updating provides a mechanism for semantic adaptation.

Future directions. Although our experiments focus on image–image retrieval in vision–language models, the construction is not intrinsically tied to images. The key ingredient is that a representation in one space defines a softmax distribution over anchors in another space through query–key scores. The theoretical structure developed here follows from this softmax, or equivalently exponential-family, form rather than from any image-specific property. In this view, a distribution over keys induces a geometry on the query space. This suggests that anchor divergences extend naturally to other contrastive and multimodal settings with the same structure. More broadly, this points to the importance of understanding the geometry of softmax-based models as a general problem in AI: recent work argues that representations defining softmax distributions are naturally studied through information geometry, with implications for probing and steering (Park et al., 2026). Progress on anchor divergences may therefore be of independent interest beyond retrieval, contributing to a broader theory of semantic geometry in softmax models.

AI USE STATEMENT

We used ChatGPT, GitHub Copilot, and Google Antigravity to assist with experimental code implementation. The research questions, conceptual framework, theoretical results, experimental design, evaluation protocols, and core helper code were developed by the authors. AI tools were primarily used to modify author-written implementations across datasets, anchor sets, anchor counts, and anchor models, and to implement bootstrap evaluation. As a validation step, AI-assisted implementations were first used to reproduce experiments that had already been implemented independently by the authors, obtaining matching results; the resulting code was also manually inspected.

AI was not used to formulate the main mathematical results or proofs. ChatGPT was used to assist in writing out some of the model-specific posterior update equations used in the adaptation framework; these expressions were subsequently checked by the authors.

We also used ChatGPT to generate the clinical report-style text anchors used in the CheXpert experiments. We generated multiple independent anchor sets from a fixed prompt and report results averaged across these generations; the prompt and evaluation procedure are provided in the appendix. The synthetic dog images used in Figure 1 were generated using the OpenAI image-generation API.

Additionally, AI was used to improve writing clarity and to assist with literature discovery. Closely related works and claims were checked against the original sources. Scientific interpretation, conclusions, and final manuscript content were determined and reviewed by the authors.

ETHICS STATEMENT

Our work studies the geometry of pretrained contrastive representations and does not involve the collection of new human-subject data. Experiments use existing benchmark datasets, including CheXpert for the clinical retrieval experiments.

The clinical experiments are intended to evaluate the proposed semantic-geometry framework and should not be interpreted as validation for clinical deployment or medical decision-making. Anchor divergences make the notion of similarity explicitly dependent on the supplied anchor distribution; inappropriate or biased anchors could therefore induce undesirable notions of semantic similarity. Any deployment in sensitive domains would require domain-specific validation, careful anchor construction, and assessment of potential biases and downstream harms.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce both the theoretical and empirical results. Complete proofs of the main theoretical claims are included in the appendix, together with derivations for the parametric and posterior anchor models. Section 5.1 and the appendix specify the datasets, pretrained models, anchor construction procedures, prompts, evaluation metrics, retrieval protocols, hyperparameters, and model-fitting details used in the experiments.

Starter code for the experiments is present in supplementary materials. The reported experiments use publicly available datasets and pretrained models, and all dataset splits and evaluation conventions are documented in the appendix.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9308–9317. IEEE, 2025.

Shun-Ichi Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, pp. 357–385, 1982.

Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15292–15301. IEEE, 2023.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. Metaclip 2: A worldwide scaling recipe. In *Advances in Neural Information Processing Systems*, 2025. to appear.

Stefan Denner, David Zimmerer, Dimitrios Bounias, Markus Bujotzek, Shuhan Xiao, Raphael Stock, Lisa Kausch, Philipp Schader, Tobias Penzkofer, Paul F Jäger, et al. Leveraging foundation models for content-based image retrieval in radiology. *Computers in Biology and Medicine*, 196: 110640, 2025.

Bradley Efron. The geometry of exponential families. *The Annals of Statistics*, pp. 362–376, 1978.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Yossi Gandelsman, Alexei Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *International conference on learning representations*, volume 2024, pp. 18395–18416, 2024.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Jonas Herzog and Yue Wang. Reevaluating the intra-modal misalignment hypothesis in clip. *arXiv preprint arXiv:2603.16100*, 2026.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *International Conference on Learning Representations*, volume 2025, pp. 77750–77771, 2025.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 360–368. IEEE, 2017.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pp. 681–699. Springer, 2020.

James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. Parts of speech–grounded subspaces in vision-language models. *Advances in Neural Information Processing Systems*, 36:2700–2724, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. *arXiv preprint arXiv:2602.15293*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International conference on machine learning*, pp. 8242–8252. PMLR, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *Acm Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2725–2736. IEEE, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *European conference on computer vision*, pp. 316–333. Springer, 2018.

Chao Yi, Lu Ren, De-Chuan Zhan, and Han-Jia Ye. Leveraging cross-modal neighbor representation for improved clip classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27392–27401. IEEE, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Sunyi Zheng, Xiaonan Cui, Yuxuan Sun, Jingxiong Li, Honglin Li, Yunlong Zhang, Pingyi Chen, Xueping Jing, Zhaoxiang Ye, and Lin Yang. Benchmarking pathclip for pathology image analysis. *Journal of Imaging Informatics in Medicine*, 38(1):422–438, 2025.

A Proof of Theorem 1

Theorem 1 (Softmax anchor divergence converges to a population Bregman divergence)**.** *Let* $\Gamma_N = \{\gamma_1, \dots, \gamma_N\}$ *be sampled i.i.d. from an anchor distribution* q *on* $\mathbb{S}^{d-1}$*. Then, as* $N \to \infty$*,*

$$\sup_{\lambda_1, \lambda_2 \in \mathbb{S}^{d-1}} \left| d_{\mathrm{SM}, \Gamma_N}(\lambda_1 \| \lambda_2) - B_{\kappa_q}\left(\frac{\lambda_2}{\tau} \middle\| \frac{\lambda_1}{\tau}\right) \right| \xrightarrow{\text{a.s.}} 0. \tag{1}$$

Proof. We begin by introducing some notation and definitions. We define

$$\eta = \frac{\lambda}{\tau}, \qquad \mathcal{E} := \left\{ \frac{\lambda}{\tau} : \lambda \in \mathcal{S}^{d-1} \right\} \subset \mathbb{R}^d.$$

Since $\mathcal{S}^{d-1}$ is compact, $\mathcal{E}$ is compact as well. We rewrite the population cumulant generating function as

$$\kappa(\eta) = \log \int e^{\eta^\top \gamma}\, q(d\gamma) = \log \mathbb{E}_q \left[e^{\eta^\top \gamma} \right].$$

For an anchor set $\Gamma_N = \{\gamma_1, \dots, \gamma_N\}$, the cumulant generating function of the empirical measure is

$$\widehat{\kappa}_N(\eta) := \log \left(\frac{1}{N} \sum_{i=1}^{N} e^{\eta^\top \gamma_i} \right).$$

The factor $1/N$ does not affect the induced softmax probabilities, but it is convenient for the law of large numbers. We also define

$$Z_N(\eta) := \frac{1}{N} \sum_{i=1}^{N} e^{\eta^\top \gamma_i}, \qquad Z(\eta) := \mathbb{E}_q[e^{\eta^\top \gamma}],$$

and

$$M_N(\eta) := \frac{1}{N} \sum_{i=1}^{N} \gamma_i e^{\eta^\top \gamma_i}, \qquad M(\eta) := \mathbb{E}_q[\gamma e^{\eta^\top \gamma}].$$

Throughout this proof, the anchors are sampled i.i.d. from q.

Now we are ready to prove the theorem. The first step is to rewrite the discrete softmax family as an empirical exponential family. Given $\Gamma_N = \{\gamma_1, \dots, \gamma_N\}$ and $\eta \in \mathbb{R}^d$,

$$\begin{aligned} p_N(i \mid \eta) &= \frac{e^{\eta^\top \gamma_i}}{\sum_{j=1}^{N} e^{\eta^\top \gamma_j}} \\ &= \frac{\frac{1}{N} e^{\eta^\top \gamma_i}}{\frac{1}{N} \sum_{j=1}^{N} e^{\eta^\top \gamma_j}} \\ &= \exp\left(\eta^\top \gamma_i - \widehat{\kappa}_N(\eta)\right) \cdot \frac{1}{N}. \end{aligned}$$

Thus, the discrete softmax is exactly the exponential tilt of the empirical anchor measure.

