Alleviating Performance Degradation Caused by Out-of-Distribution Issues in Embedding-Based Retrieval

Haotong Bao*, **Jianjin Zhang**, **Qi Chen**†, **Weihao Han**, **Zhengxin Zeng**, **Ruiheng Chang**, **Mingzheng Li**, **Hao Sun**, **Weiwei Deng**, **Feng Sun**, **Qi Zhang**
Microsoft, China
baosfw@gmail.com
{jianjzh, cheqi, weihan, zhze, ruihengchang, mingzhengli, hasun, dedeng, sunfeng, qizhang}@microsoft.com

Abstract

In Embedding Based Retrieval (EBR), Approximate Nearest Neighbor (ANN) algorithms are widely adopted for efficient large-scale search. However, recent studies reveal a query out-of-distribution (OOD) issue, where query and base embeddings follow mismatched distributions, significantly degrading ANN performance. In this work, we empirically verify the generality of this phenomenon and provide a quantitative analysis. To mitigate the distributional gap, we introduce a distribution regularizer into the encoder training objective, encouraging alignment between query and base embeddings. Extensive experiments across multiple datasets, encoders, and ANN indices show that our method consistently improves retrieval performance.

1 Introduction

Embedding-Based Retrieval (EBR) leverages deep encoders, especially Pre-trained Language Models (PLMs) (Devlin et al., 2018; Liu et al., 2019), to convert text into high-dimensional embeddings (Huang et al., 2020), enabling fast similarity search via metrics like inner product or Euclidean distance. EBR has been widely used in web search (Mitra et al., 2017; Zhang et al., 2024), recommendation (Zhang et al., 2023), QA (Karpukhin et al., 2020; Sachan et al., 2023), and dialogue systems (Lewis et al., 2020). To support large-scale scenarios with billions of items, modern retrieval systems combine EBR with Approximate Nearest Neighbor (ANN) algorithms (Johnson et al., 2019; Shrivastava and Li, 2014), achieving sub-linear search time with acceptable accuracy.

Training objectives in EBR typically aim to enhance retrieval accuracy by separating positive and negative examples in the embedding space (Zhao et al., 2024). Contrastive learning with bi-encoder models (Hadsell et al., 2006) is widely adopted: it pulls positive pairs closer and pushes negatives apart, leading to more discriminative embeddings. This paradigm has proven effective across many PLM-based retrievers (Karpukhin et al., 2020; Qu et al., 2020; Gao and Callan, 2021). The trained encoder supports efficient online search by embedding queries for Approximate Nearest Neighbor (ANN) based retrieval over a precomputed candidate index (Huang et al., 2020).

While contrastive training improves discriminative power between positives and negatives, it also amplifies the distributional gap between query and base embeddings. This issue, first identified in multi-modal training (Liang et al., 2022) and later observed in text-only settings (Chen et al., 2024b), causes query embeddings to become out-of-distribution (OOD) relative to the base data. Such distributional mismatch violates the assumption—underlying most ANN algorithms—that query and base embeddings follow the same distribution, leading to degraded retrieval accuracy (Chen et al., 2024a).

In practice, retrieval encoders are often trained without considering ANN-specific constraints, under the assumption that embeddings optimized for KNN will generalize to ANN. However, this overlooks the impact of OOD queries, resulting in substantial yet under-recognized performance drops.

Recent methods like OOD-DiskANN (Jaiswal et al., 2022) and RoarGraph (Chen et al., 2024a) improve ANN robustness by incorporating query information during index construction. Though effective, these methods focus on the index side and leave the root cause—OOD query embeddings—largely unaddressed.

This work empirically shows that query embeddings in retrieval tasks are often OOD relative to base embeddings, degrading ANN performance. We employ Maximum Mean Discrepancy (MMD) (Gretton et al., 2006) to quantify the discrepancy,

*Work done while at Microsoft.
†Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6418–6427
November 4-9, 2025

and derive a training-time regularizer to reduce the discrepancy and enhance retrieval robustness.

Unlike prior work (Jaiswal et al., 2022; Chen et al., 2024a) that improves ANN indices to tolerate OOD queries, we aim to mitigate the root cause by reducing the OOD effect at the embedding level during encoder training. To our knowledge, this is among the first works to explicitly target this overlooked distribution gap in an end-to-end text-to-text retrieval setting, where query OOD has critical impact on ANN accuracy.

Contributions. We summarize our main contributions as:

- We reveal and quantify the distribution gap between query and base embeddings, showing query OOD is common and harms ANN search.
- We propose a simple training-time distribution regularizer to reduce this gap without adding inference overhead.
- Extensive experiments on diverse encoders, datasets, and ANN methods verify consistent retrieval improvements.

2 OOD in ANN Search

Retrieval Task Setting. In embedding-based retrieval, query q and base data y are encoded by E_q and E_y into embeddings $\mathbf{q} = E_q(q)$ and $\mathbf{y} = E_y(y)$, with distributions $\mathcal{P}_q$ and $\mathcal{P}_y$. Similarity $\text{sim}(\cdot,\cdot)$, usually inner product, measures their closeness.

