

GLOBE: Training-Free Calibration for Image-Text Retrieval with Global-Local Bias and Graph Evidence

Xuzhe Zhao^{123*}, Zhirun Yue^{1*}, Xingyue Xu⁴, Houde Liu^{1†}, Kaichen Dong^{123†}

¹Institute of Data and Information; ²Center of Double Helix

^{1,2}Tsinghua Shenzhen International Graduate School, Tsinghua University

³Intelligent Passive Thermal Control Center, Research Institute of Tsinghua University

⁴Faculty of Information Technology, Monash University

Correspondence: dkc22@sz.tsinghua.edu.cn, liu.hd@sz.tsinghua.edu.cn

Abstract

Multimodal contrastive models are widely used for bidirectional image-text retrieval (Radford et al., 2021), but their rankings remain vulnerable to attraction bias, where certain candidates become hubs for many unrelated queries. Existing training-free normalization methods suppress such hubs (Chowdhury et al., 2024), but may also penalize candidates that are genuinely relevant. We propose GLOBE, a training-free plug-in module that combines global-local bias calibration with graph-based semantic evidence to suppress unsupported hubs while preserving true matches. Across MS-COCO and Flickr30K, GLOBE improves diverse frozen retrieval backbones, achieving up to 7.9% absolute Recall@1 gain over the original model and up to 5.2% over strong training-free baselines. We further construct IRB-COCO-500, a mixed real-and-synthetic benchmark with AI-generated near-positive distractors, and show that GLOBE improves retrieval robustness in this challenging setting.

1 Introduction

Contrastive vision-language models, pioneered by CLIP, are standard for large-scale image-text retrieval. These dual-encoders learn a shared space by aligning matched pairs and separating mismatched ones, enabling efficient similarity-based retrieval (Radford et al., 2021; Zhou et al., 2022). However, raw embedding similarities often yield suboptimal ranking when downstream retrieval requires distinguishing semantically similar candidates with fine-grained differences (Lin et al., 2025). One important reason is hubness: in high-dimensional embedding spaces, some candidates become nearest neighbors for many unrelated queries (Radovanović et al., 2010).

To mitigate hubness-induced errors, recent training-free methods calibrate retrieval scores

without updating parameters (Pan et al., 2025). QB-Norm and DBNorm use query or gallery banks to normalize hub-prone scores (Bogolin et al., 2022; Wang et al., 2023b), Distribution Normalization uses global reference statistics (Zhou et al., 2023), and Nearest Neighbor Normalization (NNN) estimates candidate-level bias from nearest reference queries (Chowdhury et al., 2024). Beyond bias-only normalization, neighborhood-based methods such as CSLS (Lample et al., 2018), query expansion (Chum et al., 2007), and k-reciprocal re-ranking (Zhong et al., 2017) exploit local structure to refine retrieval.

Treating hubness as purely harmful is incomplete. As shown in Figure 1(L), the same high-attraction candidate may be a spurious hub for one query, such as the red-boxed candidate with respect to q_1 , but a genuine match for another query, such as q_2 . Bias-only correction cannot distinguish these cases, since it penalizes candidates according to their overall attraction and may exclude true matches from top ranks.

We propose GLOBE, a training-free calibration method that combines global-local bias correction with graph-based semantic evidence. GLOBE operates on cached embeddings and requires no model updates. Unlike bias-only calibration, it uses neighborhood structure to determine whether a high-scoring query-candidate pair is semantically supported, thereby suppressing unsupported hubs while recovering true matches that would otherwise be over-penalized. Figure 1(R) empirically supports this design: under a CLIP ViT-B/32 COCO setting, GLOBE recovers 178 of 461 (38.6%) ground-truth pairs degraded by bias-only calibration, improving their average rank by 4.62.

We evaluate GLOBE on text-to-image and image-to-text retrieval using Flickr30k and MS-COCO (Plummer et al., 2015; Lin et al., 2014) across CLIP, SigLIP, BLIP, ALBEF, and BEiT-3 backbones (Radford et al., 2021; Zhai et al., 2023;

*Equal contribution.

†Co-corresponding authors.

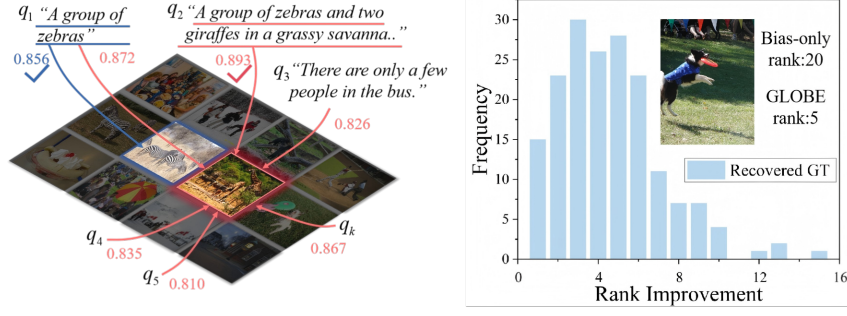


Figure 1: Motivation of GLOBE. (L) Hub-like candidates can be either spurious matches or true positives. (R) GLOBE recovers ground-truth pairs pushed out of the top- K by bias-only calibration on CLIP+COCO.

Li et al., 2022, 2021; Wang et al., 2023a). GLOBE improves retrieval accuracy over original models and strong training-free baselines in multiple settings. We further construct IRB-COCO-500, a hybrid real-synthetic benchmark featuring generated images that align with coarse caption semantics but lack original ground-truth identity, and show that GLOBE improves robustness to AI-generated near-positive distractors.

Our contributions are threefold: (1) we identify that bias-only hubness correction can over-penalize true semantic matches; (2) we introduce GLOBE, a training-free global-local graph calibration method for frozen image-text retrievers; and (3) we demonstrate improvements on T2I/I2T retrieval, recovered-pair analysis, and robustness to AI-generated near-positive distractors on our self-constructed IRB-COCO-500 benchmark.

2 GLOBE: Global-Local Bias Calibration with Graph Evidence

Given a query q and a candidate r , a contrastive retriever computes a similarity score $S(q, r)$, typically the dot product between normalized text and image embeddings. GLOBE is a training-free post-hoc calibration method operating on cached embeddings. As shown in Figure 2, it re-ranks candidates by combining the original similarity, global-local bias suppression, and graph-based semantic recovery:

$$\hat{S}(q, r) = N(S(q, r)) - \mathcal{B}(q, r) + \mathcal{G}(q, r), \quad (1)$$

where $N(\cdot)$ denotes query-wise min-max normalization over the candidate list, $\mathcal{B}(q, r)$ penalizes attraction bias, and $\mathcal{G}(q, r)$ restores neighborhood-supported matches.

The bias term combines global candidate bias and local candidate density:

$$\mathcal{B}(q, r) = \mu N(b_{K_g}(r)) + \lambda N(d_m(r)). \quad (2)$$

Here, $b_{K_g}(r)$ is estimated from the top- K_g reference queries most similar to candidate r . The local density term is

$$d_m(r) = \frac{1}{|\mathcal{A}_m(r)|} \sum_{q' \in \mathcal{A}_m(r)} z_{q'}(r), \quad (3)$$

where $\mathcal{A}_m(r)$ contains reference queries whose top- m candidate list includes r , and $z_{q'}(r)$ is the local z-score of r . Thus, b_{K_g} captures global attraction, while d_m captures repeated local over-density.

To avoid over-penalizing valid hub-like matches, GLOBE computes graph evidence from the retrieval score matrix:

$$\begin{aligned} g_{k_q}^Q(q, r) &= \sum_{q' \in \mathcal{N}_{k_q}^Q(q)} w_{qq'}^Q S(q', r), \\ g_{k_r}^R(q, r) &= \sum_{r' \in \mathcal{N}_{k_r}^R(r)} w_{rr'}^R S(q, r'). \end{aligned} \quad (4)$$

The recovery term is $\mathcal{G}(q, r) = \eta_q N(g_{k_q}^Q(q, r)) + \eta_r N(g_{k_r}^R(q, r))$, where $\mathcal{N}_{k_q}^Q(q)$ and $\mathcal{N}_{k_r}^R(r)$ are the nearest query and candidate neighborhoods, and w^Q, w^R are normalized neighbor weights. Query-side evidence checks whether neighboring queries support r , while candidate-side evidence checks whether candidates near r remain compatible with q . Overall, GLOBE suppresses globally or locally over-attractive candidates through $\mathcal{B}$, while recovering graph-supported semantic matches through $\mathcal{G}$.