Let $\eta_a = \lambda_a/\tau$ for $a = 1, 2$. Then

$$\begin{aligned} d_{\mathrm{SM}}(\lambda_1 \| \lambda_2 \mid \Gamma_N) &= \sum_{i=1}^{N} p_N(i \mid \eta_1) \log \frac{p_N(i \mid \eta_1)}{p_N(i \mid \eta_2)} \\ &= \sum_{i=1}^{N} p_N(i \mid \eta_1) \Big((\eta_1 - \eta_2)^\top \gamma_i - \widehat{\kappa}_N(\eta_1) + \widehat{\kappa}_N(\eta_2) \Big). \end{aligned}$$

Since $\sum_i p_N(i \mid \eta_1) = 1$, this becomes

$$d_{\mathrm{SM}}(\lambda_1 \| \lambda_2 \mid \Gamma_N) = (\eta_1 - \eta_2)^\top \sum_{i=1}^{N} p_N(i \mid \eta_1) \gamma_i - \widehat{\kappa}_N(\eta_1) + \widehat{\kappa}_N(\eta_2). \tag{3}$$

But

$$\nabla \widehat{\kappa}_N(\eta) = \frac{\sum_{i=1}^{N} \gamma_i e^{\eta^\top \gamma_i}}{\sum_{j=1}^{N} e^{\eta^\top \gamma_j}} = \sum_{i=1}^{N} p_N(i \mid \eta) \gamma_i,$$

so Equation (3) can be written as

$$d_{\mathrm{SM}}(\lambda_1 \| \lambda_2 \mid \Gamma_N) = \widehat{\kappa}_N(\eta_2) - \widehat{\kappa}_N(\eta_1) + (\eta_1 - \eta_2)^\top \nabla \widehat{\kappa}_N(\eta_1). \tag{4}$$

Equivalently,

$$d_{\mathrm{SM}}(\lambda_1 \| \lambda_2 \mid \Gamma_N) = B_{\widehat{\kappa}_N}(\eta_2 \| \eta_1), \tag{5}$$

where $B_f(y \| x) := f(y) - f(x) - \nabla f(x)^\top (y - x)$.

We now show uniform convergence of Z_N to Z on $\mathcal{E}$. Since $\|\eta\| = 1/\tau$ for $\eta \in \mathcal{E}$ and $\|\gamma\| = 1$, we have

$$0 < e^{\eta^\top \gamma} \le e^{1/\tau} \qquad \text{for all } \eta \in \mathcal{E},\ \gamma \in \mathcal{S}^{d-1}.$$

Moreover, for any $\eta, \eta' \in \mathcal{E}$ and $\gamma \in \mathcal{S}^{d-1}$,

$$\left| e^{\eta^\top \gamma} - e^{(\eta')^\top \gamma} \right| \le e^{1/\tau} \, |(\eta - \eta')^\top \gamma| \le e^{1/\tau} \|\eta - \eta'\|.$$

Therefore, both Z_N and Z are Lipschitz on $\mathcal{E}$ with deterministic Lipschitz constant at most $e^{1/\tau}$:

$$|Z_N(\eta) - Z_N(\eta')| \le e^{1/\tau} \|\eta - \eta'\|, \qquad |Z(\eta) - Z(\eta')| \le e^{1/\tau} \|\eta - \eta'\|.$$

Fix $\varepsilon > 0$, and let $T_\varepsilon \subset \mathcal{E}$ be a finite ε-net of the compact set $\mathcal{E}$. For each fixed $\eta \in T_\varepsilon$, the strong law of large numbers gives

$$Z_N(\eta) \xrightarrow{\text{a.s.}} Z(\eta).$$

Since T_ε is finite, it follows that

$$\max_{\eta \in T_\varepsilon} |Z_N(\eta) - Z(\eta)| \xrightarrow{\text{a.s.}} 0.$$

Now let $\eta \in \mathcal{E}$, and choose $\eta_\varepsilon \in T_\varepsilon$ with $\|\eta - \eta_\varepsilon\| \le \varepsilon$. Then

$$\begin{aligned} |Z_N(\eta) - Z(\eta)| &\le |Z_N(\eta) - Z_N(\eta_\varepsilon)| + |Z_N(\eta_\varepsilon) - Z(\eta_\varepsilon)| + |Z(\eta_\varepsilon) - Z(\eta)| \\ &\le 2e^{1/\tau} \varepsilon + \max_{\eta' \in T_\varepsilon} |Z_N(\eta') - Z(\eta')|. \end{aligned}$$

Taking the supremum over $\eta \in \mathcal{E}$ yields

$$\sup_{\eta \in \mathcal{E}} |Z_N(\eta) - Z(\eta)| \le 2e^{1/\tau} \varepsilon + \max_{\eta' \in T_\varepsilon} |Z_N(\eta') - Z(\eta')|.$$

Hence, almost surely,

$$\limsup_{N\to\infty} \sup_{\eta\in\mathcal{E}} |Z_N(\eta) - Z(\eta)| \le 2e^{1/\tau}\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude

$$\sup_{\eta\in\mathcal{E}} |Z_N(\eta) - Z(\eta)| \xrightarrow{\text{a.s.}} 0. \tag{6}$$

Using the same Lipschitz argument, for any $\eta, \eta' \in \mathcal{E}$ and $\gamma \in \mathcal{S}^{d-1}$,

$$\|\gamma e^{\eta^\top \gamma} - \gamma e^{(\eta')^\top \gamma}\| \le e^{1/\tau}\|\eta - \eta'\|.$$

Therefore, both M_N and M are Lipschitz on $\mathcal{E}$ with deterministic Lipschitz constant at most $e^{1/\tau}$:

$$\|M_N(\eta) - M_N(\eta')\| \le e^{1/\tau}\|\eta - \eta'\|, \qquad \|M(\eta) - M(\eta')\| \le e^{1/\tau}\|\eta - \eta'\|.$$

Applying the same finite-net argument componentwise gives

$$\sup_{\eta\in\mathcal{E}} \|M_N(\eta) - M(\eta)\| \xrightarrow{\text{a.s.}} 0. \tag{7}$$

Since

$$\|\gamma e^{\eta^\top \gamma}\| \le e^{1/\tau} \qquad \text{for all } \eta \in \mathcal{E},\ \gamma \in \mathcal{S}^{d-1},$$

dominated convergence justifies differentiation under the expectation, so

$$\nabla Z(\eta) = \mathbb{E}_q[\gamma e^{\eta^\top \gamma}] = M(\eta).$$

Hence,

$$\nabla\kappa(\eta) = \frac{\nabla Z(\eta)}{Z(\eta)} = \frac{M(\eta)}{Z(\eta)}, \qquad \nabla\widehat{\kappa}_N(\eta) = \frac{M_N(\eta)}{Z_N(\eta)}.$$

By Equation (6), $Z_N \to Z$ uniformly almost surely on $\mathcal{E}$. Also, $Z(\eta) > 0$ for every η, and Z is continuous on compact $\mathcal{E}$, so

$$m := \inf_{\eta\in\mathcal{E}} Z(\eta) > 0.$$

Therefore, almost surely for all sufficiently large N,

$$\inf_{\eta\in\mathcal{E}} Z_N(\eta) \ge \frac{m}{2} > 0.$$

Because $\log$ is Lipschitz on $[m/2, \infty)$, Equation (6) implies

$$\sup_{\eta\in\mathcal{E}} |\widehat{\kappa}_N(\eta) - \kappa(\eta)| \xrightarrow{\text{a.s.}} 0. \tag{8}$$

Using Equation (6) and Equation (7), together with the lower bound on the denominators, we obtain

$$\sup_{\eta\in\mathcal{E}} \|\nabla\widehat{\kappa}_N(\eta) - \nabla\kappa(\eta)\| \xrightarrow{\text{a.s.}} 0. \tag{9}$$

Recall that

$$B_\kappa(\eta_2\|\eta_1) = \kappa(\eta_2) - \kappa(\eta_1) - \nabla\kappa(\eta_1)^\top(\eta_2 - \eta_1) = \kappa(\eta_2) - \kappa(\eta_1) + (\eta_1 - \eta_2)^\top\nabla\kappa(\eta_1).$$

Hence,

$$\begin{aligned}
&\sup_{\eta_1,\eta_2\in\mathcal{E}} |d_{\mathrm{SM}}(\tau\eta_1\|\tau\eta_2 \mid \Gamma_N) - B_\kappa(\eta_2\|\eta_1)| \\
&\le 2\sup_{\eta\in\mathcal{E}} |\widehat{\kappa}_N(\eta) - \kappa(\eta)| + \sup_{\eta_1,\eta_2\in\mathcal{E}} \left|(\eta_1 - \eta_2)^\top\left(\nabla\widehat{\kappa}_N(\eta_1) - \nabla\kappa(\eta_1)\right)\right|.
\end{aligned}$$

Since $\mathcal{E} = \tau^{-1}\mathcal{S}^{d-1}$,

$$\sup_{\eta_1,\eta_2\in\mathcal{E}} \|\eta_1 - \eta_2\| \le \frac{2}{\tau}.$$

Therefore,

$$\sup_{\eta_1,\eta_2\in\mathcal{E}} |d_{\mathrm{SM}}(\tau\eta_1\|\tau\eta_2 \mid \Gamma_N) - B_\kappa(\eta_2\|\eta_1)|$$
$$\le 2\sup_{\eta\in\mathcal{E}} |\widehat{\kappa}_N(\eta) - \kappa(\eta)| + \frac{2}{\tau}\sup_{\eta\in\mathcal{E}} \|\nabla\widehat{\kappa}_N(\eta) - \nabla\kappa(\eta)\|.$$

By Equation (8) and Equation (9), the right-hand side converges almost surely to zero. Replacing $\eta_a = \lambda_a/\tau$ gives

$$\sup_{\lambda_1,\lambda_2\in\mathcal{S}^{d-1}} \left| d_{\mathrm{SM}}(\lambda_1\|\lambda_2 \mid \Gamma_N) - B_\kappa\left(\frac{\lambda_2}{\tau}\middle\|\frac{\lambda_1}{\tau}\right)\right| \xrightarrow[N\to\infty]{\text{a.s.}} 0.$$

This proves the claim. □

B Structured approximations to the population anchor geometry

The population anchor divergence is determined by the anchor distribution through its cumulant generating function. To obtain tractable estimators of this geometry, we consider structured parametric approximations to the anchor distribution. These models should be understood as pragmatic structured surrogates for the population anchor geometry, not as literal distributional models of normalized contrastive embeddings. The Gaussian and Gaussian-mixture models place mass off the unit sphere, while the latent low-rank models additionally idealize the semantically relevant variation as concentrated in a lower-dimensional subspace. We use these models because they yield tractable geometries with useful statistical tradeoffs in fidelity, robustness, and sample efficiency.