The training of the encoders involves triplets $\{q, y^+, Y^-\}$ with positive y^+ and negatives Y^-, optimized by InfoNCE loss (Oord et al., 2018):

$$\mathcal{L}_{\text{con}} = -\log \frac{e^{\mathbf{q}^\top \mathbf{y}^+/\tau}}{e^{\mathbf{q}^\top \mathbf{y}^+/\tau} + \sum_{\mathbf{y}\in\mathcal{N}} e^{\mathbf{q}^\top \mathbf{y}/\tau}}, \quad (1)$$

where τ is temperature. After training, base embeddings $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ are indexed with ANN for efficient search given $\mathbf{q}$.

The Out-Of-Distribution (OOD) Issue. Our definition of OOD follows (Jaiswal et al., 2022; Chen et al., 2024a), where query embeddings $\mathbf{q}$ are OOD relative to base embeddings $\mathbf{y}$ if their distributions differ substantially. This differs from treating individual outliers as OOD; here, OOD describes the entire query embedding distribution $\mathcal{P}_q$ lies outside the base embedding distribution $\mathcal{P}_y$.

To quantify this discrepancy, we use the squared Maximum Mean Discrepancy (MMD) (Gretton et al., 2006) between the query and base embedding distributions, denoted as:

$$\text{MMD}^2(\mathcal{P}_q, \mathcal{P}_y) = \left\| \mathbb{E}_{\mathcal{P}_q}[\mathbf{q}] - \mathbb{E}_{\mathcal{P}_y}[\mathbf{y}] \right\|^2, \quad (2)$$

Based on the base embedding distribution, we compute single-point MMD scores for each query embedding. As shown in Figure 1, in-distribution (ID) embeddings (sampled from the base set) yield lower scores, while OOD embeddings exhibit significantly higher values, confirming the presence of OOD in text-to-text retrieval consistent with prior ANN observations (Jaiswal et al., 2022).

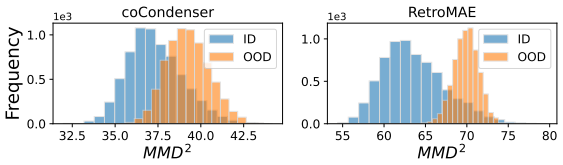


Figure 1: MMD^2 between ID/OOD query and base embeddings.

Challenges of Serving OOD Queries in ANN Search. OOD queries are challenging to serve in ANN searches because they break the fundamental assumption that queries and base data share the same distribution, leading to significantly larger search spaces (Chen et al., 2024a). This results in inefficient search convergence, requiring more computations, memory access, and extended search paths.

To demonstrate this issue, we evaluated two ANN algorithms, IVF and HNSW, on the MS-MARCO dataset (Nguyen et al., 2017) using embeddings from coCondenser (Gao and Callan, 2021) and RetroMAE (Xiao et al., 2022). OOD queries were taken from MSMARCO dev set, while the ANN indices were built on the base embeddings. We compared the search complexity—measured by candidates accessed—for ID and OOD queries at equal recall. As shown in Figure 2, OOD queries require significantly higher complexity, resulting in longer search times and making it difficult to achieve satisfactory recall efficiently. This suggests ANN indexes constructed on base distributions poorly generalize to OOD queries.

Contrastive Learning Amplifies Distribution Gap. We empirically observe that contrastive learning with InfoNCE increases the distance between query and base embeddings. During fine-tuning on MSMARCO with coCondenser and RetroMAE, we monitor the average ℓ_2 distance

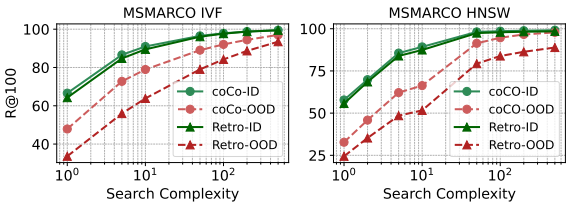


Figure 2: Search efficiency for ID and OOD queries. "coCO" denotes coCondenser, "Retro" denotes RetroMAE.

between query–positive pairs. Since the squared ℓ_2 distance can serve as a proxy measure for MMD with a linear kernel (proof in Appendix A), it enables us to track distributional shift more conveniently. As shown in Figure 3, the ℓ_2 distance steadily increases during training, indicating that InfoNCE implicitly enlarges the query–base distribution gap.

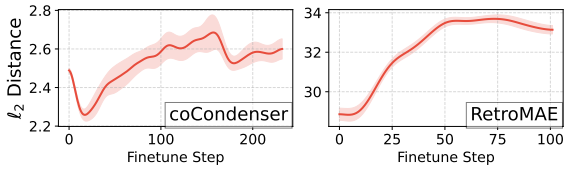


Figure 3: Increasing ℓ_2 distance in contrastive learning.

3 Solution to the OOD Issue

Section 2 highlights the OOD issue in retrieval and its harm to ANN search, with the squared ℓ_2 distance showing a growing trend as a proxy for distributional drift. In this section, we propose a method to mitigate OOD by explicitly regularizing the drift via instance-level decomposition of MMD.

For a single query embedding $\mathbf{q} \sim \mathcal{P}_q$ and its corresponding document embeddings $\{\mathbf{y}^+ \cup \mathcal{Y}^-\} \sim \mathcal{P}_y$, we construct the regularization loss directly from the MMD decomposition:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{\mathbf{y} \in \{\mathbf{y}^+ \cup \mathcal{Y}^-\}} \|\mathbf{q} - \mathbf{y}\|_2^2 \qquad (3)$$

where N denotes the number of document samples. This formulation preserves the core property of MMD while avoiding computational overhead from covariance estimation.