3 Experiments and Ablations

3.1 Retrieval Accuracy of GLOBE

We evaluate GLOBE on standard text-to-image (T2I) and image-to-text (I2T) retrieval benchmarks using Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014) in table 1 and table 2. Following prior training-free normalization work, we

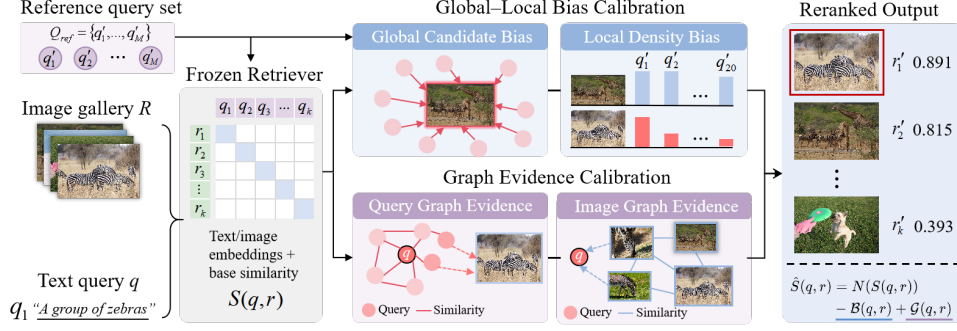


Figure 2: Overview of GLOBE. GLOBE calibrates frozen image-text retrieval scores by suppressing global candidate and local density biases, while using query- and image-side graph evidence to recover relevant pairs.

Table 1: Text-to-image retrieval Recall@1 results on Flickr30k and COCO. Best results within each retrieval dataset and model row are bolded. “ft.” indicates finetuned.

Model	Flickr30k retrieval					COCO retrieval				
	Orig.	DN	DBNorm	NNN	GLOBE	Orig.	DN	DBNorm	NNN	GLOBE
CLIP ViT-B/32	56.52	60.42 (+3.9)	61.56 (+5.0)	62.92 (+6.4)	64.38 (+7.9)	30.24	32.35 (+2.1)	33.49 (+3.3)	36.80 (+6.6)	37.92 (+7.7)
SigLIP ViT-B/16	73.94	73.18 (-0.8)	75.96 (+2.0)	76.10 (+2.2)	76.68 (+2.7)	47.78	47.76 (+0.0)	49.67 (+1.9)	51.05 (+3.3)	51.50 (+3.7)
BLIP pretrained	67.02	69.20 (+2.2)	67.64 (+0.6)	70.28 (+3.3)	72.66 (+5.6)	44.87	48.79 (+3.9)	47.35 (+2.5)	51.61 (+6.7)	51.61 (+6.7)
BLIP ft. Flickr	82.54	80.16 (-2.4)	82.04 (-0.5)	83.44 (+0.9)	83.54 (+1.0)	54.68	53.83 (-0.9)	55.67 (+1.0)	57.74 (+3.1)	58.05 (+3.4)
BLIP ft. COCO	78.66	76.64 (-2.0)	79.02 (+0.4)	80.38 (+1.7)	80.58 (+1.9)	57.31	57.06 (-0.3)	58.31 (+1.0)	60.18 (+2.9)	60.24 (+2.9)
ALBEF ft. COCO	72.74	72.34 (-0.4)	74.34 (+1.6)	75.34 (+2.6)	75.66 (+2.9)	53.06	53.45 (+0.4)	54.36 (+1.3)	55.75 (+2.7)	55.74 (+2.7)
ALBEF ft. Flickr	78.62	76.74 (-1.9)	80.08 (+1.5)	79.74 (+1.1)	80.08 (+1.5)	49.52	50.15 (+0.6)	53.53 (+4.0)	53.07 (+3.6)	53.42 (+3.9)
BEiT3-base-ipc	74.36	72.86 (-1.5)	75.42 (+1.1)	76.52 (+2.1)	77.54 (+3.2)	47.52	45.13 (-2.4)	47.96 (+0.4)	50.25 (+2.8)	50.79 (+3.3)
BEiT3 ft. Flickr	85.52	81.46 (-4.1)	84.52 (-1.0)	85.52 (+0.0)	85.50 (+0.0)	52.78	49.33 (-3.5)	52.57 (-0.2)	55.57 (+2.8)	56.01 (+3.2)
BEiT3 ft. COCO	82.16	78.80 (-3.4)	81.68 (-0.5)	82.94 (+0.8)	83.52 (+1.4)	61.38	58.76 (-2.6)	60.87 (-0.5)	61.79 (+0.4)	61.27 (-0.1)
BEiT3-large-ipc	69.82	68.40 (-1.4)	71.94 (+2.1)	71.74 (+1.9)	72.12 (+2.3)	43.07	41.70 (-1.4)	45.81 (+2.7)	45.65 (+2.6)	45.86 (+2.8)
BEiT3 large ft. COCO	76.46	74.00 (-2.5)	77.40 (+0.9)	77.52 (+1.1)	77.58 (+1.1)	56.35	53.45 (-2.9)	55.43 (-0.9)	56.71 (+0.4)	56.91 (+0.6)
BEiT3 large ft. Flickr	77.74	74.16 (-3.6)	76.80 (-0.9)	77.84 (+0.1)	77.90 (+0.2)	47.98	45.70 (-2.3)	50.73 (+2.8)	50.00 (+2.0)	50.77 (+2.8)

compare against the original retriever, Distribution Normalization (DN), DBNorm, and Nearest Neighbor Normalization (NNN). For NNN, we use the model-, dataset-, reference-, and retrieval-direction-specific optimal hyperparameters reported in the original NNN appendix, rather than re-tuning them. The main tables report Recall@1, while complete T2I and I2T results with Recall@5 and Recall@10 are provided in Appendix E and Appendix D, respectively. We additionally evaluate CSLS and the recent Dual Bank Sinkhorn Normalization (DBSN) on CLIP and SigLIP; these complementary comparisons are reported in Appendix A.4.

Table 1 shows that GLOBE improves T2I retrieval across diverse backbones and datasets. On CLIP ViT-B/32, GLOBE improves Recall@1 by +7.86 on Flickr30K and +7.68 on COCO over the original baseline, and also outperforms strong normalization baselines. Table 2 reports similar trends for I2T retrieval, where GLOBE yields larger gains on pretrained backbones and smaller but consistent improvements on strongly fine-tuned models whose Recall@1 is already close to saturation.

Fine-tuning and GLOBE are complementary rather than competing: fine-tuning can remove much of the representation error and leave less headroom for post-hoc calibration, whereas GLOBE targets remaining retrieval bias without updating model parameters.

Additional comparisons with CSLS and the recent DBSN baseline on CLIP and SigLIP show that GLOBE achieves the best or tied-best Recall@1 across all evaluated settings (Appendix A.4).

These results indicate that GLOBE is not simply a stronger hubness penalty: by combining global-local bias suppression with graph consistency, it suppresses unsupported hubs while preserving semantically valid pairs.

Cross-dataset hyperparameter transfer. To evaluate whether GLOBE requires target-specific hyperparameter re-selection, we directly transfer hyperparameters between Flickr30K and COCO while retaining each target dataset’s own in-domain reference bank.

As shown in Table 3, transferred hyperparameters improve the original retriever in all 12 settings

Table 2: Image-to-text retrieval Recall@1 results on Flickr30k and COCO. Best results within each retrieval dataset and model row are bolded. “ft.” indicates finetuned.

Model	Flickr30k retrieval					COCO retrieval				
	Orig.	DN	DBNorm	NNN	GLOBE	Orig.	DN	DBNorm	NNN	GLOBE
CLIP ViT-B/32	77.7	73.8 (-3.9)	78.1 (+0.4)	80.1 (+2.4)	81 (+3.3)	49.14	45.06 (-4.1)	49.3 (+0.2)	53.1 (+4.0)	53.86 (+4.7)
SigLIP ViT-B/16	88.9	84.7 (-4.2)	90.1 (+1.2)	91.1 (+2.2)	92.2 (+3.3)	65.72	60.1 (-5.6)	65.22 (-0.5)	69.96 (+4.2)	70.37 (+4.7)
BLIP pretrained	83	86.9 (+3.9)	88.4 (+5.4)	89.1 (+6.1)	90.2 (+7.2)	61.88	67.24 (+5.4)	68.74 (+6.9)	71.68 (+9.8)	72.18 (+10.3)
BLIP ft. Flickr	93.4	93.6 (+0.2)	95 (+1.6)	95.8 (+2.4)	95.9 (+2.5)	69.9	71.84 (+1.9)	73.1 (+3.2)	76.4 (+6.5)	76.8 (+6.9)
BLIP ft. COCO	90.8	90.8 (+0.0)	92.1 (+1.3)	92.8 (+2.0)	93 (+2.2)	75.76	77.34 (+1.6)	78 (+2.2)	79.1 (+3.3)	79.1 (+3.3)
ALBEF ft. COCO	86.6	84.7 (-1.9)	89.4 (+2.8)	89.5 (+2.9)	90.2 (+3.6)	71.42	70.1 (-1.3)	73.04 (+1.6)	74.1 (+2.7)	74.24 (+2.8)
ALBEF ft. Flickr	90.6	89.2 (-1.4)	91.4 (+0.8)	90.8 (+0.2)	91.6 (+1.0)	66.32	65.42 (-0.9)	71.7 (+5.4)	70.64 (+4.3)	71.7 (+5.4)
BEiT3-base-its	90	87.1 (-2.9)	90.2 (+0.2)	92.1 (+2.1)	92.9 (+2.9)	59.74	60.54 (+0.8)	64.82 (+5.1)	69.34 (+9.6)	69.84 (+10.1)
BEiT3 ft. Flickr	95.6	93.5 (-2.1)	95.1 (-0.5)	95.3 (-0.3)	95.1 (-0.5)	71.02	65.64 (-5.4)	70.86 (-0.2)	74.9 (+3.9)	75.2 (+4.2)
BEiT3 ft. COCO	93.8	92.2 (-1.6)	94.5 (+0.7)	95.1 (+1.3)	95.2 (+1.4)	79.06	74.12 (-4.9)	78.04 (-1.0)	80.52 (+1.5)	80.24 (+1.2)
BEiT3-large	82.9	83.4 (+0.5)	87.2 (+4.3)	87 (+4.1)	87.5 (+4.6)	55.46	56.04 (+0.6)	62 (+6.5)	62.66 (+7.2)	62.98 (+7.5)
BEiT3 large ft. COCO	87.8	86.5 (-1.3)	89 (+1.2)	89 (+1.2)	89.8 (+2.0)	73.74	69.14 (-4.6)	74.94 (+1.2)	75.3 (+1.6)	75.66 (+1.9)
BEiT3 large ft. Flickr	88	87.9 (-0.1)	89.7 (+1.7)	89.9 (+1.9)	90.3 (+2.3)	63.28	61.26 (-2.0)	68.08 (+4.8)	67.98 (+4.7)	68.16 (+4.9)