B.1 General framework

We begin with a latent semantic variable $z \in \mathcal{Z} \subseteq \mathbb{R}^k$, representing text-side semantic information, and a corresponding latent natural parameter $\eta \in \mathcal{H} \subseteq \mathbb{R}^k$, representing the image-side semantic parameter. Given a base measure q_Z on $\mathcal{Z}$, we let η parameterize z through the exponential family

$$p(dz \mid \eta) = \exp\big(\eta^\top z - \kappa_Z(\eta)\big)\, q_Z(dz), \tag{10}$$

whose cumulant generating function is

$$\kappa_Z(\eta) := \log \int \exp(\eta^\top z)\, q_Z(dz). \tag{11}$$

The associated latent semantic geometry is the Bregman divergence

$$d_Z(\eta_1\|\eta_2) := B_{\kappa_Z}(\eta_2\|\eta_1). \tag{12}$$

To relate this latent construction to the anchor divergence used in the main text, let

$$\gamma = T(z), \tag{13}$$

where T maps latent text-side semantics into ambient text embeddings. λ is then defined as the reparameterization of η so that λ/τ is the natural parameter for the exponential family with sufficient statistic γ, aligning with the CLIP scoring rule. We consider two cases:

Identity case. When $T = \mathrm{Id}$, we have $z = \gamma$, so

$$p(d\gamma \mid \eta) = \exp\big(\eta^\top \gamma - \kappa_Z(\eta)\big)\, q_Z(d\gamma), \tag{14}$$

and hence $\frac{\lambda}{\tau} = \eta$. In this case, Equation (11) reduces to the cumulant generating function of the anchor distribution, and Equation (12) recovers the population anchor divergence from the main text.

Low-rank affine case. For the latent low-rank models, we take

$$T(z) = \mu + Uz, \tag{15}$$

where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns. Thus, latent semantics are realized in the affine subspace $\mu + \mathrm{Im}(U) \subset \mathbb{R}^d$. Since T is injective onto this subspace, the latent exponential family induces a pushed-forward family on ambient anchors. Writing $q = T_{\#}q_Z$ for the pushforward base measure, we obtain

$$p(d\gamma \mid \eta) = \exp\big(\eta^\top U^\top(\gamma - \mu) - \kappa_Z(\eta)\big)\, q(d\gamma). \tag{16}$$

To align this family with the CLIP score, we parameterize the image side by

$$\frac{\lambda}{\tau} = U\eta. \tag{17}$$

So the pushed-forward family can be written in the score-aligned form

$$p(d\gamma \mid \lambda) = \exp\left(\frac{1}{\tau}\lambda^\top\gamma - \left[\kappa_Z\left(\frac{1}{\tau}U^\top\lambda\right) + \frac{1}{\tau}\lambda^\top\mu\right]\right) q(d\gamma). \tag{18}$$

The relation $\lambda/\tau = U\eta$ defines an exact inverse only on the semantic subspace $\mathrm{Im}(U)$. Because the low-rank models are idealized approximations, general ambient image embeddings need not lie exactly in this subspace, so we use a model-specific extended inverse or projection map $\eta(\lambda)$. The induced image–image semantic divergence is then defined by pulling back the latent geometry:

$$d(\lambda_1\|\lambda_2) := d_Z\big(\eta(\lambda_1)\|\eta(\lambda_2)\big) = B_{\kappa_Z}\big(\eta(\lambda_2)\|\eta(\lambda_1)\big). \tag{19}$$

Thus the semantic geometry is computed in latent space, while ambient text and image embeddings provide its realization and parameterization.

B.2 Gaussian anchor model

We first recover the ambient-space Gaussian model as the identity case of the general framework. Take

$$z \in \mathbb{R}^d, \qquad q_Z = \mathcal{N}(\mu, \Sigma), \qquad T = \mathrm{Id}, \tag{20}$$

so that $z = \gamma$. By the identity-case discussion in Appendix B.1, the image-side natural parameter is

$$\eta = \frac{\lambda}{\tau}. \tag{21}$$

Thus the Gaussian anchor model is simply the case in which the latent semantic space coincides with the ambient anchor space.

Proposition 1 (Gaussian anchor model)**.** *For the Gaussian anchor model,*

$$d_{\mathrm{G}}(\lambda_1\|\lambda_2) = \frac{1}{2\tau^2}(\lambda_1 - \lambda_2)^\top\Sigma(\lambda_1 - \lambda_2). \tag{22}$$

Proof. For $q_Z = \mathcal{N}(\mu, \Sigma)$, the cumulant generating function is

$$\kappa_Z(\eta) = \log \int \exp(\eta^\top z)\, \mathcal{N}(dz; \mu, \Sigma) \tag{23}$$

$$= \mu^\top\eta + \frac{1}{2}\eta^\top\Sigma\eta. \tag{24}$$

Hence

$$\nabla\kappa_Z(\eta) = \mu + \Sigma\eta. \tag{25}$$

Substituting into Equation (12) gives

$$d_Z(\eta_1\|\eta_2) = B_{\kappa_Z}(\eta_2\|\eta_1) \tag{26}$$

$$= \frac{1}{2}(\eta_2 - \eta_1)^\top\Sigma(\eta_2 - \eta_1). \tag{27}$$

Using $\eta(\lambda) = \lambda/\tau$ in Equation (19),

$$d_{\mathrm{G}}(\lambda_1\|\lambda_2) = d_Z\left(\frac{\lambda_1}{\tau}\middle\|\frac{\lambda_2}{\tau}\right) \tag{28}$$

$$= \frac{1}{2\tau^2}(\lambda_1 - \lambda_2)^\top\Sigma(\lambda_1 - \lambda_2), \tag{29}$$

which proves Equation (22). □

B.3 LATENT GAUSSIAN MODEL

We now impose Gaussian structure only on a lower-dimensional latent semantic variable. Take

$$z \in \mathbb{R}^k, \qquad q_Z = \mathcal{N}(0, I_k), \qquad T(z) = \mu + Uz, \tag{30}$$

where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns. By the low-rank affine construction in Appendix B.1, the image-side parameter satisfies

$$\frac{\lambda}{\tau} = U\eta, \tag{31}$$

and for a general ambient image embedding we use the canonical projection

$$\eta(\lambda) = \frac{1}{\tau}U^\top\lambda. \tag{32}$$

Thus the image–image divergence is obtained by pulling back the latent Gaussian geometry to the semantic subspace.

Proposition 2 (Latent Gaussian model)**.** *For the latent Gaussian model with semantic subspace U,*

$$d_{\mathrm{LG}}(\lambda_1\|\lambda_2) = \frac{1}{2\tau^2}(\lambda_1-\lambda_2)^\top UU^\top(\lambda_1-\lambda_2). \tag{33}$$

Proof. For the standard Gaussian base measure on $\mathbb{R}^k$,

$$\kappa_Z(\eta) = \log\int \exp(\eta^\top z)\,\mathcal{N}(dz; 0, I_k) \tag{34}$$

$$= \frac{1}{2}\eta^\top\eta. \tag{35}$$

Hence

$$\nabla\kappa_Z(\eta) = \eta, \tag{36}$$

and therefore

$$d_Z(\eta_1\|\eta_2) = B_{\kappa_Z}(\eta_2\|\eta_1) = \frac{1}{2}\|\eta_2-\eta_1\|_2^2. \tag{37}$$

Using Equation (19) together with Equation (32),

$$d_{\mathrm{LG}}(\lambda_1\|\lambda_2) = \frac{1}{2}\left\|\frac{1}{\tau}U^\top\lambda_2 - \frac{1}{\tau}U^\top\lambda_1\right\|_2^2 \tag{38}$$

$$= \frac{1}{2\tau^2}\|U^\top(\lambda_2-\lambda_1)\|_2^2 \tag{39}$$

$$= \frac{1}{2\tau^2}(\lambda_1-\lambda_2)^\top UU^\top(\lambda_1-\lambda_2), \tag{40}$$

which proves Equation (33). □

B.4 LATENT SPHERICAL MODEL

The latent spherical model keeps only directional semantic variation within the same low-dimensional semantic subspace. We take

$$z \in \mathbb{S}^{k-1}, \qquad \eta \in \mathbb{S}^{k-1}, \qquad q_Z = \sigma := \mathrm{Unif}(\mathbb{S}^{k-1}), \qquad T(z) = \mu + Uz, \tag{41}$$

where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns. By the low-rank affine construction in Appendix B.1, the image-side parameter satisfies

$$\frac{\lambda}{\tau} = U\eta, \tag{42}$$

and for a general ambient image embedding we use the projected direction

$$\eta(\lambda) := \frac{U^\top\lambda}{\|U^\top\lambda\|_2}, \qquad \text{whenever } U^\top\lambda \neq 0. \tag{43}$$

Thus the image–image divergence is obtained by pulling back the latent directional geometry to the semantic subspace.