The complete training objective combines contrastive learning with MMD-aware regularization:

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda \mathcal{L}_{\text{reg}}, \qquad (4)$$

where λ governs the trade-off between instance discrimination and distributional alignment. This principled design ensures that the learned embeddings simultaneously maximize semantic discriminability and minimize population deviation from the target distribution.

Unlike contrastive learning loss, which optimizes relative distances for discrimination, our regularization imposes a global constraint by uniformly reducing the distance between all query and base embeddings. This suppresses distributional drift—reflected by rising MMD—and preserves distributional alignment. Incorporating $\mathcal{L}_{\text{reg}}$ alone improves ANN search performance with negligible overhead and no added training or search complexity. Further discussion of the ℓ_2 regularization is provided in the Appendix B.

4 Experiments

In this section, we evaluate our method on two standard retrieval benchmarks, MSMARCO (Nguyen et al., 2017) and Natural Questions (Kwiatkowski et al., 2019), measuring recall versus queries per second (QPS) following (Aumüller et al., 2020). Experiments use two strong PLM-based encoders, *coCondenser* (Gao and Callan, 2021) and *RetroMAE* (Xiao et al., 2022). ANN search is conducted with five algorithms: IVF, IVFPQ, HNSW (via FAISS), DiskANN, and RoarGraph (Chen et al., 2024a). The distribution regularization weight λ is set to 0.01 to balance with contrastive loss, and training follows official fine-tuning setup. More implementation details are provided in Appendix E.

Main Results Figure 4 shows the impact of applying the distribution regularization when fine-tuning coCondenser and RetroMAE on MSMARCO. Similar trends are observed on Natural Questions in Figure 5, confirming that our method consistently improves ANN search performance across datasets, especially in the high-QPS, low-recall region where the impact of OOD is more significant.

Graph-based methods such as HNSW, DiskANN and RoarGraph benefit most from the regularization. Even RoarGraph—designed for OOD queries—sees substantial gains, showing that aligning embeddings addresses the root cause of query–base mismatch and complements index-level robustness. IVF and IVFPQ also show clear recall improvements despite heavy approximations. In the low-QPS, high-recall region, improvements

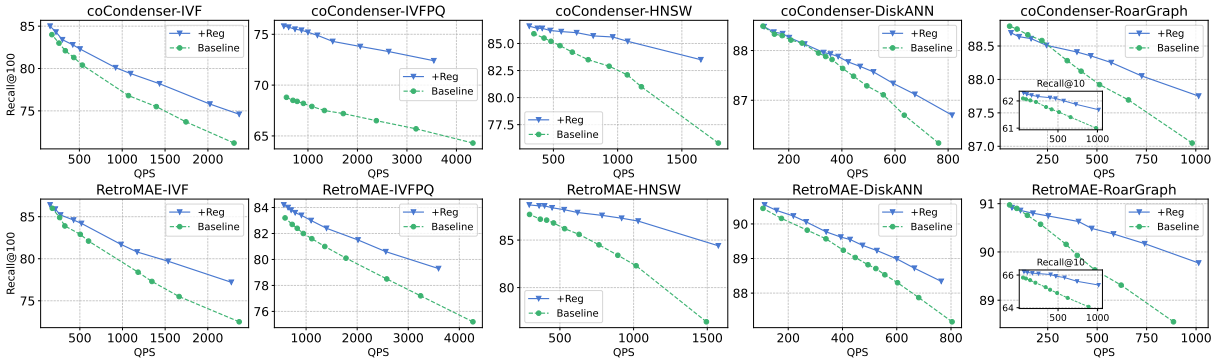


Figure 4: QPS-Recall curves of the distribution regularizer method on the MSMARCO dataset. Curves closer to the top right of the chart indicate better performance.

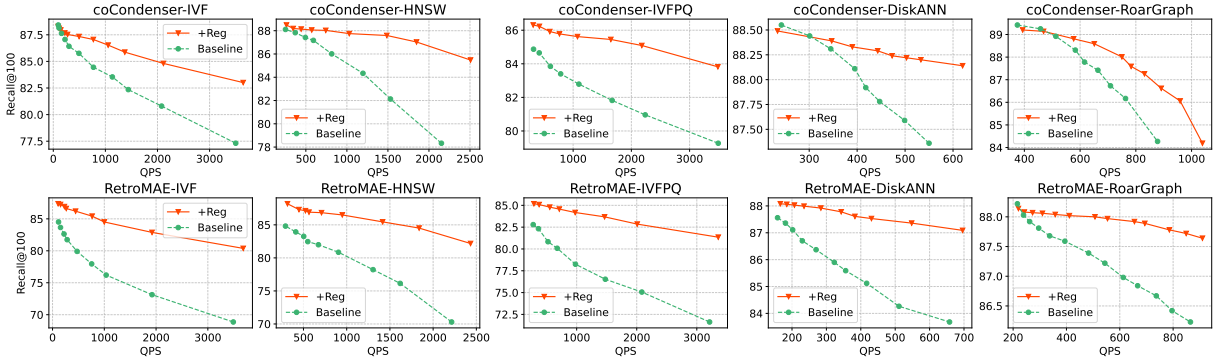
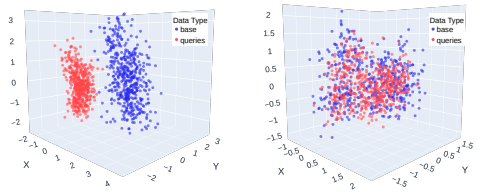


Figure 5: QPS-Recall curves of the distribution regularization method on the NQ dataset.

are less pronounced due to reduced ANN approximation, but RoarGraph still shows significant gains in Recall@10, demonstrating the regularization's advantage in tight-target scenarios. At the same recall level, the regularizer enables up to 2.6× and 4.1× speedups on coCondenser and RetroMAE, respectively, highlighting efficiency improvements under strict latency constraints.