Table 3: Cross-dataset transfer of GLOBE hyperparameters between Flickr30K and COCO. Each target uses its own in-domain reference bank; “Target” and “Transfer” denote target-specific and transferred hyperparameters, respectively.

Source → Target	Model	Dir.	Orig.	NNN	Target	Transfer
F30K → COCO	CLIP	T2I	30.24	36.80	37.92	37.46
F30K → COCO	SigLIP	T2I	47.78	51.05	51.50	51.02
F30K → COCO	BEiT3-base	T2I	47.52	50.25	50.79	50.43
COCO → F30K	CLIP	T2I	56.52	62.92	64.38	63.98
COCO → F30K	SigLIP	T2I	73.94	76.10	76.68	76.17
COCO → F30K	BEiT3-base	T2I	74.36	76.52	77.54	77.29
F30K → COCO	CLIP	I2T	49.14	53.10	53.86	53.29
F30K → COCO	SigLIP	I2T	65.72	69.96	70.37	70.06
F30K → COCO	BEiT3-base	I2T	59.74	69.34	69.84	69.40
COCO → F30K	CLIP	I2T	77.70	80.10	81.00	80.31
COCO → F30K	SigLIP	I2T	88.90	91.10	92.20	91.85
COCO → F30K	BEiT3-base	I2T	90.00	92.10	92.90	92.23

and outperform NNN in 11 of 12 settings, while remaining only 0.46 R@1 points below target-specific GLOBE on average. These results indicate that target-specific selection improves peak accuracy but is not required to retain most of GLOBE’s gains. Additional cross-model transfer results are provided in Appendix B.4.

The nearest-neighbor operations in GLOBE can be accelerated with Faiss; the corresponding runtime analysis is provided in Appendix G.

3.2 Robustness to AI-generated Near-positive Distractors

We further evaluate GLOBE under mixed real-and-synthetic retrieval using IRB-COCO-500, a stress test built from 500 COCO image-caption pairs. For each real image, we generate one SDXL image from the corresponding human caption and construct a hybrid gallery containing both real images and caption-consistent synthetic near-positives. We compare the original CLIP retriever,

NNN-calibrated retrieval, and GLOBE-calibrated retrieval to test whether calibration preserves the annotated real target without increasing preference for AI-generated images.

Figure 3(L) shows a randomly sampled query example. CLIP and NNN rank the paired AI-generated counterpart ahead of the annotated real image, while GLOBE promotes the true ground-truth image and reduces the tendency to place the generated image at the top of the list. Figure 3(R) quantifies this behavior: GLOBE achieves the highest RealTop1Rate while obtaining the lowest GenPreferenceRate, where GenPreferenceRate measures preference for AI-generated counterparts over the real target.

We define *RealTop1Rate* as the fraction of queries for which the annotated real COCO image is ranked first, and *GenPreferenceRate* as the fraction for which the paired AI-generated distractor is ranked above the real target. Formal definitions are provided in Appendix F.



Figure 3: Robustness to AI-generated near-positive distractors. (L) Top-10 retrieval results for a randomly sampled IRB-COCO-500 query under CLIP, NNN, and GLOBE. Red boxes mark the paired ground-truth image. (R) Quantitative IRB-COCO-500 results.

Table 4: Ablation study of GLOBE components on text-to-image retrieval using CLIP ViT-B/32.

Method	Flickr30k		COCO	
	R@1	Δ R@1	R@1	Δ R@1
Original	56.52	0.00	30.24	0.00
Global only	62.94	+6.42	36.51	+6.27
Local only	59.92	+3.40	35.13	+4.89
Graph only	56.68	+0.16	29.70	-0.54
Global + Local	62.98	+6.46	36.49	+6.25
Global + Graph	63.62	+7.10	36.05	+5.81
Local + Graph	59.92	+3.40	34.36	+4.12
Full GLOBE	64.38	+7.86	37.92	+7.68

3.3 Ablation Results

We conduct ablations on CLIP and SigLIP to assess the contribution of each component in GLOBE. As shown in Table 4, global bias calibration provides the largest individual gain, confirming that broad candidate attraction is a dominant failure mode. Local density calibration also improves performance, suggesting that retrieval bias is amplified by locally dense candidate neighborhoods.

In contrast, graph evidence alone provides only a small gain on Flickr30K (+0.16 R@1) and slightly degrades COCO performance (-0.54 R@1). Since graph evidence is computed on the uncalibrated similarity space, high neighborhood connectivity in hub-dominated regions may reflect popularity bias rather than semantic consistency. Global-local calibration first suppresses these distortions, making graph evidence more reliable for recovering semantically supported matches.

On CLIP, full GLOBE improves R@1 by +7.86 on Flickr30K and +7.68 on COCO over the original baseline. These results show that global-local bias calibration and graph-based semantic evidence are complementary. Full ablation results on CLIP and SigLIP are reported in Appendix C.

4 Conclusion

We presented GLOBE, a training-free global-local graph calibration method for frozen image-text retrieval. By combining bias suppression with semantic recovery, GLOBE mitigates hubness-induced errors while preserving genuine cross-modal matches that may otherwise be over-penalized by bias-only correction. Across standard T2I/I2T benchmarks and the mixed real-synthetic IRB-COCO-500 setting, GLOBE improves retrieval accuracy and AI-generated robustness without updating model parameters.

5 Limitations

GLOBE is a training-free reranking method and therefore depends on the quality of the frozen retriever and the reference set used for calibration. If the reference distribution is substantially mismatched with the target retrieval setting, the estimated attraction bias may be less reliable and the gain can be reduced. GLOBE also introduces several calibration hyperparameters and additional test-time computation for reference statistics and graph evidence, although embeddings and statistics can be cached in practice. Future work may explore adaptive parameter selection and more memory-efficient implementations for very large-scale retrieval. Graph evidence may also be unreliable when visually similar but semantically incorrect candidates form misleading neighborhoods. Such false semantic support may cause an incorrect candidate to be promoted by the recovery term.

6 Ethical considerations

GLOBE does not generate new content, modify input images or captions, or require additional user data beyond frozen retrieval scores, embeddings, and reference sets. However, reranking can affect the visibility of retrieved items (Wang et al., 2022; Berg et al., 2022). If the gallery or reference set contains demographic, cultural, geographic, or synthetic-data biases, GLOBE may inherit these biases or penalize genuinely relevant candidates that are common in the reference distribution. Therefore, the reference set should be selected carefully, especially in provenance-sensitive or high-stakes applications.

We construct IRB-COCO-500 as a small diagnostic benchmark derived from COCO to study robustness to AI-generated near-positive distractors. The generated images are used for research evaluation, not to claim that generated images are always less relevant than real images. When sharing this benchmark, we will report the construction protocol and avoid redistributing COCO images directly; users should follow the original COCO/Flickr licensing terms. We do not collect new personal data, and IRB-COCO-500 will be released as metadata, prompts, generation seeds, scripts, and evaluation code rather than redistributing original images. We also avoid adding prompts, annotations, or metadata that intentionally contain personally identifying or offensive content. GLOBE should be used together with dataset auditing, provenance meta-

data, and human review in sensitive settings. We report both positive and negative results to avoid overstating the effectiveness of the method and to support transparent evaluation.

7 Acknowledgements

X. Zhao and K. Dong acknowledge support from National Key R&D Program of China (grant 2023YFB3208700), National Natural Science Foundation of China (grant 62375151), Shenzhen Science and Technology Program (grant KJZD20240903095721028), Start-up funding in Tsinghua Shenzhen International Graduate School, Tsinghua University. Z. Yue and H. Liu were supported by the Shenzhen Science and Technology Program (Grant No. RCJC20210706091946001) and the Shenzhen Science and Technology Program (Grant No. ZDCY20250901104207008).

AI Assistance Disclosure

We used AI assistants for language polishing, code debugging, and checklist preparation. All scientific claims, experiments, analyses, and final writing decisions were reviewed and verified by the authors.