Proposition 3 (Latent spherical model)**.** *For the latent spherical model with semantic subspace U, and assuming $U^\top \lambda_1 \neq 0$ and $U^\top \lambda_2 \neq 0$,*

$$d_{\mathrm{LS}}(\lambda_1 \| \lambda_2) = A_k(1) \left(1 - \frac{(U^\top \lambda_1)^\top (U^\top \lambda_2)}{\|U^\top \lambda_1\|_2 \, \|U^\top \lambda_2\|_2}\right), \tag{44}$$

where

$$A_k(r) = \frac{I_{k/2}(r)}{I_{k/2-1}(r)},$$

and I_ν denotes the modified Bessel function of the first kind.

Proof. For the uniform base measure on $\mathbb{S}^{k-1}$, the cumulant generating function depends only on the norm:

$$\kappa_Z(\eta) = \psi_k(\|\eta\|_2), \tag{45}$$

where ψ_k is the usual von Mises–Fisher log-normalizer. Its gradient is radial:

$$\nabla \kappa_Z(\eta) = A_k(\|\eta\|_2) \, \frac{\eta}{\|\eta\|_2}. \tag{46}$$

Now let $\eta_1, \eta_2 \in \mathbb{S}^{k-1}$. Since $\|\eta_1\|_2 = \|\eta_2\|_2 = 1$,

$$\kappa_Z(\eta_1) = \kappa_Z(\eta_2) = \psi_k(1), \qquad \nabla \kappa_Z(\eta_1) = A_k(1)\eta_1. \tag{47}$$

Therefore,

$$\begin{aligned} d_Z(\eta_1 \| \eta_2) &= B_{\kappa_Z}(\eta_2 \| \eta_1) && (48)\\ &= \kappa_Z(\eta_2) - \kappa_Z(\eta_1) - \nabla \kappa_Z(\eta_1)^\top (\eta_2 - \eta_1) && (49)\\ &= -A_k(1)\eta_1^\top (\eta_2 - \eta_1) && (50)\\ &= A_k(1)\big(1 - \eta_1^\top \eta_2\big). && (51) \end{aligned}$$

Using Equation (19) together with Equation (43),

$$\begin{aligned} d_{\mathrm{LS}}(\lambda_1 \| \lambda_2) &= A_k(1) \left(1 - \eta(\lambda_1)^\top \eta(\lambda_2)\right) && (52)\\ &= A_k(1) \left(1 - \frac{(U^\top \lambda_1)^\top (U^\top \lambda_2)}{\|U^\top \lambda_1\|_2 \, \|U^\top \lambda_2\|_2}\right), && (53) \end{aligned}$$

which proves Equation (44). □

B.5 Gaussian mixture model

Finally, we obtain a more expressive ambient-space approximation by modeling the anchor distribution as a Gaussian mixture. Take

$$z \in \mathbb{R}^d, \qquad q_Z(dz) = \sum_{m=1}^{M} \pi_m \, \mathcal{N}(dz; \mu_m, \Sigma_m), \qquad T = \mathrm{Id}, \tag{54}$$

where $\pi_m > 0$, $\sum_{m=1}^{M} \pi_m = 1$, and $\Sigma_m \succ 0$. Thus, $z = \gamma$. By the identity-case construction in Appendix B.1, the image-side natural parameter is

$$\eta = \frac{\lambda}{\tau}. \tag{55}$$

This yields a multimodal ambient approximation to the population anchor geometry.

Proposition 4 (Gaussian mixture anchor model)**.** *For the Gaussian mixture anchor model,*

$$d_{\mathrm{GMM}}(\lambda_1 \| \lambda_2) = B_{\kappa_{\mathrm{GMM}}}\left(\frac{\lambda_2}{\tau} \Big\| \frac{\lambda_1}{\tau}\right), \tag{56}$$

where

$$\kappa_{\mathrm{GMM}}(\eta) = \log \sum_{m=1}^{M} \pi_m \exp\left(\mu_m^\top \eta + \frac{1}{2}\eta^\top \Sigma_m \eta\right). \tag{57}$$

Proof. Starting from Equation (54),

$$\kappa_Z(\eta) = \log \int \exp(\eta^\top z)\, q_Z(dz) \tag{58}$$

$$= \log \sum_{m=1}^{M} \pi_m \int \exp(\eta^\top z)\, \mathcal{N}(dz; \mu_m, \Sigma_m). \tag{59}$$

Each inner integral is the Gaussian MGF, so

$$\kappa_Z(\eta) = \log \sum_{m=1}^{M} \pi_m \exp\left(\mu_m^\top \eta + \frac{1}{2}\eta^\top \Sigma_m \eta\right), \tag{60}$$

which is exactly Equation (57). Hence,

$$d_Z(\eta_1 \| \eta_2) = B_{\kappa_{\mathrm{GMM}}}(\eta_2 \| \eta_1). \tag{61}$$

Using $\eta(\lambda) = \lambda/\tau$ in Equation (19),

$$d_{\mathrm{GMM}}(\lambda_1 \| \lambda_2) = d_Z\left(\frac{\lambda_1}{\tau} \middle\| \frac{\lambda_2}{\tau}\right) \tag{62}$$

$$= B_{\kappa_{\mathrm{GMM}}}\left(\frac{\lambda_2}{\tau} \middle\| \frac{\lambda_1}{\tau}\right), \tag{63}$$

which proves Equation (56). □

C Parameter estimation for structured anchor models

Given anchor embeddings $\{\gamma_i\}_{i=1}^n \subset \mathbb{R}^d$, we estimate the parameters of each structured anchor model directly from the anchor set.

Gaussian anchor model. For the Gaussian model, we fit the mean and covariance by maximum likelihood:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \gamma_i, \tag{64}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\gamma_i - \hat{\mu})(\gamma_i - \hat{\mu})^\top. \tag{65}$$

The fitted semantic divergence is obtained by substituting $\hat{\Sigma}$ into Equation (22).

Low-rank models. For the latent Gaussian and latent spherical models, we estimate the semantic subspace $U \in \mathbb{R}^{d\times k}$ by the top-k eigenvectors of the anchor covariance matrix

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\gamma_i - \bar{\gamma})(\gamma_i - \bar{\gamma})^\top, \qquad \bar{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \gamma_i. \tag{66}$$

We choose k to explain 95% of the anchor covariance. The fitted divergences are then obtained by substituting $\hat{U}$ into Equation (33) and Equation (44), respectively.

Gaussian mixture model. For the Gaussian mixture model, we fit the mixture weights, means, and covariances,

$$\{(\pi_m, \mu_m, \Sigma_m)\}_{m=1}^{M},$$

by expectation–maximization (EM) on the anchor embeddings. The fitted parameters are then substituted into Equation (57) and Equation (56).

D MODEL-SPECIFIC POSTERIOR AND POSTERIOR-LIKE UPDATES FOR POSTERIOR ANCHOR MODELING

This appendix gives model-specific posterior and posterior-like updates for the posterior anchor modeling framework introduced in Section 4. Broad anchors provide prior semantic structure for a parametric anchor model, precise task-aligned anchors refine that structure, and the resulting model summary induces an adapted geometry through $B_{\kappa_{\hat{\theta}}}$.

D.1 GENERAL ADAPTATION FRAMEWORK

Suppose we are given a broad anchor set

$$\Gamma_{\text{broad}} = \{\gamma_1^{\text{broad}}, \dots, \gamma_m^{\text{broad}}\}$$

and a smaller set of precise task-aligned anchors

$$\Gamma_{\text{target}} = \{\gamma_1^{\text{target}}, \dots, \gamma_n^{\text{target}}\}.$$

Let $\{q_\theta : \theta \in \Theta\}$ be a parametric family of anchor distributions together with model-specific prior structure on Θ. A generic adaptation procedure is:

1. fit the parametric family on Γ_{broad} to obtain a broad-anchor estimate θ_{broad};
2. choose hyperparameters or regularization structure $\Omega = \Omega(\theta_{\text{broad}})$ so that the update is centered around the broad semantic structure encoded by θ_{broad};
3. combine this prior structure with the precise anchors to obtain a model-specific posterior or posterior-like update over θ;
4. use a model-specific summary $\hat{\theta}$ to define the adapted anchor model $q_{\hat{\theta}}$;
5. induce the adapted geometry through

$$\kappa_{\hat{\theta}}(\xi) = \log \int \exp(\xi^\top \gamma)\, q_{\hat{\theta}}(d\gamma), \qquad d_{\hat{\theta}}(\lambda_1 \| \lambda_2) = B_{\kappa_{\hat{\theta}}}\left(\frac{\lambda_2}{\tau} \middle\| \frac{\lambda_1}{\tau}\right).$$

We now specialize this framework to the model families used in the paper.