Overall, the proposed method offers a simple yet effective enhancement to encoder training, improving recall and latency across datasets, models, and ANN methods.



(a) only contrastive loss (b) with regularization

Figure 6: Distributions of query and base embeddings visualized with t-SNE, comparing (a) only contrastive loss and (b) using distribution regularization.

Impact of Regularization Figure 6 illustrates the effect of the distribution regularization: without it, query and base embeddings form two separate clusters with a clear gap, consistent with (Chen et al., 2024b); with it, the distributions become aligned. Figure 7 further shows that the regularization reduces the MMD between query and base embeddings, as reflected by the similarity in their distribution centers and overall shapes (cf. Figure 1). This demonstrates that our method effectively mitigates OOD issues by reducing MMD, ultimately leading to improved ANN search recall.

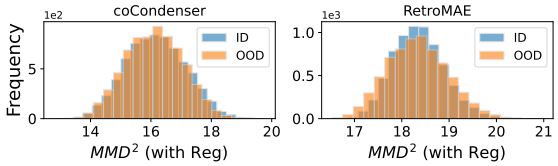


Figure 7: MMD^2 between ID/OOD query and base embeddings with regularization.

5 Conclusion

We identify and quantify the query OOD issue in embedding-based retrieval and show its adverse impact on ANN performance. To address this, we propose a simple ℓ_2-based regularization that narrows the query–base embedding gap with-

out adding inference-time overhead. Experiments across datasets, encoders, and ANN methods show consistent recall gains, especially in high-QPS regimes. As future work, we plan to include direct comparisons with recent embedding models such as GRITLM (Muennighoff et al., 2024) , and to conduct a more systematic analysis of training and inference efficiency.

Limitations

Our approach, while effective within the tested scope, still presents several avenues for future improvement:

Scope of distribution shift. We currently monitor drift only via the *first-order* linear-kernel MMD. Other forms of mismatch are not explicitly handled and could, in some situations, influence retrieval quality.

Single-modality evaluation. Empirical validation is limited to text-to-text retrieval. Extending the regularizer to cross-modal, multilingual, or speech/image settings remains an open question that warrants additional experimentation.

Tuning the trade-off parameter λ. The hyperparameter that balances discrimination ($\mathcal{L}_{\text{con}}$) and alignment ($\mathcal{L}_{\text{reg}}$) is selected on held-out data. While stable in our study, the optimal λ may vary across tasks and domains, suggesting the benefit of automated or adaptive tuning strategies.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Martin Aumüller and Matteo Ceccarello. 2023. Recent approaches and trends in approximate nearest neighbor search, with remarks on benchmarking. *IEEE Data Eng. Bull.*, 46:89–105.

Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2020. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87:101374.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Meng Chen, Kai Zhang, Zhenying He, Yinan Jing, and X. Sean Wang. 2024a. Roargraph: A projected bipartite graph for efficient cross-modal approximate nearest neighbor search. *Proceedings of the VLDB Endowment*, 17(11):2735–2749.

Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul Bennett, and 1 others. 2024b. Ms marco web search: a large-scale information-rich web dataset with millions of real click labels. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 292–301.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pages 3887–3896. PMLR.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Shikhar Jaiswal, Ravishankar Krishnaswamy, Ankit Garg, Harsha Vardhan Simhadri, and Sheshansh Agrawal. 2022. Ood-diskann: Efficient and scalable graph anns for out-of-distribution queries. *arXiv preprint arXiv:2211.12850*.

Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. 2019. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in Neural Information Processing Systems*, 32.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th international conference on world wide web*, pages 1291–1299.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2017. MS MARCO: A human-generated MAchine reading COmprehension dataset.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2023. Questions are all you need to train a dense passage retriever. *Transactions of the Association for Computational Linguistics*, 11:600–616.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). *Advances in neural information processing systems*, 27.

Harsha Vardhan Simhadri, Ravishankar Krishnaswamy, Gopal Srinivasa, Suhas Jayaram Subramanya, Andrija Antonijevic, Dax Pryce, David Kaczynski, Shane Williams, Siddarth Gollapudi, Varun Sivashankar, Neel Karia, Aditi Singh, Shikhar Jaiswal, Neelam Mahapatro, Philip Adams, Bryan Tower, and Yash Patel. 2023. DiskANN: Graph-structured Indices for Scalable, Fast, Fresh and Filtered Approximate Nearest Neighbor Search.

Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. 2023. Contextual masked auto-encoder for dense passage retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4738–4746.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. *arXiv preprint arXiv:2205.12035*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Yandex Research. 2021. Text-to-image-1b: Benchmarks for billion-scale similarity search. https://research.yandex.com/blog/benchmarks-for-billion-scalesimilarity-search.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Yanan Zhang, Xiaoling Bai, and Tianhua Zhou. 2024. Event-enhanced retrieval in real-time search. *arXiv preprint arXiv:2404.05989*.

Yuan Zhang, Xue Dong, Weijie Ding, Biao Li, Peng Jiang, and Kun Gai. 2023. Divide and conquer: Towards better embedding-based retrieval for recommender systems from a multi-task perspective. In *Companion Proceedings of the ACM Web Conference 2023*, pages 366–370.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Appendix

A Average ℓ_2 Distance *vs.* Linear–Kernel MMD

We prove in detail that the batch–average squared Euclidean distance is an affine surrogate of the linear–kernel maximum mean discrepancy (MMD).[1]

A.1 Linear–kernel MMD revisited

Write $\boldsymbol{\mu}_q = \mathbb{E}_{\mathcal{P}_q}[\mathbf{q}]$, $\boldsymbol{\mu}_y = \mathbb{E}_{\mathcal{P}_y}[\mathbf{y}]$. For the linear kernel $k(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}$ the squared MMD is[2]

$$\mathrm{MMD}^2(\mathcal{P}_q, \mathcal{P}_y) = \left\|\boldsymbol{\mu}_q - \boldsymbol{\mu}_y\right\|^2. \tag{5}$$

For brevity, we will refer to it as $\mathrm{MMD}^2_{\mathrm{lin}}$ in the following.

[1]Throughout we assume $\mathbf{q} \sim \mathcal{P}_q$ and $\mathbf{y} \sim \mathcal{P}_y$ are *independent* draws, which is the standard setting for MMD.

[2]Eq. (2) is copied from the main paper for completeness.

Derivation. Let $\mathbf{q}, \mathbf{q}' \overset{\text{i.i.d.}}{\sim} \mathcal{P}_q$ and $\mathbf{y}, \mathbf{y}' \overset{\text{i.i.d.}}{\sim} \mathcal{P}_y$, all mutually independent. Plugging the linear kernel into the population MMD formula yields:

$$\begin{aligned}\mathrm{MMD}^2_{\mathrm{lin}} &= \mathbb{E}_{\mathbf{q},\mathbf{q}'}[\mathbf{q}^\top \mathbf{q}'] + \mathbb{E}_{\mathbf{y},\mathbf{y}'}[\mathbf{y}^\top \mathbf{y}'] \\ &\quad - 2\mathbb{E}_{\mathbf{q},\mathbf{y}}[\mathbf{q}^\top \mathbf{y}] \\ &= \boldsymbol{\mu}_q^\top \boldsymbol{\mu}_q + \boldsymbol{\mu}_y^\top \boldsymbol{\mu}_y - 2\,\boldsymbol{\mu}_q^\top \boldsymbol{\mu}_y \\ &= \left\|\boldsymbol{\mu}_q - \boldsymbol{\mu}_y\right\|^2, \end{aligned} \tag{6}$$

where independence implies $\mathbb{E}[\mathbf{q}^\top \mathbf{q}'] = \mathbb{E}[\mathbf{q}]^\top \mathbb{E}[\mathbf{q}'] = \boldsymbol{\mu}_q^\top \boldsymbol{\mu}_q$, and likewise for the other expectations, no covariance terms survive.

A.2 ℓ_2 Distance Decomposition

Draw a single query–document pair $(\mathbf{q}, \mathbf{y}) \sim \mathcal{P}_q \times \mathcal{P}_y$. Expanding the squared Euclidean distance gives

$$\|\mathbf{q} - \mathbf{y}\|^2 \;=\; \mathbf{q}^\top \mathbf{q} \;+\; \mathbf{y}^\top \mathbf{y} \;-\; 2\,\mathbf{q}^\top \mathbf{y}. \tag{7}$$

Let $\boldsymbol{\mu}_q = \mathbb{E}[\mathbf{q}]$, $\boldsymbol{\mu}_y = \mathbb{E}[\mathbf{y}]$, $\boldsymbol{\Sigma}_q = \mathrm{Cov}(\mathbf{q})$, $\boldsymbol{\Sigma}_y = \mathrm{Cov}(\mathbf{y})$, $\boldsymbol{\Sigma}_{qy} = \mathrm{Cov}(\mathbf{q}, \mathbf{y})$. Taking expectations term by term yields

$$\begin{aligned}\mathbb{E}[\mathbf{q}^\top \mathbf{q}] &= \|\boldsymbol{\mu}_q\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_q), \\ \mathbb{E}[\mathbf{y}^\top \mathbf{y}] &= \|\boldsymbol{\mu}_y\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_y), \\ \mathbb{E}[\mathbf{q}^\top \mathbf{y}] &= \boldsymbol{\mu}_q^\top \boldsymbol{\mu}_y + \mathrm{tr}(\boldsymbol{\Sigma}_{qy}). \end{aligned} \tag{8}$$

Substituting (8) into (7) gives the decomposition

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{q} - \mathbf{y}\|^2\big] &= \|\boldsymbol{\mu}_q - \boldsymbol{\mu}_y\|^2 \\ &\quad + \mathrm{tr}\big(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_y - 2\boldsymbol{\Sigma}_{qy}\big). \end{aligned} \tag{9}$$

Independence. When the query and document are sampled independently—the situation at inference time—$\boldsymbol{\Sigma}_{qy} = \mathbf{0}$, and (9) reduces to

$$\mathbb{E}\big[\|\mathbf{q} - \mathbf{y}\|^2\big] \;=\; \mathrm{MMD}^2_{\mathrm{lin}} + \mathrm{tr}(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_y), \tag{10}$$

i.e. an affine function of the linear kernel MMD (§A, Eq. (5)).