References

- Hugo Berg, Siobhan Hall, Yash Bhargat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822.
- Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. 2022. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. 2024. Nearest neighbor normalization improves multimodal retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22571–22582.
- Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE International Conference on Computer Vision*, pages 1–8.

- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*.
- Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755.
- Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. 2025. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9263–9273.
- Yu. A. Malkov and D. A. Yashunin. 2020. 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.
- Zhengxin Pan, Haishuai Wang, Fangyu Wu, Peng Zhang, and Jiajun Bu. 2025. Hubness reduction with dual bank sinkhorn normalization for cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6153–6162.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*.
- Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.
- Junyang Wang, Yi Zhang, and Jitao Sang. 2022. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*.
- Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023a. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yimu Wang, Xiangru Jian, and Bo Xue. 2023b. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348.
- Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser-Nam Lim. 2023. Test-time distribution normalization for contrastively learned vision-language models. In *Advances in Neural Information Processing Systems*.

Appendix

A Baselines

Our main experiments compare GLOBE with three training-free calibration baselines: Distribution Normalization (DN), DBNorm, and Nearest Neighbor Normalization (NNN). We additionally compare with CSLS and DBSN on CLIP and SigLIP.

All methods operate on frozen image-text embeddings and do not update model parameters. Let q denote a query, r denote a retrieval candidate, and $s(q, r)$ denote the original contrastive similarity score, computed as the dot product or cosine similarity between normalized embeddings.

A.1 Distribution Normalization(DN)

Distribution Normalization (DN) approximates test-time softmax normalization by shifting embeddings using reference distribution statistics. In practice, DN can be viewed as normalizing candidate scores with respect to their reference-query score distribution. Given a reference query set $\hat{Q}$, we compute the mean and standard deviation of scores assigned to candidate r :

$$\mu_r = \frac{1}{|\hat{Q}|} \sum_{\hat{q} \in \hat{Q}} s(\hat{q}, r), \quad (5)$$

$$\sigma_r = \sqrt{\frac{1}{|\hat{Q}|} \sum_{\hat{q} \in \hat{Q}} (s(\hat{q}, r) - \mu_r)^2}. \quad (6)$$

The DN score is then

$$s_{\text{DN}}(q, r) = \frac{s(q, r) - \mu_r}{\sigma_r + \epsilon}, \quad (7)$$

where ϵ is a small constant for numerical stability. This correction penalizes candidates that receive high scores from many reference queries, while keeping the method efficient at inference time. In the original formulation, DN is described as a first-order approximation that normalizes embeddings by subtracting the mean reference embedding.

A.2 DBNorm

DBNorm mitigates hubness by normalizing retrieval scores with respect to reference query and gallery banks. We use the DualIS scoring function following prior work. Given a reference query database $\hat{Q}$ and a reference candidate database $\hat{R}$, DBNorm computes two normalized scores:

$$\hat{s}_{q,r}^{\hat{R}} = \frac{\exp(\beta_1 s(q, r))}{\sum_{\hat{r} \in \hat{R}} \exp(\beta_1 s(q, \hat{r}))} \quad (8)$$

$$\hat{s}_{q,r}^{\hat{Q}} = \frac{\exp(\beta_2 s(q, r))}{\sum_{\hat{q} \in \hat{Q}} \exp(\beta_2 s(\hat{q}, r))} \quad (9)$$

The final DBNorm score is

$$s_{\text{DBNorm}}(q, r) = \hat{s}_{q,r}^{\hat{R}} \cdot \hat{s}_{q,r}^{\hat{Q}} \quad (10)$$

Here, β_1 and β_2 are temperature-like hyperparameters controlling the sharpness of the softmax normalization. The $\hat{R}$ -side term calibrates query-to-gallery competition, while the $\hat{Q}$ -side term penalizes candidates that are frequently preferred by many reference queries. This formulation follows the DualIS score used in DBNorm.

A.3 Nearest Neighbor Normalization(NNN)

Nearest Neighbor Normalization (NNN) estimates a candidate-specific bias using only the nearest reference queries. Let D be a reference query database. For each candidate r , NNN selects the k reference queries with the highest similarity to r :

$$D_{\text{top } k}(r) = \arg \text{top } k_{q' \in D} s(q', r). \quad (11)$$

The candidate bias is computed as

$$b(r) = \alpha \cdot \frac{1}{k} \sum_{q' \in D_{\text{top } k}(r)} s(q', r), \quad (12)$$

where k is the number of nearest reference queries and α controls the strength of the bias correction. The normalized score is

$$s_{\text{NNN}}(q, r) = s(q, r) - b(r). \quad (13)$$

NNN therefore subtracts an additive candidate-level penalty from the original score. Compared with DBNorm, NNN only requires the top- k reference queries for each candidate and thus avoids full softmax normalization over the entire reference database.

A.4 Additional Neighborhood-Based and Recent Baselines

CSLS. Cross-domain similarity local scaling (CSLS) uses local neighborhood statistics to rescale cross-modal similarities. Rather than applying a candidate-level global penalty, CSLS accounts for the average similarity of each representation to its local cross-domain neighborhood, thereby reducing the preference for locally popular candidates.

Model	Setting	Orig.	CSLS	DBSN	GLOBE
CLIP	COCO-I2T	49.14	53.50	52.82	53.86
CLIP	COCO-T2I	30.24	35.99	37.70	37.92
CLIP	F30K-I2T	77.70	80.60	80.10	81.00
CLIP	F30K-T2I	56.52	63.46	63.78	64.38
SigLIP	COCO-I2T	65.72	70.28	70.16	70.37
SigLIP	COCO-T2I	47.78	51.50	50.85	51.50
SigLIP	F30K-I2T	88.90	91.90	89.70	92.20
SigLIP	F30K-T2I	73.94	75.98	75.78	76.68

Table 5: Additional R@1 comparison with CSLS and DBSN on CLIP and SigLIP. F30K denotes Flickr30K. Best results are bolded.

Relation to query expansion and k-reciprocal re-ranking.

Query-expansion methods refine a retrieval representation using its high-ranked neighbors, while k-reciprocal re-ranking exploits reciprocal nearest-neighbor relationships to refine pairwise distances or rankings. GLOBE uses neighborhood structure for a different purpose. Its graph term does not replace the original similarity with a neighborhood-based representation or reciprocal distance; instead, it acts as an additive semantic recovery signal on top of explicit bias calibration. Therefore, neighborhood consistency in GLOBE is used to determine whether a hub-like match is semantically supported, rather than serving as a standalone re-ranking criterion.

NeighborRetr versus GLOBE. NeighborRetr and GLOBE share the high-level observation that hubness is not uniformly harmful. However, they address this issue through different formulations. NeighborRetr adjusts retrieval behavior by balancing hub centrality through neighbor relationships. In contrast, GLOBE is a training-free post-hoc calibration method operating directly on frozen embeddings and similarity scores. It explicitly decomposes the calibrated score into bias suppression and semantic recovery: global and local terms estimate candidate attraction and repeated local over-density, while graph evidence evaluates whether a query-candidate match remains semantically supported. Thus, the two approaches share a motivation but differ in their calibration formulation and the role assigned to neighborhood information.

Additional comparison with CSLS and DBSN. GLOBE achieves the best Recall@1 in seven of the eight evaluated settings and ties with CSLS on SigLIP COCO T2I (table 5). These results suggest that the gains of GLOBE are maintained when compared with both neighborhood-based similarity

scaling and the recent DBSN method.

B Hyperparameter selection

Reference-set construction. Unless otherwise specified, the reference bank is constructed by randomly sampling 20% of the corresponding training split using random seed 42. All calibration methods operate on frozen image-text embeddings and similarity scores. Once the reference subset and hyperparameters are fixed, the nearest-neighbor operations and subsequent score calibration are deterministic.

B.1 DBNorm

For DBNorm, we use the DualIS scoring function and tune its two temperature parameters β_1 and β_2 over the fixed grid

$$\beta_1, \beta_2 \in \{0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 5, 10, 25, 50, 100\}.$$

The best configuration is selected according to Recall@1 for each model, target dataset, and retrieval direction. For T2I retrieval, the reference query and candidate banks are training captions and training images, respectively; for I2T retrieval, they are training images and training captions. We compute DualIS scores in log-space for numerical stability.

B.2 Nearest Neighbor Normalization(NNN)

For the NNN baseline, we do not perform additional hyperparameter search in our experiments. Instead, we directly adopt the best hyperparameters reported in the NNN paper (Chowdhury et al., 2024). Specifically, for each model, retrieval direction, target dataset, and reference dataset, we use the corresponding optimal (α, k) selected by the original NNN study. This protocol avoids re-tuning NNN on our experimental setup and ensures that GLOBE is compared against the strongest reported NNN configuration for each setting. The NNN hyperparameters used in our experiments are summarized in Table 6.