D.2 GAUSSIAN ANCHOR MODEL WITH NORMAL–INVERSE-WISHART UPDATE

For the Gaussian anchor model,

$$q_\theta(\gamma) = \mathcal{N}(\gamma; \mu, \Sigma), \qquad \theta = (\mu, \Sigma), \qquad \Sigma \succ 0.$$

As shown in Appendix B.2, the induced geometry is

$$d_{\text{G}}(\lambda_1 \| \lambda_2) = \frac{1}{2\tau^2}(\lambda_1 - \lambda_2)^\top \Sigma (\lambda_1 - \lambda_2),$$

so adaptation corresponds to updating the covariance structure of the anchor distribution.

We place a Normal–Inverse-Wishart prior on (μ, Σ):

$$\Sigma \sim \mathcal{IW}(\Psi_0, \nu_0),$$
$$\mu \mid \Sigma \sim \mathcal{N}\left(\mu_0, \frac{1}{\kappa_0}\Sigma\right).$$

The adaptation steps are:

1. **Fit the broad Gaussian.** Fit a Gaussian model to the broad anchors:

$$\bar{\gamma}_{\text{broad}} = \frac{1}{m}\sum_{i=1}^{m} \gamma_i^{\text{broad}},$$

$$S_{\text{broad}} = \sum_{i=1}^{m} (\gamma_i^{\text{broad}} - \bar{\gamma}_{\text{broad}})(\gamma_i^{\text{broad}} - \bar{\gamma}_{\text{broad}})^\top.$$

This gives the broad-anchor estimate $(\mu_{\text{broad}}, \Sigma_{\text{broad}})$.

2. **Initialize the prior.** Use the broad-anchor estimate in an empirical-Bayes step:

$$\mu_0 = \bar{\gamma}_{\text{broad}},$$
$$\Psi_0 = (\nu_0 - d - 1)\Sigma_{\text{broad}},$$

where κ_0 and ν_0 control prior strength.

3. **Update with precise anchors.** Given precise anchors Γ_{target}, compute

$$\bar{\gamma}_{\text{target}} = \frac{1}{n}\sum_{i=1}^{n} \gamma_i^{\text{target}},$$
$$S_{\text{target}} = \sum_{i=1}^{n} (\gamma_i^{\text{target}} - \bar{\gamma}_{\text{target}})(\gamma_i^{\text{target}} - \bar{\gamma}_{\text{target}})^\top.$$

The posterior is again NIW with parameters

$$\kappa_n = \kappa_0 + n,$$
$$\nu_n = \nu_0 + n,$$
$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{\gamma}_{\text{target}}}{\kappa_0 + n},$$
$$\Psi_n = \Psi_0 + S_{\text{target}} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\gamma}_{\text{target}} - \mu_0)(\bar{\gamma}_{\text{target}} - \mu_0)^\top.$$

4. **Form the adapted model.** Use a posterior summary $(\hat{\mu}, \hat{\Sigma})$, such as the posterior mean or MAP estimate, to define the adapted Gaussian anchor model

$$q_{\hat{\theta}}(\gamma) = \mathcal{N}(\gamma; \hat{\mu}, \hat{\Sigma}).$$

5. **Induce the adapted geometry.** The adapted geometry is

$$d_{\text{G}}^{\text{adapt}}(\lambda_1 \| \lambda_2) = \frac{1}{2\tau^2}(\lambda_1 - \lambda_2)^\top \hat{\Sigma}(\lambda_1 - \lambda_2).$$

D.3 Latent Gaussian model with matrix Bingham update

For the latent Gaussian model, the anchor distribution is determined by a low-dimensional semantic subspace. Concretely, we use the ambient surrogate

$$q_\theta(\gamma) = \mathcal{N}(\gamma; \mu, UU^\top + \varepsilon I), \qquad \theta = (\mu, U),$$

where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns, $\mu \in \mathbb{R}^d$, and $\varepsilon > 0$ is a small isotropic regularization used in implementation. As shown in Appendix B.3:

$$d_{\text{LG}}(\lambda_1 \| \lambda_2) = \frac{1}{2\tau^2}(\lambda_1 - \lambda_2)^\top UU^\top (\lambda_1 - \lambda_2).$$

Thus, adaptation amounts to updating the subspace U.

We place a matrix Bingham prior on U over the Stiefel manifold

$$\text{St}(d, k) = \{U \in \mathbb{R}^{d\times k} : U^\top U = I_k\},$$

with density

$$p(U \mid A) \propto \exp\big(\text{tr}(U^\top A U)\big).$$

Under the Gaussian covariance model above, this prior is conjugate for the subspace parameter U.

The adaptation steps are:

1. **Fit the broad semantic subspace.** Compute the empirical covariance of the broad anchors and let U_{broad} be its top-k eigenvectors:

$$\widehat{\text{Cov}}(\Gamma_{\text{broad}}) = \frac{1}{m}\sum_{i=1}^{m} (\gamma_i^{\text{broad}} - \bar{\gamma}_{\text{broad}})(\gamma_i^{\text{broad}} - \bar{\gamma}_{\text{broad}})^\top.$$

2. **Initialize the Bingham prior.** Use the broad subspace to define
$$A = \alpha\, U_{\text{broad}} U_{\text{broad}}^\top,$$
where $\alpha > 0$ controls prior strength.
3. **Update with precise anchors.** Let
$$\widehat{\text{Cov}}(\Gamma_{\text{target}}) = \frac{1}{n}\sum_{i=1}^{n}(\gamma_i^{\text{target}} - \bar{\gamma}_{\text{target}})(\gamma_i^{\text{target}} - \bar{\gamma}_{\text{target}})^\top.$$
Under the Gaussian covariance model, the posterior over U remains in matrix-Bingham form, with updated parameter matrix
$$A_{\text{post}} = A + \beta\, \widehat{\text{Cov}}(\Gamma_{\text{target}}),$$
where
$$\beta = \frac{n}{2}\left(\frac{1}{\varepsilon} - \frac{1}{1+\varepsilon}\right).$$
4. **Form the adapted model via MAP.** We use the MAP estimator under the Bingham posterior:
$$U_{\text{post}} = \arg\max_{U \in \text{St}(d,k)} \text{tr}(U^\top A_{\text{post}} U),$$
which is given by the top-k eigenvectors of A_{post}.
5. **Induce the adapted geometry.** The adapted geometry is
$$d_{\text{LG}}^{\text{adapt}}(\lambda_1 \| \lambda_2) = \frac{1}{2\tau^2}(\lambda_1 - \lambda_2)^\top U_{\text{post}} U_{\text{post}}^\top (\lambda_1 - \lambda_2).$$

D.4 Latent spherical model with posterior-like Bingham update

For the latent spherical model, adaptation again reduces to updating the semantic subspace U. We use the same matrix-Bingham form as a tractable posterior-like update, driven by the broad-anchor subspace and the precise-anchor covariance. As in the latent Gaussian case, we initialize
$$A = \alpha\, U_{\text{broad}} U_{\text{broad}}^\top$$
from the broad-anchor subspace U_{broad}, then combine this with the precise-anchor covariance to form
$$A_{\text{post}} = A + \beta\, \widehat{\text{Cov}}(\Gamma_{\text{target}}),$$
where $\beta > 0$ controls the influence of the precise anchors. The adapted subspace is the MAP estimate
$$U_{\text{post}} = \arg\max_{U \in \text{St}(d,k)} \text{tr}(U^\top A_{\text{post}} U),$$
given by the top-k eigenvectors of A_{post}. The adapted geometry is obtained by substituting U_{post} into the latent spherical divergence.

D.5 Gaussian mixture model with Dirichlet–NIW update

For the Gaussian mixture model, we parameterize the anchor distribution as
$$q_\theta(d\gamma) = \sum_{m=1}^{M} \pi_m\, \mathcal{N}(d\gamma; \mu_m, \Sigma_m), \qquad \theta = \{(\pi_m, \mu_m, \Sigma_m)\}_{m=1}^{M},$$
where $\pi_m > 0$, $\sum_{m=1}^{M} \pi_m = 1$, and $\Sigma_m \succ 0$. The induced geometry is the Bregman divergence generated by
$$\kappa_{\text{GMM}}(\xi) = \log \sum_{m=1}^{M} \pi_m \exp\left(\mu_m^\top \xi + \frac{1}{2}\xi^\top \Sigma_m \xi\right).$$

Unlike the single-Gaussian case, this geometry is generally nonquadratic and can represent multiple semantic modes.