In conclusion, under independence, the expectation of $\|\mathbf{q} - \mathbf{y}\|_2^2$ equals the linear–kernel MMD plus a constant that depends solely on the marginal variances $\mathrm{tr}(\boldsymbol{\Sigma}_q)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_y)$. Since $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_y$ are covariance matrices and thus positive semi-definite, their traces are non-negative. It follows that

$$\mathbb{E}\big[\|\mathbf{q} - \mathbf{y}\|^2\big] \geq \mathrm{MMD}^2_{\mathrm{lin}}. \tag{11}$$

This implies that an increase in the squared ℓ_2 distance necessarily indicates an increase in MMD.

B In-Batch ℓ_2 Regularization and MMD

A training batch contains one query embedding $\mathbf{q} = E_q(q)$ and N document embeddings $\mathbf{y}_1, \ldots, \mathbf{y}_N \overset{\text{i.i.d.}}{\sim} \mathcal{P}_y$. The additional loss is

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{q} - \mathbf{y}_i\|^2.$$

Because every $\mathbf{y}_i$ is independently drawn from $\mathcal{P}_y$,

$$\mathbb{E}[\|\mathbf{q} - \mathbf{y}_i\|^2] = \|\boldsymbol{\mu}_q - \boldsymbol{\mu}_y\|^2 + \operatorname{tr}\big(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_y\big),$$

and taking the expectation of the whole sum gives

$$\mathbb{E}\big[\mathcal{L}_{\text{reg}}\big] = \text{MMD}^2_{\text{lin}} + \operatorname{tr}\big(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_y\big). \tag{12}$$

The second term on the right-hand side depends only on the within-distribution covariance of the two encoders. Since every covariance matrix is positive semidefinite, its trace—the sum of marginal variances—is *non-negative*. Therefore

$$\mathcal{L}_{\text{reg}} \;\geq\; \text{MMD}^2_{\text{lin}}.$$

Minimizing $\mathcal{L}_{\text{reg}}$ thus tightens a valid upper bound on the linear-kernel MMD: any decrease in $\mathcal{L}_{\text{reg}}$ must include an equal or larger decrease in $\text{MMD}^2_{\text{lin}}$. Conversely, reducing $\text{MMD}^2_{\text{lin}}$ immediately reduces $\mathcal{L}_{\text{reg}}$ by the same amount, establishing the two quantities as *tightly coupled*.

Adding

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda\, \mathcal{L}_{\text{reg}}$$

to the training objective therefore preserves instance-level discrimination through $\mathcal{L}_{\text{con}}$, while $\mathcal{L}_{\text{reg}}$ continuously narrows the population-level gap between query and document embeddings, achieving distributional alignment without explicit covariance estimation or added indexing cost.

C InfoNCE Loss Enlarges Linear Kernel MMD

This subsection provides a theoretical proof that training only with the InfoNCE loss drives the MMD between query and document embeddings upward when the inner product is used as the metric. Consider one query embedding $\mathbf{q} = E_q(q)$ and $K+1$ document embeddings $\mathbf{y}^+, \mathbf{y}_1^-, \ldots, \mathbf{y}_K^- \overset{\text{i.i.d.}}{\sim} \mathcal{P}_y$. With inner-product similarity $s(\mathbf{q}, \mathbf{y}) = \mathbf{q}^\top \mathbf{y}$, the InfoNCE loss is

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{k=1}^{K} \exp(s_k^-/\tau)},$$

where $s^+ = s(\mathbf{q}, \mathbf{y}^+)$ and $s_k^- = s(\mathbf{q}, \mathbf{y}_k^-)$.

Let $d_t = \|\boldsymbol{\mu}_q^{(t)} - \boldsymbol{\mu}_y^{(t)}\|^2$ and update the means by $\boldsymbol{\mu}_q^{(t+1)} = \boldsymbol{\mu}_q^{(t)} + \Delta_q$, $\boldsymbol{\mu}_y^{(t+1)} = \boldsymbol{\mu}_y^{(t)} + \Delta_y$. The difference evolves as

$$\begin{aligned} d_{t+1} - d_t &= \|\boldsymbol{\mu}_q^{(t+1)} - \boldsymbol{\mu}_y^{(t+1)}\|^2 - \|\boldsymbol{\mu}_q^{(t)} - \boldsymbol{\mu}_y^{(t)}\|^2 \\ &= \|(\boldsymbol{\mu}_q^{(t)} + \Delta_q) - (\boldsymbol{\mu}_y^{(t)} + \Delta_y)\|^2 - \|\boldsymbol{\mu}_q^{(t)} - \boldsymbol{\mu}_y^{(t)}\|^2 \\ &= 2(\boldsymbol{\mu}_q - \boldsymbol{\mu}_y)^\top (\Delta_q - \Delta_y) + \|\Delta_q - \Delta_y\|^2 \end{aligned} \tag{13}$$