B.3 GLOBE

For GLOBE, we perform a compact grid search over the main hyperparameters in the global-local bias term $\mathcal{B}(q, r)$ and the graph evidence term $\mathcal{G}(q, r)$. The best configuration is selected according to Recall@1 for each model, target dataset, reference dataset, and retrieval direction. GLOBE

Table 6: NNN hyperparameters used in our experiments. Each entry reports the pair (α, k) . F/F, F/C, C/F, and C/C denote target/reference pairs: Flickr/Flickr, Flickr/COCO, COCO/Flickr, and COCO/COCO, respectively. For each model, retrieval direction, target dataset, and reference dataset, we use the best hyperparameters reported in the NNN paper (Chowdhury et al., 2024) without additional search.

Model	T2I / Image Retrieval				I2T / Text Retrieval			
	F/F	F/C	C/F	C/C	F/F	F/C	C/F	C/C
CLIP ViT-B/32	(0.75, 128)	(0.75, 16)	(0.5, 8)	(0.75, 256)	(0.75, 16)	(0.5, 2)	(0.5, 8)	(0.75, 128)
SigLIP ViT-B/16	(0.75, 128)	(0.5, 128)	(0.5, 16)	(0.75, 128)	(0.5, 64)	(0.75, 256)	(0.25, 32)	(0.75, 128)
BLIP ft. Flickr	(0.5, 32)	(0.25, 4)	(0.5, 64)	(0.75, 16)	(0.75, 512)	(0.75, 64)	(0.75, 32)	(0.75, 64)
BLIP ft. COCO	(0.5, 16)	(0.25, 4)	(0.25, 4)	(0.75, 64)	(1, 512)	(0.75, 16)	(0.5, 16)	(0.75, 32)
ALBEF ft. Flickr	(0.75, 32)	(0.25, 16)	(0.5, 4)	(0.75, 256)	(0.25, 512)	(0.25, 64)	(0.5, 16)	(0.75, 128)
ALBEF ft. COCO	(0.75, 32)	(0.5, 16)	(0.25, 8)	(0.75, 128)	(0.75, 32)	(0.5, 64)	(0.25, 8)	(0.75, 32)
BEiT-3	(0.75, 32)	(0.5, 64)	(0.25, 4)	(0.75, 128)	(0.75, 64)	(0.5, 32)	(0.5, 32)	(0.75, 256)
BEiT-3 ft. Flickr	(0.25, 8)	(0.25, 64)	(0.25, 4)	(0.75, 256)	(1, 32)	(0.75, 4)	(0.25, 16)	(0.75, 256)
BEiT-3 ft. COCO	(0.75, 16)	(0.25, 2)	(0.25, 32)	(0.25, 128)	(0.5, 32)	(0.5, 4)	(0.25, 4)	(0.5, 8)
BEiT-3 Large	(0.5, 256)	(0.5, 32)	(0.25, 32)	(0.75, 128)	(0.5, 64)	(0.5, 512)	(0.5, 16)	(0.75, 512)
BEiT-3 Large ft. Flickr	(0.5, 16)	(0.25, 1)	(0.25, 16)	(0.75, 512)	(0.5, 64)	(0.75, 16)	(0.5, 16)	(0.75, 128)
BEiT-3 Large ft. COCO	(0.5, 8)	(0.25, 128)	(0.25, 8)	(0.5, 64)	(0.5, 64)	(0.75, 32)	(0.25, 64)	(0.5, 16)

Table 7: GLOBE search space. We fix $m = 150$, $\lambda = 0.02$, $k_q = 3$, and $k_r = 10$.

Term	Param.	Values
Global bias	μ	$\{0.1, 0.25, 0.375\}$
Global bias	K_g	$\{16, 32, 64, 128\}$
Query graph	η_q	$\{0, 0.05, 0.1, 0.15\}$
Cand. graph	η_r	$\{0.05, 0.1, 0.15, 0.2\}$

uses the same predefined search space across models for a consistent training-free comparison. In all main experiments, we fix the local candidate size $m = 150$, the local density weight $\lambda = 0.02$, the query-side graph neighborhood size $k_q = 3$, and the candidate-side graph neighborhood size $k_r = 10$. We search four core hyperparameters: the global bias strength μ , the global bias neighborhood size K_g , the query-side graph evidence weight η_q , and the candidate-side graph evidence weight η_r .

The effective search grid contains

$$3 \times 4 \times 4 \times 4 = 192$$

configurations per target/reference setting, corresponding to μ , K_g , η_q , and η_r . Tables 8 and 9 report the selected GLOBE hyperparameters for the best Recall@1 result of each model in the I2T and T2I main tables, respectively.

Although the main results use target-specific configurations for peak performance, Sections 3.1 and B.4 show that most gains are retained without target-specific hyperparameter re-selection.

B.4 Hyperparameter Transferability

Cross-model transfer. We further evaluate whether a fixed hyperparameter configuration selected on CLIP can transfer to other retrieval backbones. We directly apply the same CLIP-selected configuration to SigLIP and BEiT3-base without model-specific hyperparameter re-selection.

As shown in Table 10, the transferred configuration improves all four original retrieval results and remains within 0.26–0.73 R@1 points of target-specific GLOBE, indicating that GLOBE’s hyperparameters are not tightly coupled to a particular backbone.

C Full Ablation Results

Appendix C reports the full ablation results of GLOBE on CLIP and SigLIP. We evaluate the contribution of each component by comparing the original retrieval score with variants using only the global bias term, only the local density term, only the graph evidence term, and their pairwise combinations. The final row in each setting corresponds to the full GLOBE model, which combines global-local bias calibration with graph-based semantic evidence. We report Recall@1, Recall@5, Recall@10, and the absolute Recall@1 improvement over the original model.

Tables 11 and 12 show that the global bias term provides the largest individual gain in most settings, confirming that attraction bias is a major source of retrieval errors. The local and graph terms alone are less consistently effective, but they improve performance when combined with global calibration. In particular, the full GLOBE model achieves the best

Table 8: Selected GLOBE hyperparameters for the best Recall@1 result of each model in the I2T main table. For each model, we select the target/reference setting with the highest GLOBE Recall@1. Here, K_g denotes the number of reference queries used for global candidate-bias estimation.

Model	Best setting	Best R@1	μ	K_g	η_q	η_r
CLIP ViT-B/32	Flickr/Flickr	81.00	0.375	16	0.10	0.10
SigLIP ViT-B/16	Flickr/Flickr	92.20	0.25	64	0.0	0.20
BLIP pretrained	Flickr/Flickr	90.20	0.375	64	0.15	0.20
BLIP ft. Flickr	Flickr/Flickr	95.70	0.25	64	0.10	0.20
BLIP ft. COCO	Flickr/COCO	94.00	0.25	16	0.0	0.10
ALBEF ft. COCO	Flickr/Flickr	90.20	0.375	128	0.15	0.15
ALBEF ft. Flickr	Flickr/Flickr	91.60	0.10	16	0.15	0.05
BEiT3-base-its	Flickr/Flickr	92.90	0.25	64	0.10	0.10
BEiT3 Large	Flickr/Flickr	87.50	0.25	32	0.0	0.15
BEiT3 Large ft. COCO	Flickr/Flickr	89.80	0.10	16	0.10	0.15
BEiT3 Large ft. Flickr	Flickr/Flickr	90.30	0.25	128	0.15	0.10

Table 9: Selected GLOBE hyperparameters for the best Recall@1 result of each model in the T2I main table. For each model, we select the target/reference setting with the highest GLOBE Recall@1. Here, K_g denotes the number of reference queries used for global candidate-bias estimation.

Model	Best setting	Best R@1	μ	K_g	η_q	η_r
CLIP ViT-B/32	Flickr/Flickr	64.38	0.375	64	0.15	0.10
SigLIP ViT-B/16	Flickr/Flickr	76.68	0.375	64	0.15	0.10
BLIP pretrained	Flickr/Flickr	72.66	0.375	64	0.00	0.20
BLIP ft. Flickr	Flickr/Flickr	95.70	0.25	64	0.15	0.10
BLIP ft. COCO	Flickr/COCO	94.00	0.25	16	0.00	0.10
ALBEF ft. COCO	Flickr/Flickr	75.66	0.375	64	0.15	0.10
ALBEF ft. Flickr	Flickr/Flickr	80.08	0.10	16	0.00	0.15
BEiT3-base-its	Flickr/Flickr	77.12	0.25	64	0.10	0.10
BEiT3 ft. Flickr	Flickr/Flickr	85.50	0.25	64	0.10	0.15
BEiT3 ft. COCO	Flickr/Flickr	83.52	0.375	32	0.10	0.10
BEiT3 Large	Flickr/Flickr	72.12	0.25	32	0.15	0.15
BEiT3 Large ft. COCO	Flickr/Flickr	77.58	0.25	32	0.10	0.15
BEiT3 Large ft. Flickr	Flickr/COCO	78.20	0.10	64	0.15	0.05

Model	Dir.	Orig.	NNN	Target	Transfer
SigLIP	T2I	47.78	51.05	51.50	50.77
BEiT3-B	T2I	47.52	50.25	50.79	50.21
SigLIP	I2T	65.72	69.96	70.37	70.11
BEiT3-B	I2T	59.74	69.34	69.84	69.47

Table 10: Cross-model transfer of a fixed CLIP-selected GLOBE configuration on COCO. “Target” denotes target-specific hyperparameter selection, while “Transfer” uses the fixed CLIP-selected configuration without model-specific re-selection.

or near-best Recall@1 across both Flickr30k and COCO, indicating that graph-based evidence complements bias suppression by recovering semantically supported pairs rather than simply penalizing all hub-like candidates.