We place a symmetric Dirichlet prior on the mixture weights and independent Normal–Inverse-Wishart priors on the component parameters:

$$\begin{aligned}(\pi_1, \dots, \pi_M) &\sim \mathrm{Dir}(\alpha, \dots, \alpha),\\ \Sigma_m &\sim \mathcal{IW}(\Psi_{0,m}, \nu_{0,m}),\\ \mu_m \mid \Sigma_m &\sim \mathcal{N}\left(\mu_{0,m}, \frac{1}{\kappa_{0,m}}\Sigma_m\right), \qquad m = 1, \dots, M.\end{aligned}$$

The adaptation steps are:

1. **Fit the broad mixture.** Fit an M-component GMM to the broad anchors, producing
$$\left\{\pi_m^{\mathrm{broad}}, \mu_m^{\mathrm{broad}}, \Sigma_m^{\mathrm{broad}}\right\}_{m=1}^{M}.$$
2. **Initialize the prior.** Use the broad-anchor fit to set the component-specific hyperparameters
$$\begin{aligned}\mu_{0,m} &= \mu_m^{\mathrm{broad}},\\ \Psi_{0,m} &= (\nu_{0,m} - d - 1)\Sigma_m^{\mathrm{broad}},\end{aligned}$$
while resetting the mixture weights to a symmetric prior
$$(\pi_1, \dots, \pi_M) \sim \mathrm{Dir}(\alpha, \dots, \alpha).$$
Thus, the broad anchors determine the semantic modes and local covariance scales, but not their prior prevalence.
3. **Update with precise anchors using soft assignments.** Given precise anchors Γ_{target}, compute responsibilities
$$r_{im} = p(m \mid \gamma_i^{\mathrm{target}})$$
under the current mixture model. Soft assignments allow precise anchors to refine multiple broad semantic modes simultaneously. Define weighted sufficient statistics
$$\begin{aligned}N_m &= \sum_{i=1}^{n} r_{im},\\ \bar{\gamma}_m &= \frac{1}{N_m}\sum_{i=1}^{n} r_{im}\gamma_i^{\mathrm{target}},\\ S_m &= \sum_{i=1}^{n} r_{im}(\gamma_i^{\mathrm{target}} - \bar{\gamma}_m)(\gamma_i^{\mathrm{target}} - \bar{\gamma}_m)^\top.\end{aligned}$$
The updated hyperparameters are
$$\begin{aligned}\alpha'_m &= \alpha + N_m,\\ \kappa'_m &= \kappa_{0,m} + N_m,\\ \nu'_m &= \nu_{0,m} + N_m,\\ \mu'_m &= \frac{\kappa_{0,m}\mu_{0,m} + N_m\bar{\gamma}_m}{\kappa_{0,m} + N_m},\\ \Psi'_m &= \Psi_{0,m} + S_m + \frac{\kappa_{0,m}N_m}{\kappa_{0,m} + N_m}(\bar{\gamma}_m - \mu_{0,m})(\bar{\gamma}_m - \mu_{0,m})^\top.\end{aligned}$$
4. **Form the adapted model.** Use posterior means to define the adapted mixture parameters
$$\begin{aligned}\hat{\pi}_m &= \frac{\alpha'_m}{\sum_{\ell=1}^{M}\alpha'_\ell},\\ \hat{\mu}_m &= \mu'_m,\\ \hat{\Sigma}_m &= \frac{\Psi'_m}{\nu'_m - d - 1},\end{aligned}$$

provided $\nu'_m > d + 1$. This gives the adapted anchor model

$$q_{\hat{\theta}}(d\gamma) = \sum_{m=1}^{M} \hat{\pi}_m \mathcal{N}(d\gamma; \hat{\mu}_m, \hat{\Sigma}_m).$$

5. **Induce the adapted geometry.** The adapted geometry is

$$d^{\text{adapt}}_{\text{GMM}}(\lambda_1 \| \lambda_2) = B_{\kappa_{\hat{\theta}}}\left(\frac{\lambda_2}{\tau} \middle\| \frac{\lambda_1}{\tau}\right),$$

where

$$\kappa_{\hat{\theta}}(\xi) = \log \sum_{m=1}^{M} \hat{\pi}_m \exp\left(\hat{\mu}_m^\top \xi + \frac{1}{2}\xi^\top \hat{\Sigma}_m \xi\right).$$

D.6 Discussion

These model-specific updates instantiate the same principle: posterior anchor modeling updates a parametric anchor model using broad anchors as prior semantic structure and precise anchors as task-specific evidence, which in turn updates the induced Bregman geometry in image space. The models differ in the semantic structure they can transfer: a single global covariance in the Gaussian case, a shared low-dimensional semantic subspace in the low-rank case, and multiple reweighted semantic modes in the mixture case. This is precisely why richer models such as the GMM can support forms of semantic transfer that would not arise from simple projection heuristics.

E Implementation details

E.1 Datasets

We evaluate on ImageNet, ImageNet-R, CUB-200, Sketchy, and CheXpert.

ImageNet. We use the standard training split as the retrieval gallery/keys and the validation split as queries. Relevance is defined by class agreement. ImageNet class-name anchors are formed from the class labels using the template "a photo of a {x}".

ImageNet-R. Queries and keys are both drawn from ImageNet-R, with each query removed from its own key set. Object relevance is defined by ImageNet class agreement, while style labels for the style-control experiments are parsed from the ImageNet-R filenames. Object-name anchors use the template "a photo of a {x}", and style anchors use "a {s} rendition of an object". For style retrieval we discard images that have "misc" style labels, since these are not well-defined styles and do not have corresponding anchors.

CUB-200. We use the standard train/test split, with training images as the keys and test images as queries. Relevance is defined by bird-class agreement. CUB anchors are formed from the class names using the same class-name template.

Sketchy. We use the Sketchy Extended Split 2 protocol for zero-shot sketch-based image retrieval, following Yelamarthi et al. (2018). The test set consists of 21 classes chosen to be absent from ImageNet-1K. We use sketches from these unseen classes as queries and photos from the same classes as the retrieval gallery; no training-class images are used. Relevance is defined by class agreement, and anchors are formed from the class names using the template "a photo of a {x}".

CheXpert. We evaluate multi-label retrieval on CheXpert using pathology labels as relevance targets. The official training split is used as keys and the validation split as queries. We discard the training images with uncertain labels. For each target pathology, we construct report-style text anchors describing that pathology. The exact LLM prompt is given in Appendix E.6.

Synthetic collection for Figure 1. Figure 1 uses a synthetic image collection in which dog breed, background, and clothing are systematically varied. Images were generated using OpenAI's `gpt-image-2` model. The displayed neighbors are obtained directly by retrieval under cosine similarity and the corresponding anchor divergences.

E.2 Models

All natural-image experiments use MetaCLIP-2. All image and text embeddings are normalized to unit ℓ_2 norm, and we use the model-provided logit scale, $\exp(\texttt{clip_model.logit_scale})$, as the inverse temperature.

For CheXpert, we use CheXzero and report results averaged over the 10 released model weights.

E.3 Image-text inversion baseline

As a retrieval baseline, we use image-text inversion (ITI), which converts a query image into a text-side proxy by selecting the anchor with highest image–text compatibility. Given an anchor set $\Gamma = \{\gamma_1, \dots, \gamma_N\}$ and a query image embedding λ_q, we first choose

$$i^\star(\lambda_q) = \arg\max_{1 \le i \le N} \lambda_q^\top \gamma_i. \tag{67}$$

We then take the corresponding anchor

$$\tilde{\gamma}(\lambda_q) = \gamma_{i^\star(\lambda_q)} \tag{68}$$

as the inverted text representation of the query. Each gallery image embedding λ_k is then scored by cosine similarity to this inferred text embedding:

$$s_{\mathrm{ITI}}(\lambda_q, \lambda_k) = \tilde{\gamma}(\lambda_q)^\top \lambda_k. \tag{69}$$

Keys are ranked in decreasing order of $s_{\mathrm{ITI}}(\lambda_q, \lambda_k)$.

Thus, ITI uses the anchor set only to select a single best-matching text proxy for the query, after which retrieval is performed by standard image-text cosine similarity.

E.4 Evaluation metric

We evaluate retrieval performance with the area under the precision–recall curve (AUPRC). For a given query, let

$$\mathcal{G} = \{x_1, \dots, x_N\}$$

denote the key set, let s_i be the retrieval score assigned to key item x_i, and let

$$y_i \in \{0, 1\}$$

indicate whether x_i is semantically relevant to the query. Sorting the keys by score induces a ranked list, and for any threshold t on the score we define

$$\mathrm{Precision}(t) = \frac{\sum_{i=1}^N \mathbf{1}\{s_i \ge t\} y_i}{\sum_{i=1}^N \mathbf{1}\{s_i \ge t\}}, \tag{70}$$

$$\mathrm{Recall}(t) = \frac{\sum_{i=1}^N \mathbf{1}\{s_i \ge t\} y_i}{\sum_{i=1}^N y_i}. \tag{71}$$

The precision–recall curve is the set of points

$$\big(\mathrm{Recall}(t), \mathrm{Precision}(t)\big)$$

obtained as the threshold varies over the ranked list, and AUPRC is the area under this curve:

$$\mathrm{AUPRC} = \int_0^1 \mathrm{Precision}(r)\, dr, \tag{72}$$

with the discrete implementation given by the corresponding finite-sum approximation over ranked thresholds.

E.5 Parametric models and adaptation

Gaussian model. We fit the Gaussian anchor model by maximum likelihood on the anchor embeddings.