We omit the superscript t since all variables involved are at time t in Eq 13. Given that $\|\Delta_q - \Delta_y\|^2 \geq 0$ always holds, to prove that the MMD increases, we only need to show that $2(\boldsymbol{\mu}_q - \boldsymbol{\mu}_y)^\top (\Delta_q - \Delta_y) \geq 0$

Proof: $2(\boldsymbol{\mu}_q - \boldsymbol{\mu}_y)^\top (\Delta_q - \Delta_y) \geq 0$

We define softmax weights as:

$$\sigma^+ = \frac{e^{s^+/\tau}}{Z}, \quad \sigma_k^- = \frac{e^{s_k^-/\tau}}{Z}, \tag{14}$$

where $Z = \exp(s^+/\tau) + \sum_{k=1}^{K} \exp(s_k^-/\tau)$. Because all documents are drawn from the same distribution $\mathcal{P}_y$, the scores $s^+, s_1^-, \ldots, s_K^-$ are identically distributed and exchangeable. This symmetry yields

$$\mathbb{E}[\sigma^+] = \mathbb{E}[\sigma_1^-] = \cdots = \mathbb{E}[\sigma_K^-],$$

and with $\sigma^+ + \sum_{k=1}^{K} \sigma_k^- = 1$ we obtain

$$\mathbb{E}[\sigma^+] = \mathbb{E}[\sigma_k^-] = \frac{1}{K+1}. \tag{15}$$

For the query embedding, omitting the constant $1/\tau$, the gradient is

$$\partial_{\mathbf{q}} \mathcal{L}_{\text{con}} = \sigma^+(-\mathbf{y}^+) + \sum_{k=1}^{K} \sigma_k^- \mathbf{y}_k^-,$$

whose expectation, using (15), becomes

$$\begin{aligned} \mathbb{E}[\partial_{\mathbf{q}} \mathcal{L}_{\text{con}}] &= \mathbb{E}[\sigma^+](-\boldsymbol{\mu}_y) + K\mathbb{E}[\sigma_k^-]\, \boldsymbol{\mu}_y \\ &= \left(\tfrac{K-1}{K+1}\right) \boldsymbol{\mu}_y. \end{aligned}$$

Therefore, the expected increment Δ_q of μ_q in gradient descent optimization can be written as:

$$\Delta_q = \mathbb{E}[-\eta \partial_{\mathbf{q}} \mathcal{L}_{\text{con}}] = -\eta \left(\tfrac{K-1}{K+1}\right) \boldsymbol{\mu}_y, \tag{16}$$

where $\eta > 0$ is the learning rate and K the number of negatives.

For documents, the positive receives gradient $-\sigma^{+}\mathbf{q}$ and each negative $\sigma_k^{-}\mathbf{q}$. Averaging over the batch and over all documents gives

$$\Delta_y = -\eta\Big[-\mathbb{E}[\sigma^{+}] + \sum_{k=1}^{K}\mathbb{E}[\sigma_k^{-}]\Big]\boldsymbol{\mu}_q$$
$$= -\eta\left(\tfrac{K-1}{K+1}\right)\boldsymbol{\mu}_q. \quad (17)$$

Substituting (16) and (17) into $\Delta_q - \Delta_y$ yields

$$\Delta_q - \Delta_y = \eta\Big(\tfrac{K-1}{K+1}\Big)(-\boldsymbol{\mu}_y + \boldsymbol{\mu}_q). \quad (18)$$

Plugging (18) into the first term of (13) gives

$$2(\boldsymbol{\mu}_q - \boldsymbol{\mu}_y)^{\top}(\Delta_q - \Delta_y)$$
$$= 2\eta\Big(\tfrac{K-1}{K+1}\Big)(\boldsymbol{\mu}_q - \boldsymbol{\mu}_y)^{\top}(-\boldsymbol{\mu}_y + \boldsymbol{\mu}_q)$$
$$= 2\eta\Big(\tfrac{K-1}{K+1}\Big)\|\boldsymbol{\mu}_q - \boldsymbol{\mu}_y\|^2 \;\geq\; 0. \quad (19)$$

Since this non-negative quantity is the dominant term in (13), the squared mean gap—and hence the linear kernel MMD—does not decrease and increases unless it is already zero. Therefore, the InfoNCE loss with an inner-product metric drives the query and document distributions farther apart as training proceeds. This theoretical prediction aligns with the empirical curves presented in the main text.

D Detailed Related Work

Embedding Based Retrieval The basic process of EBR involves generating a semantic embedding from an encoder model, followed by a retrieval process that recalls the corresponding embeddings. The remarkable efficacy of Pre-trained Language Models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) has established them as foundational encoder architectures. The bi-encoder (Bromley et al., 1993) architecture is a common choice that typically includes a query encoder and a base encoder to generate embeddings for queries and the bases (search corpus), respectively, offering a flexible and efficient framework. Recently, advanced models based on the bi-encoder architecture (Gao and Callan, 2021; Wu et al., 2023; Qu et al., 2020; Xiong et al., 2020; Zhang et al., 2022; Ren et al., 2021) have been proposed. These approaches have effectively enhanced the embedding representation capabilities of models, significantly boosting search performance.