D Full I2T Results

Table 13,14,15,16,17,18,19,20,21,22,23,24,25 reports the full image-to-text retrieval results on Flickr30k and MS-COCO. In addition to the Re-

call@1 results summarized in the main paper, we provide Recall@5 and Recall@10 to evaluate retrieval quality beyond the top-ranked prediction. We include all evaluated backbones: CLIP ViT-B/32, SigLIP ViT-B/16, BLIP pretrained, BLIP fine-tuned on Flickr30k, BLIP fine-tuned on COCO, ALBEF fine-tuned on Flickr30k, ALBEF fine-tuned on COCO, BEiT3-base-its, BEiT3 fine-tuned on Flickr30k, BEiT3 fine-tuned on COCO, BEiT3-large-its, BEiT3 Large fine-tuned on Flickr30k, and BEiT3 Large fine-tuned on COCO. For each model, we compare the original retrieval score, DN, DBNorm, NNN, and GLOBE under the same evaluation protocol as the main experiments. The results show that GLOBE generally improves or preserves Recall@5 and Recall@10 while producing the clearest gains at Recall@1, indicating that the proposed graph-based calibration mainly improves the precision of top-ranked predictions without degrading broader retrieval quality.

Table 11: Full Ablation Result of GLOBE on CLIP.

Backbone	Target	Ref	Method	R@1	R@5	R@10	$\Delta R@1$
CLIP	Flickr	Flickr	Original	56.52	81.38	87.72	0.00
CLIP	Flickr	Flickr	Global only	62.94	86.24	92.02	6.42
CLIP	Flickr	Flickr	Local only	59.92	83.42	89.66	3.40
CLIP	Flickr	Flickr	Graph only	56.68	81.36	88.32	0.16
CLIP	Flickr	Flickr	Global + Local	62.98	86.34	92.48	6.46
CLIP	Flickr	Flickr	Global + Graph	63.62	86.36	92.14	7.10
CLIP	Flickr	Flickr	Local + Graph	59.92	83.38	89.80	3.40
CLIP	Flickr	Flickr	Full GLOBE	64.38	87.06	92.76	7.86
CLIP	Flickr	COCO	Original	56.52	81.38	87.72	0.00
CLIP	Flickr	COCO	Global only	61.76	84.40	90.90	5.24
CLIP	Flickr	COCO	Local only	58.74	82.34	88.80	2.22
CLIP	Flickr	COCO	Graph only	56.68	81.36	88.32	0.16
CLIP	Flickr	COCO	Global + Local	60.86	84.12	90.60	4.34
CLIP	Flickr	COCO	Global + Graph	62.32	85.16	91.20	5.80
CLIP	Flickr	COCO	Local + Graph	58.44	82.38	88.86	1.92
CLIP	Flickr	COCO	Full GLOBE	62.32	85.26	91.10	5.80
CLIP	COCO	Flickr	Original	30.24	54.94	65.72	0.00
CLIP	COCO	Flickr	Global only	32.37	57.37	68.33	2.13
CLIP	COCO	Flickr	Local only	30.24	54.94	65.72	0.00
CLIP	COCO	Flickr	Graph only	29.70	54.65	65.35	-0.54
CLIP	COCO	Flickr	Global + Local	32.37	57.37	68.33	2.13
CLIP	COCO	Flickr	Global + Graph	32.34	57.41	68.27	2.10
CLIP	COCO	Flickr	Local + Graph	29.70	54.65	65.35	-0.54
CLIP	COCO	Flickr	Full GLOBE	32.37	57.41	68.27	2.13
CLIP	COCO	COCO	Original	30.24	54.94	65.72	0.00
CLIP	COCO	COCO	Global only	36.51	62.03	72.17	6.27
CLIP	COCO	COCO	Local only	35.13	60.54	70.94	4.89
CLIP	COCO	COCO	Graph only	29.70	54.65	65.35	-0.54
CLIP	COCO	COCO	Global + Local	36.49	62.03	72.32	6.25
CLIP	COCO	COCO	Global + Graph	36.05	61.41	71.48	5.81
CLIP	COCO	COCO	Local + Graph	34.36	59.78	70.22	4.12
CLIP	COCO	COCO	Full GLOBE	37.92	63.21	73.33	7.68

E Full T2I Results

Table 26,27,28,29,30,31,32,33,34,35,36,37,38 reports the full text-to-image retrieval results on Flickr30k and MS-COCO. Beyond the Recall@1 results presented in the main paper, we include Recall@5 and Recall@10 to examine whether GLOBE improves retrieval performance across broader ranking depths. We evaluate all backbones used in the main experiments, including CLIP ViT-B/32, SigLIP ViT-B/16, BLIP pretrained, BLIP fine-tuned on Flickr30k, BLIP fine-tuned on COCO, ALBEF fine-tuned on Flickr30k, ALBEF fine-tuned on COCO, BEiT3-base-its, BEiT3 fine-tuned on Flickr30k, BEiT3 fine-tuned on

COCO, BEiT3-large-its, BEiT3 Large fine-tuned on Flickr30k, and BEiT3 Large fine-tuned on COCO. For each model, we compare the original retrieval scores with DN, DBNorm, NNN, and GLOBE under the same reference-database protocol as the main experiments. The results show that GLOBE consistently improves top-ranked retrieval performance and generally maintains or improves Recall@5 and Recall@10, suggesting that the proposed calibration does not simply reorder the first prediction but also preserves the quality of the broader candidate list.

Table 12: Full Ablation Result of GLOBE on SigLIP.

Backbone	Target	Ref	Method	R@1	R@5	R@10	$\Delta R@1$
SigLIP	Flickr	Flickr	Original	73.94	91.88	95.46	0.00
SigLIP	Flickr	Flickr	Global only	75.50	92.96	95.94	1.56
SigLIP	Flickr	Flickr	Local only	73.94	91.88	95.46	0.00
SigLIP	Flickr	Flickr	Graph only	74.10	91.98	95.62	0.16
SigLIP	Flickr	Flickr	Global + Local	75.50	92.96	95.94	1.56
SigLIP	Flickr	Flickr	Global + Graph	76.68	93.44	96.14	2.74
SigLIP	Flickr	Flickr	Local + Graph	74.10	91.98	95.62	0.16
SigLIP	Flickr	Flickr	Full GLOBE	76.68	93.44	96.14	2.74
SigLIP	Flickr	COCO	Original	73.94	91.88	95.46	0.00
SigLIP	Flickr	COCO	Global only	74.10	91.94	95.40	0.16
SigLIP	Flickr	COCO	Local only	73.94	91.88	95.46	0.00
SigLIP	Flickr	COCO	Graph only	74.10	91.98	95.62	0.16
SigLIP	Flickr	COCO	Global + Local	74.10	91.94	95.40	0.16
SigLIP	Flickr	COCO	Global + Graph	75.24	92.78	95.70	1.30
SigLIP	Flickr	COCO	Local + Graph	74.10	91.98	95.62	0.16
SigLIP	Flickr	COCO	Full GLOBE	75.24	92.78	95.70	1.30
SigLIP	COCO	Flickr	Original	47.78	72.43	81.03	0.00
SigLIP	COCO	Flickr	Global only	47.93	72.17	80.82	0.15
SigLIP	COCO	Flickr	Local only	47.78	72.43	81.03	0.00
SigLIP	COCO	Flickr	Graph only	47.29	72.10	80.61	-0.49
SigLIP	COCO	Flickr	Global + Local	47.93	72.17	80.82	0.15
SigLIP	COCO	Flickr	Global + Graph	48.34	72.57	81.26	0.56
SigLIP	COCO	Flickr	Local + Graph	47.29	72.10	80.61	-0.49
SigLIP	COCO	Flickr	Full GLOBE	48.34	72.57	81.26	0.56
SigLIP	COCO	COCO	Original	47.78	72.43	81.03	0.00
SigLIP	COCO	COCO	Global only	51.05	75.09	83.19	3.27
SigLIP	COCO	COCO	Local only	48.92	73.43	81.88	1.15
SigLIP	COCO	COCO	Graph only	47.29	72.10	80.61	-0.49
SigLIP	COCO	COCO	Global + Local	51.03	75.03	83.16	3.25
SigLIP	COCO	COCO	Global + Graph	51.11	75.12	83.10	3.33
SigLIP	COCO	COCO	Local + Graph	48.31	72.97	81.27	0.53
SigLIP	COCO	COCO	Full GLOBE	51.50	75.32	83.42	3.72

F IRB-COCO-500 Dataset

We construct IRB-COCO-500 to evaluate retrieval robustness in mixed real-and-synthetic galleries. The benchmark is built from MS-COCO and contains 500 query instances. For each selected COCO image-caption pair, we keep the original COCO image as the ground-truth candidate and generate one near-positive synthetic distractor from the corresponding caption using Stable Diffusion XL base 1.0 (Podell et al., 2023) through the Diffusers StableDiffusionXLPipeline. Each synthetic image is generated at 1024×1024 resolution with 30 denoising steps, guidance scale 7.0, and a deterministic per-sample seed $42 + \text{pair_id}$.