Low-rank models. For the latent Gaussian and latent spherical models, we estimate the semantic subspace U by the top-k eigenvectors of the centered anchor covariance matrix. Unless otherwise specified, we choose k to explain 95% of the anchor covariance. In the semantic adaptation experiments, where a fixed latent dimension is required, we use $k = 128$. For the ambient surrogate used in latent Gaussian adaptation, we set the isotropic regularization to $\varepsilon = 10^{-8}$.

Gaussian mixture model. We fit the Gaussian mixture model by expectation–maximization, with the number of mixture components specified in the corresponding plots. EM is initialized with k-means, each covariance matrix is regularized by adding $10^{-5}I$, and optimization is terminated when the improvement in log-likelihood falls below 10^{-3} or when 200 iterations are reached.

Adaptation. Gaussian adaptation uses the Normal–Inverse-Wishart update in Appendix D.2. Latent Gaussian adaptation uses the matrix-Bingham update in Appendix D.3. Latent spherical adaptation uses the posterior-like Bingham update in Appendix D.4. Gaussian mixture adaptation uses the Dirichlet–NIW update in Appendix D.5.

Adaptation hyperparameters. For Gaussian adaptation, we use $\kappa_0 = 5$ and $\nu_0 = d + 20$. The low-rank Bingham updates have prior strength $\alpha = 50$ and $\varepsilon = 10^{-8}$. For GMM adaptation, we use Dirichlet concentration $\alpha = 1$, $\kappa_{0,m} = 5$, and $\nu_{0,m} = d + 20$, shared across mixture components unless otherwise specified.

E.6 CheXpert anchor generation

For CheXpert, we generate pathology-specific report-style anchors with ChatGPT 5.4 Thinking. The prompt asks for a Python dictionary whose keys are the CheXpert conditions and whose values are lists of possible doctor report notes for the corresponding condition. Each list contains both positive and negative entries and is sized to contain roughly 50 strings. The resulting pathology-specific anchor lists are then embedded with CheXzero and used as the anchor sets defining the corresponding retrieval geometries.

Prompt.

> For the conditions in LABELS = [“Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Opacity”, “Lung Lesion”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”] Give a python dictionary where the keys are the different conditions and the values are a list of strings of different possible doctor report notes for that condition. Have positive and negative entries. Have each list be of size around 50

Usage in retrieval. For each target pathology, we extract the corresponding list from the generated dictionary, embed those report-style strings with CheXzero, and use the resulting text embeddings as the pathology-specific anchor set for retrieval.

F Additional Experiments

F.1 Parametric anchor models on CheXpert

The main text reports CheXpert results for the pathology-specialized softmax anchor divergence. Here we include the corresponding parametric anchor models on the same retrieval task. Table 3 shows that Gaussian anchor divergence performs comparably to softmax on average, while the latent models also improve over cosine, indicating that parametric anchor geometries can preserve clinically relevant coordinates beyond the natural-image setting.

Table 3: **Parametric CheXpert anchor divergences.** Entries are pathology-wise mean AUPRC differences relative to cosine similarity, with 95% confidence interval. Results average over independently generated anchor sets and CheXzero checkpoints, with intervals accounting for bootstrap and anchor-generation variability. Softmax and Gaussian anchor divergences give the largest average gains, while latent models also improve over cosine on average.

Pathology	ITI − Cosine	Softmax − Cosine	Gaussian − Cosine	Latent G. − Cosine	Latent S. − Cosine
Enlarged Cardiomediastinum	0.0042 ± 0.0009	-0.0002 ± 0.0003	-0.0002 ± 0.0003	0.0000 ± 0.0002	-0.0004 ± 0.0003
Cardiomegaly	$\mathbf{0.0211 \pm 0.0083}$	$\mathbf{0.0279 \pm 0.0034}$	$\mathbf{0.0327 \pm 0.0030}$	$\mathbf{0.0152 \pm 0.0015}$	$\mathbf{0.0194 \pm 0.0020}$
Lung Opacity	-0.0047 ± 0.0050	$\mathbf{0.0090 \pm 0.0032}$	$\mathbf{0.0092 \pm 0.0031}$	0.0037 ± 0.0016	$\mathbf{0.0083 \pm 0.0019}$
Lung Lesion	-0.0050 ± 0.0011	0.0001 ± 0.0005	-0.0000 ± 0.0004	0.0003 ± 0.0003	0.0013 ± 0.0004
Edema	0.0030 ± 0.0064	$\mathbf{0.0316 \pm 0.0044}$	$\mathbf{0.0309 \pm 0.0045}$	$\mathbf{0.0127 \pm 0.0024}$	$\mathbf{0.0194 \pm 0.0027}$
Consolidation	-0.0011 ± 0.0020	$\mathbf{0.0062 \pm 0.0014}$	$\mathbf{0.0056 \pm 0.0012}$	0.0026 ± 0.0006	0.0020 ± 0.0007
Pneumonia	-0.0037 ± 0.0017	$\mathbf{0.0081 \pm 0.0013}$	$\mathbf{0.0072 \pm 0.0010}$	0.0038 ± 0.0005	0.0035 ± 0.0007
Atelectasis	-0.0018 ± 0.0030	$\mathbf{0.0095 \pm 0.0027}$	$\mathbf{0.0101 \pm 0.0025}$	0.0029 ± 0.0012	$\mathbf{0.0058 \pm 0.0017}$
Pneumothorax	-0.0048 ± 0.0020	$\mathbf{0.0080 \pm 0.0013}$	$\mathbf{0.0090 \pm 0.0013}$	0.0018 ± 0.0005	0.0005 ± 0.0005
Pleural Effusion	$\mathbf{0.0137 \pm 0.0083}$	$\mathbf{0.0465 \pm 0.0060}$	$\mathbf{0.0441 \pm 0.0055}$	$\mathbf{0.0194 \pm 0.0030}$	$\mathbf{0.0228 \pm 0.0030}$
Pleural Other	-0.0034 ± 0.0006	-0.0009 ± 0.0003	-0.0009 ± 0.0003	-0.0004 ± 0.0002	0.0006 ± 0.0003
Fracture	-0.0027 ± 0.0005	0.0003 ± 0.0002	0.0003 ± 0.0002	0.0003 ± 0.0001	0.0016 ± 0.0003
Support Devices	-0.0158 ± 0.0079	$\mathbf{0.0208 \pm 0.0056}$	$\mathbf{0.0277 \pm 0.0050}$	$\mathbf{0.0126 \pm 0.0026}$	$\mathbf{0.0183 \pm 0.0030}$
Average	-0.0001 ± 0.0012	$\mathbf{0.0128 \pm 0.0011}$	$\mathbf{0.0135 \pm 0.0008}$	$\mathbf{0.0058 \pm 0.0005}$	$\mathbf{0.0079 \pm 0.0006}$

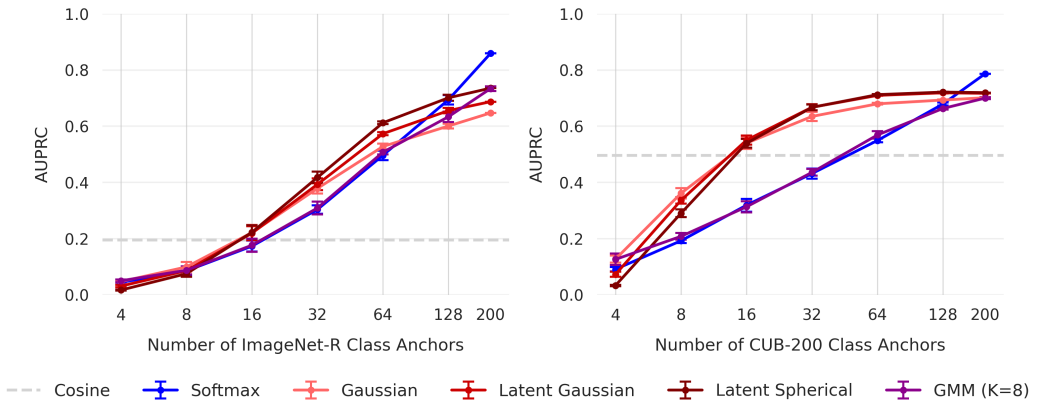

Figure 6: **Parametric structure is useful when anchor coverage is limited.** AUPRC as the number of class-name anchors varies on **ImageNet-R** (left) and **CUB-200** (right). Parametric anchor divergences are strongest across much of the limited- and moderate-anchor regime, while softmax improves with anchor coverage and is strongest when the full class-anchor set is available.

F.2 Parametric models across anchor budgets on additional datasets

The main text studies how empirical and parametric anchor models behave as the number of available anchors varies on ImageNet. We observe the same qualitative tradeoff on ImageNet-R and CUB-200. As shown in Figure 6, parametric anchor divergences are strongest over much of the limited- and moderate-anchor regime, while softmax improves steadily with anchor coverage and becomes strongest when the full class-anchor set is available. Thus, the anchor-coverage tradeoff observed on ImageNet is not specific to that dataset.

F.3 Exploratory model-complexity diagnostic

We include an exploratory diagnostic on how mixture complexity interacts with anchor-distribution structure. In the anchor-distribution view, parametric models are structured approximations to q: increasing flexibility can reduce approximation bias when the anchor distribution is heterogeneous, but can also introduce unnecessary complexity or estimation noise when simpler structure is sufficient.