Approximate Nearest Neighbor Search ANN search is a key component in the EBR process, achieving the task of finding the nearest k neighbors to a query embedding within the base embeddings. As data scales increasingly reach millions or even billions, the exact k-Nearest Neighbors (kNN) method, which requires traversing the entire high-dimensional embedding dataset, becomes impractically slow. Thus, Approximate Nearest Neighbor (ANN) algorithms have emerged. ANN algorithms build indexes on base embeddings and return the approximately nearest k neighbors to a query, trading off search accuracy for reduced latency. Index building methods for ANN include cluster-based (e.g., IVF, ScaNN (Guo et al., 2020)) and graph-based (e.g., HNSW (Malkov and Yashunin, 2018), DiskANN (Jayaram Subramanya et al., 2019)) approaches. Although the aforementioned methods are effective, there remain some problems for ANN search, as it depends heavily on the transitivity of near neighbors between bases. Recent work shows that OOD queries represent a new problem for ANN search (Aumüller and Ceccarello, 2023), where the query embeddings significantly diverge in distribution from the base embeddings, with more difficulty in finding nearest neighbors.

Out-of-Distribution Issue As most machine learning methods are based on the assumption of independent and identically distributed data, which is rarely met in real-world conditions (Arjovsky et al., 2019; Liu et al., 2021), OOD issues are prevalent across the machine learning field, including Natural Language Processing (NLP) (Yuan et al., 2024) and Computer Vision (CV) (Hendrycks and Gimpel, 2016; Liu et al., 2020). While extensive research has focused on OOD detection (Yang et al., 2021) and OOD generalization (Liu et al., 2021) problems, studies on how OOD affects ANN search are scarce. The recently released Yandex Text2Image (Yandex Research, 2021) dataset is naturally designed for OOD scenarios, with base embeddings from images inferred by Se-ResNext-101 (Hu et al., 2018) model and query embeddings from text produced by a variant of the DSSM (Huang et al., 2013) model, showing substantial distribution differences. In OOD-DiskANN (Jaiswal et al., 2022), they defined OOD queries in ANN search and analyzed the challenges posed by OOD queries, eventually proposing effective algorithmic improvements. Recently, RoarGraph (Chen et al., 2024a) employs a projected bipartite graph approach tai-

lored for OOD queries, achieving up to 3.3× improved search efficiency and won the NeurIPS'23 Big-ANN OOD track.

E Detailed Experimental Setup

Implementation Details. The regularization strength λ in Equation 4 is chosen such that the regularization loss remains approximately one order of magnitude smaller than the contrastive loss. We set $\lambda = 0.01$ as a general default, and provide ablation results for varying λ values.

For encoder fine-tuning, we use the official codebases and pre-trained weights of *coCondenser* and *RetroMAE*. Fine-tuning is performed on MS-MARCO using their respective training configurations and hyperparameters.

ANN Index Implementations. - **FAISS**: IVF, IVFPQ, and HNSW are implemented via FAISS (Johnson et al., 2019). - **DiskANN & RoarGraph**: We use the official repositories to build and query DiskANN (Simhadri et al., 2023) and RoarGraph (Chen et al., 2024a) indices, and they were compiled and run from their official repositories.

Hardware and Environment. All finetuning is done on 4×NVIDIA V100 GPUs (32GB) using PyTorch 1.11.0 and HuggingFace Transformers 4.40.2. Index construction and search are run on a server with dual Intel Xeon Platinum 8168 CPUs and 503GB of RAM. All search experiments are run single-threaded, where QPS is equivalent to per-query latency.

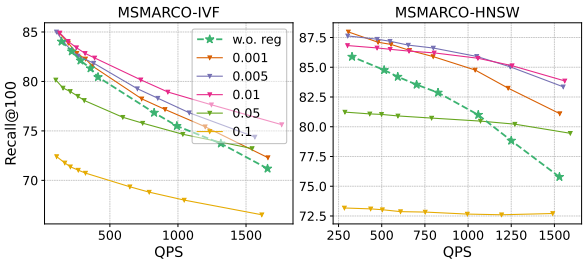


Figure 8: QPS-recall results obtained with different constraint strengths of λ. The larger the λ, the greater the reduction in the distance between the query and base distributions.

F Ablation on Impact of Regularization Strength λ

Figure 8 shows the performance of embeddings obtained with different values of λ in 4 on IVF and HNSW searches. Selecting an appropriate λ is critical for balancing the model's discriminative power, driven by $\mathcal{L}_{\text{con}}$, and the distribution alignment, enforced by $\mathcal{L}_{\text{reg}}$. In our main experiments, we set λ such that the regularization loss is approximately one order of magnitude smaller than the contrastive loss. Here, we present the impact of different λ values on ANN search performance, demonstrating that the choice of λ is relatively robust across a range of settings.

Based on the results from Figure 8, When the regularization weight is less than 0.05, performance improvements are observed across both ANN search methods, especially in high-QPS scenarios. However, if the regularization strength is too high, such as with a weight of 0.1, the ANN search curves fall significantly below the baseline (without regularization), harming search performance. This suggests that overemphasizing proximity between query and base vectors can drastically reduce the distinguishability between them, severely impacting search accuracy.

Within the effective range of regularization, higher regularization strength shows more pronounced improvements in the "high-QPS, low-recall" region, while lower regularization strength tends to improve performance in the "low-QPS, high-recall" region.