Evaluation Metrics. For each query q_i , let r_i^{real} denote the original COCO ground-truth image and r_i^{gen} denote the corresponding AI-generated distractor. Let $\pi_i(1)$ be the top-ranked image returned by a retrieval method for query q_i . We define RealTop1Rate as the fraction of queries for which the real COCO image is ranked first:

$$\text{RealTop1Rate} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\pi_i(1) = r_i^{\text{real}}]. \quad (14)$$

We define GenPreferenceRate as the fraction of queries for which the generated distractor is ranked above the real COCO image:

Table 13: Full I2T results of different algorithms on CLIP.

Model	Target	Ref.	Method	R@1	R@5	R@10
CLIP	Flickr	-	Original	77.70	94.60	97.10
CLIP	Flickr	Flickr	DN	73.80	91.50	95.40
CLIP	Flickr	Flickr	DBNorm	78.10	94.10	96.80
CLIP	Flickr	Flickr	NNN	80.10	94.30	97.10
CLIP	Flickr	Flickr	GLOBE	81.00	94.40	97.20
CLIP	Flickr	COCO	NNN	78.30	94.80	97.00
CLIP	Flickr	COCO	GLOBE	78.20	94.40	97.00
CLIP	COCO	-	Original	49.14	73.34	82.34
CLIP	COCO	COCO	DN	45.06	71.26	80.42
CLIP	COCO	COCO	DBNorm	49.30	74.44	83.02
CLIP	COCO	Flickr	NNN	49.72	73.70	82.52
CLIP	COCO	Flickr	GLOBE	50.08	73.64	82.02
CLIP	COCO	COCO	NNN	53.10	75.92	83.64
CLIP	COCO	COCO	GLOBE	53.86	76.94	84.14

Table 15: Full I2T results of different algorithms on BLIP pretrained.

Model	Target	Ref.	Method	R@1	R@5	R@10
BLIP pre.	Flickr	-	Original	83.00	96.30	98.60
BLIP pre.	Flickr	Flickr	DN	86.90	96.90	99.00
BLIP pre.	Flickr	Flickr	DBNorm	88.40	97.30	99.00
BLIP pre.	Flickr	Flickr	NNN	89.10	97.90	99.20
BLIP pre.	Flickr	Flickr	GLOBE	90.20	97.90	98.90
BLIP pre.	Flickr	COCO	NNN	85.00	97.00	98.90
BLIP pre.	Flickr	COCO	GLOBE	86.20	97.00	98.80
BLIP pre.	COCO	-	Original	61.88	85.44	92.08
BLIP pre.	COCO	Flickr	DN	67.24	88.90	94.58
BLIP pre.	COCO	Flickr	DBNorm	68.74	89.52	94.50
BLIP pre.	COCO	Flickr	NNN	65.06	87.60	93.40
BLIP pre.	COCO	Flickr	GLOBE	65.36	87.58	93.10
BLIP pre.	COCO	COCO	NNN	71.68	90.68	95.30
BLIP pre.	COCO	COCO	GLOBE	72.18	90.96	95.08

$$\text{GenPreferenceRate} = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left[\text{rank}_i(r_i^{\text{gen}}) < \text{rank}_i(r_i^{\text{real}}) \right]. \quad (15)$$

Here, $N = 500$ is the number of query instances in IRB-COCO-500, and $\text{rank}_i(r)$ denotes the rank position of candidate r for query q_i , where a smaller value indicates a higher rank.

The final gallery contains 500 real COCO images and 500 synthetic images, resulting in 1,000 candidate images. Since the generated images preserve the coarse semantics of the query but do not correspond to the original ground-truth image, they serve as challenging near-positive distractors rather than random negatives. We use IRB-COCO-500

Table 14: Full I2T results of different algorithms on SigLIP.

Model	Target	Ref.	Method	R@1	R@5	R@10
SigLIP	Flickr	-	Original	88.90	97.80	99.50
SigLIP	Flickr	Flickr	DN	84.70	97.00	98.30
SigLIP	Flickr	Flickr	DBNorm	90.10	98.30	99.00
SigLIP	Flickr	Flickr	NNN	91.10	98.30	99.40
SigLIP	Flickr	Flickr	GLOBE	92.20	98.50	99.30
SigLIP	Flickr	COCO	NNN	90.90	98.00	99.40
SigLIP	Flickr	COCO	GLOBE	91.60	98.20	99.20
SigLIP	COCO	-	Original	65.72	85.44	91.22
SigLIP	COCO	COCO	DN	60.10	82.62	89.54
SigLIP	COCO	COCO	DBNorm	65.22	85.58	91.78
SigLIP	COCO	Flickr	NNN	66.62	86.74	91.68
SigLIP	COCO	Flickr	GLOBE	67.26	86.88	91.88
SigLIP	COCO	COCO	NNN	69.96	88.18	92.94
SigLIP	COCO	COCO	GLOBE	70.36	88.32	93.08

Table 16: Full I2T results of different algorithms on BLIP ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
BLIP ft. F	Flickr	-	Original	93.40	99.40	99.80
BLIP ft. F	Flickr	Flickr	DN	93.60	99.50	99.90
BLIP ft. F	Flickr	Flickr	DBNorm	95.00	99.80	100.00
BLIP ft. F	Flickr	Flickr	NNN	95.80	99.80	100.00
BLIP ft. F	Flickr	Flickr	GLOBE	95.70	99.50	99.90
BLIP ft. F	Flickr	COCO	NNN	95.40	99.50	99.90
BLIP ft. F	Flickr	COCO	GLOBE	95.40	99.60	100.00
BLIP ft. F	COCO	-	Original	69.90	90.50	94.88
BLIP ft. F	COCO	Flickr	DN	71.84	91.08	95.64
BLIP ft. F	COCO	Flickr	DBNorm	73.10	91.88	95.84
BLIP ft. F	COCO	Flickr	NNN	73.14	91.22	95.36
BLIP ft. F	COCO	Flickr	GLOBE	73.06	91.00	95.28
BLIP ft. F	COCO	COCO	NNN	76.40	92.86	96.12
BLIP ft. F	COCO	COCO	GLOBE	76.80	92.94	96.18

to test whether retrieval methods can preserve the true image-text match under synthetic visual interference.

Figure 4 shows representative examples from IRB-COCO-500. For each query, we visualize the original COCO ground-truth image and the corresponding AI-generated distractor. The generated images are often semantically close to the query and share major objects, scenes, or actions with the ground truth, which makes them strong distractors for image-text retrieval. At the same time, they may differ in fine-grained visual details, object identity, spatial layout, or photographic realism.

For reproducibility, we will release the dataset metadata, prompts, generation seeds, generation script, and evaluation code.

Table 17: Full I2T results of different algorithms on BLIP ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
BLIP ft. C Flickr	-	-	Original	90.80	98.70	99.60
BLIP ft. C Flickr	Flickr	Flickr	DN	90.80	99.20	99.70
BLIP ft. C Flickr	Flickr	Flickr	DBNorm	92.10	99.30	99.80
BLIP ft. C Flickr	Flickr	Flickr	NNN	92.80	99.40	99.90
BLIP ft. C Flickr	Flickr	Flickr	GLOBE	93.00	99.30	99.80
BLIP ft. C Flickr	COCO	COCO	NNN	93.80	99.00	99.90
BLIP ft. C Flickr	COCO	COCO	GLOBE	94.00	99.00	99.80
BLIP ft. C COCO	-	-	Original	75.76	93.80	96.62
BLIP ft. C COCO	Flickr	Flickr	DN	77.34	93.68	96.88
BLIP ft. C COCO	Flickr	Flickr	DBNorm	78.00	94.02	96.88
BLIP ft. C COCO	Flickr	Flickr	NNN	77.50	93.58	96.70
BLIP ft. C COCO	Flickr	Flickr	GLOBE	77.44	93.68	96.66
BLIP ft. C COCO	COCO	COCO	NNN	79.10	94.00	97.02
BLIP ft. C COCO	COCO	COCO	GLOBE	79.06	94.20	97.10

Table 19: Full I2T results of different algorithms on ALBEF ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
ALBEF ft. F Flickr	-	-	Original	90.60	98.60	99.90
ALBEF ft. F Flickr	Flickr	Flickr	DN	89.20	98.70	99.40
ALBEF ft. F Flickr	Flickr	Flickr	DBNorm	91.40	99.00	99.80
ALBEF ft. F Flickr	Flickr	Flickr	NNN	90.80	98.70	99.80
ALBEF ft. F Flickr	Flickr	Flickr	GLOBE	91.60	98.80	99.70
ALBEF ft. F Flickr	COCO	COCO	NNN	90.30	99.20	99.90
ALBEF ft. F Flickr	COCO	COCO	GLOBE	91.10	99.10	99.80
ALBEF ft. F COCO	-	-	Original	66.32	88.50	94.16
ALBEF ft. F COCO	Flickr	Flickr	DN	54.80	80.44	89.18
ALBEF ft. F COCO	Flickr	Flickr	DBNorm	67.38	88.84	94.56
ALBEF ft. F COCO	Flickr	Flickr	NNN	67.54	88.86	94.60
ALBEF ft. F COCO	Flickr	Flickr	GLOBE	67.34	88.90	94.38
ALBEF ft. F COCO	COCO	COCO	NNN	70.64	90.92	95.42
ALBEF ft. F COCO	COCO	COCO	GLOBE	71.66	90.96	95.54

G Runtime and Computational Cost

GLOBE is a training-free reranking method and does not require model fine-tuning. We evaluate runtime using cached image and text embeddings and exclude the cost of feature extraction, since the reranking stage operates only on frozen embeddings. Here, we focus on the post-hoc GLOBE calibration module rather than separately reporting the base score-matrix computation. GLOBE itself introduces no trainable parameters; the model sizes are therefore those of the frozen off-the-shelf image-text retrievers evaluated in our experiments.