We compare ImageNet and CUB-200 as illustrative regimes. ImageNet class-name anchors span many coarse semantic groups, while CUB-200 anchors are narrower and concentrated around fine-grained bird categories. Figure 7 shows different trends across these datasets: increasing the number

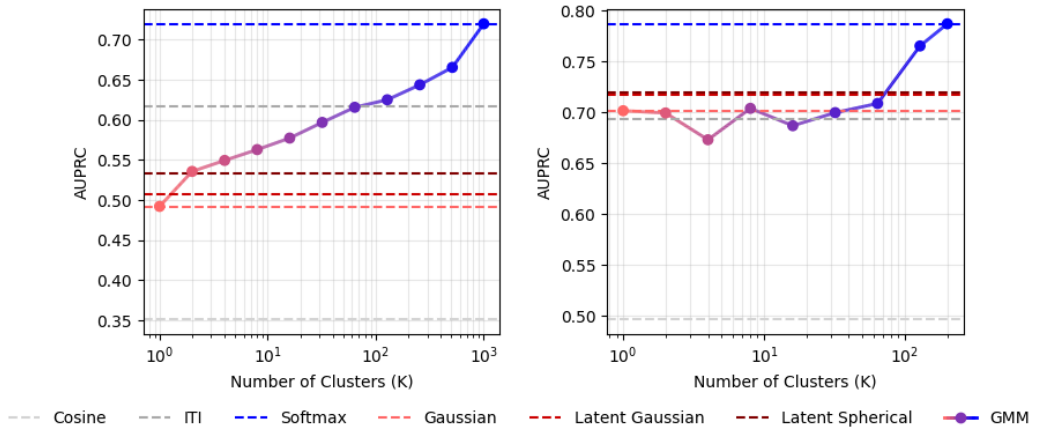

Figure 7: **Mixture complexity behaves differently across anchor sets.** On ImageNet, increasing the number of GMM components generally improves retrieval, consistent with a heterogeneous class-name anchor distribution. On CUB-200, gains are smaller and less consistent, suggesting that simpler structured geometries may already capture much of the relevant anchor variation.

Table 4: **Runtime scaling with anchor-set size.** Wall-clock retrieval time (in seconds) on a single NVIDIA H100 NVL GPU for 10,000 queries and 50,000 gallery images. Runtimes include fitting, preprocessing, and retrieval and are averaged over 10 runs following 3 warm-up runs.

Method	$\mathbf{2^{10}}$	$\mathbf{2^{12}}$	$\mathbf{2^{14}}$	$\mathbf{2^{16}}$	$\mathbf{2^{17}}$
Cosine	0.152	0.152	0.152	0.151	0.152
Softmax	0.216	0.281	0.538	1.622	3.066
Gaussian	0.364	0.363	0.367	0.370	0.375

of mixture components tends to help on ImageNet, while gains are smaller and less consistent on CUB-200. This diagnostic is not intended as a complete model-selection study, but illustrates that parametric anchor models introduce structure whose usefulness depends on the structure of the anchor set.

F.4 Computational scaling with anchor-set size

Parametric anchor models also provide a computational advantage when the number of observed anchors is large. Let Q denote the number of queries, G the number of gallery items, N the number of anchors, and d the embedding dimension. A direct implementation of softmax anchor divergence constructs an N-dimensional distribution for each image, giving computational cost $O(QGN + (Q+G)Nd)$.

For the symmetrized divergence used in our experiments, the subsequent query–gallery comparison can be computed exactly from cached d-dimensional statistics. Writing $\eta = \lambda/\tau$ and $\mu(\lambda) = \nabla\kappa_{\hat{q}_\Gamma}(\eta)$, we have

$$d_{\mathrm{sym}}(\lambda_q, \lambda_g) = \frac{1}{2}(\eta_q - \eta_g)^\top \big(\mu(\lambda_q) - \mu(\lambda_g)\big). \tag{73}$$

Thus, after caching $\mu(\lambda)$, softmax anchor divergence has total cost

$$O((Q+G)Nd + QGd),$$

so dependence on N remains in the gallery and query dependent preprocessing step. In contrast, a fitted parametric anchor model summarizes the observed anchors by a fixed set of parameters. Once this fitting step is complete, query and gallery preprocessing and pairwise evaluation need not depend on the original anchor-set size.

We examine this empirically by varying the number of anchors and timing retrieval for 10,000 queries against 50,000 gallery image embeddings on a single NVIDIA H100 NVL GPU. Runtimes include model fitting, and retrieval, and are averaged over 10 runs following 3 warm-up runs. For softmax anchor divergence, anchors are streamed in chunks when computing the cached statistics, so additional peak GPU memory does not grow substantially with N. Table 4 shows that softmax runtime increases with anchor-set size as anchor-dependent preprocessing becomes appreciable, whereas Gaussian anchor divergence remains nearly constant after the anchor distribution has been fit.

F.5 Qualitative object–style retrieval examples

Figure 8 shows additional object- and style-anchor retrieval examples. Object anchors retrieve images with the same object across styles, while style anchors retrieve images with the same style across object categories. This pattern appears for both empirical softmax and parametric anchor divergences, suggesting that parametric models preserve the semantic coordinate selected by the anchors.

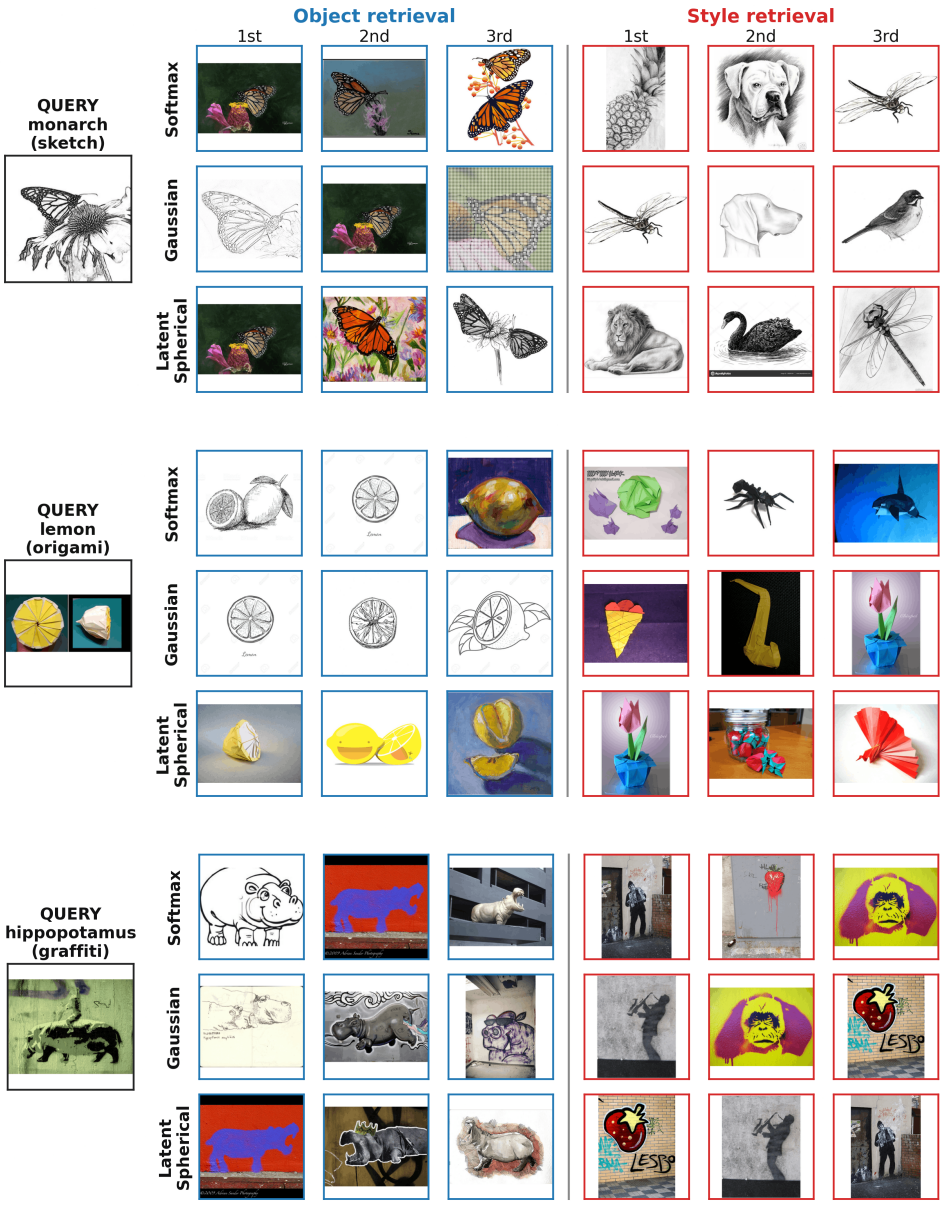

Figure 8: **Qualitative examples of anchor-controlled retrieval.** Blue boxes show object-anchor retrieval, which prioritizes object identity across styles; red boxes show style-anchor retrieval, which prioritizes style across object categories. Rows compare some of the different anchor divergences.