Several components of GLOBE require nearest-neighbor search. First, the global candidate-bias term retrieves the top- k_b reference queries for each retrieval candidate. Second, the local density term estimates neighborhood statistics within the top-ranked candidate set. Third, the query-side

Table 18: Full I2T results of different algorithms on ALBEF ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
ALBEF ft. C Flickr	-	-	Original	87.50	98.20	99.50
ALBEF ft. C Flickr	Flickr	Flickr	DN	84.90	97.70	99.00
ALBEF ft. C Flickr	Flickr	Flickr	DBNorm	87.90	98.70	99.60
ALBEF ft. C Flickr	Flickr	Flickr	NNN	89.70	98.60	99.50
ALBEF ft. C Flickr	Flickr	Flickr	GLOBE	89.90	98.60	99.30
ALBEF ft. C Flickr	COCO	COCO	NNN	87.20	97.60	99.30
ALBEF ft. C Flickr	COCO	COCO	GLOBE	87.10	97.90	99.10
ALBEF ft. C COCO	-	-	Original	71.32	91.46	95.94
ALBEF ft. C COCO	Flickr	Flickr	DN	70.12	90.94	95.68
ALBEF ft. C COCO	Flickr	Flickr	DBNorm	72.36	91.62	96.20
ALBEF ft. C COCO	Flickr	Flickr	NNN	71.76	91.62	96.20
ALBEF ft. C COCO	Flickr	Flickr	GLOBE	71.68	91.50	95.94
ALBEF ft. C COCO	COCO	COCO	NNN	74.10	92.34	96.16
ALBEF ft. C COCO	COCO	COCO	GLOBE	74.08	92.32	96.24

Table 20: Full I2T results of different algorithms on BEiT3-base-itsc.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-B ITC Flickr	-	-	Original	90.00	98.70	99.50
BEiT3-B ITC Flickr	Flickr	Flickr	DN	87.10	97.70	98.90
BEiT3-B ITC Flickr	Flickr	Flickr	DBNorm	90.20	98.60	99.70
BEiT3-B ITC Flickr	Flickr	Flickr	NNN	92.10	98.60	99.60
BEiT3-B ITC Flickr	Flickr	Flickr	GLOBE	92.90	98.70	99.70
BEiT3-B ITC Flickr	COCO	COCO	NNN	90.70	98.60	99.40
BEiT3-B ITC Flickr	COCO	COCO	GLOBE	90.90	98.60	99.40
BEiT3-B ITC COCO	-	-	Original	59.74	83.38	90.36
BEiT3-B ITC COCO	COCO	COCO	DN	60.54	83.14	89.94
BEiT3-B ITC COCO	COCO	COCO	DBNorm	64.82	86.70	92.08
BEiT3-B ITC COCO	Flickr	Flickr	NNN	64.64	86.40	92.00
BEiT3-B ITC COCO	Flickr	Flickr	GLOBE	65.08	86.22	92.04
BEiT3-B ITC COCO	COCO	COCO	NNN	69.34	87.98	93.04
BEiT3-B ITC COCO	COCO	COCO	GLOBE	69.84	87.98	93.20

and candidate-side graph evidence terms require nearest-neighbor search over query embeddings and candidate embeddings, respectively. Without acceleration, these steps require exhaustive pairwise similarity search over the corresponding embedding banks, which becomes expensive on COCO-scale retrieval. We therefore use Faiss to build vector indices over the reference-query, query, and candidate embeddings, replacing exhaustive all-pair neighbor search with indexed top- k retrieval. After these sparse neighborhoods are obtained, GLOBE only performs lightweight score aggregation for bias calibration and graph-evidence computation.

The reported runtime should not be interpreted as a strict linear scaling law with gallery size. Index construction can be performed offline, while online GLOBE operates on fixed-size retrieved neighborhoods. For galleries beyond 100K can-

Table 21: Full I2T results of different algorithms on BEiT3 ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3 ft. C Flickr	-	-	Original	93.80	99.50	99.70
BEiT3 ft. C Flickr	Flickr	Flickr	DN	92.20	99.10	99.70
BEiT3 ft. C Flickr	Flickr	Flickr	DBNorm	94.50	99.50	99.90
BEiT3 ft. C Flickr	Flickr	Flickr	NNN	95.10	99.40	99.80
BEiT3 ft. C Flickr	Flickr	Flickr	GLOBE	95.10	99.60	99.80
BEiT3 ft. C Flickr	COCO	COCO	NNN	95.20	99.20	99.90
BEiT3 ft. C Flickr	COCO	COCO	GLOBE	94.50	99.10	99.90
BEiT3 ft. C COCO	-	-	Original	79.06	94.36	97.22
BEiT3 ft. C COCO	Flickr	Flickr	DN	74.12	91.86	96.26
BEiT3 ft. C COCO	Flickr	Flickr	DBNorm	78.04	93.94	97.10
BEiT3 ft. C COCO	Flickr	Flickr	NNN	78.08	94.34	97.18
BEiT3 ft. C COCO	Flickr	Flickr	GLOBE	78.18	94.32	97.22
BEiT3 ft. C COCO	COCO	COCO	NNN	80.52	94.56	97.38
BEiT3 ft. C COCO	COCO	COCO	GLOBE	80.24	94.20	97.10

Table 23: Full I2T results of different algorithms on BEiT3 Large.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-L Flickr	-	-	Original	82.90	95.10	96.80
BEiT3-L Flickr	Flickr	Flickr	DN	83.40	94.20	96.50
BEiT3-L Flickr	Flickr	Flickr	DBNorm	87.20	95.70	97.60
BEiT3-L Flickr	Flickr	Flickr	NNN	87.00	95.80	97.40
BEiT3-L Flickr	Flickr	Flickr	GLOBE	87.50	95.80	97.50
BEiT3-L Flickr	COCO	COCO	NNN	86.70	95.20	97.50
BEiT3-L Flickr	COCO	COCO	GLOBE	87.10	95.30	97.60
BEiT3-L COCO	-	-	Original	55.46	78.56	86.16
BEiT3-L COCO	Flickr	Flickr	DN	44.26	68.96	79.10
BEiT3-L COCO	Flickr	Flickr	DBNorm	59.14	81.36	88.24
BEiT3-L COCO	Flickr	Flickr	NNN	59.18	81.64	88.14
BEiT3-L COCO	Flickr	Flickr	GLOBE	59.44	81.24	88.10
BEiT3-L COCO	COCO	COCO	NNN	62.66	83.12	89.36
BEiT3-L COCO	COCO	COCO	GLOBE	62.98	83.08	89.18

Table 25: Full I2T results of different algorithms on BEiT3 Large ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-L ft. F Flickr	-	-	Original	88.00	96.70	98.20
BEiT3-L ft. F Flickr	Flickr	Flickr	DN	87.90	96.50	97.80
BEiT3-L ft. F Flickr	Flickr	Flickr	DBNorm	89.70	97.10	98.40
BEiT3-L ft. F Flickr	Flickr	Flickr	NNN	89.90	97.30	98.50
BEiT3-L ft. F Flickr	Flickr	Flickr	GLOBE	90.30	97.20	98.40
BEiT3-L ft. F Flickr	COCO	COCO	NNN	89.00	97.00	98.10
BEiT3-L ft. F Flickr	COCO	COCO	GLOBE	89.20	96.90	98.20
BEiT3-L ft. F COCO	-	-	Original	63.28	84.38	90.84
BEiT3-L ft. F COCO	Flickr	Flickr	DN	53.18	76.76	84.98
BEiT3-L ft. F COCO	Flickr	Flickr	DBNorm	64.68	84.68	91.02
BEiT3-L ft. F COCO	Flickr	Flickr	NNN	64.80	85.00	91.14
BEiT3-L ft. F COCO	Flickr	Flickr	GLOBE	64.80	85.06	91.16
BEiT3-L ft. F COCO	COCO	COCO	NNN	67.98	86.72	91.72
BEiT3-L ft. F COCO	COCO	COCO	GLOBE	68.16	86.74	92.08

Table 22: Full I2T results of different algorithms on BEiT3 ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3 ft. F Flickr	-	-	Original	95.60	99.50	100.00
BEiT3 ft. F Flickr	Flickr	Flickr	DN	93.50	99.20	99.70
BEiT3 ft. F Flickr	Flickr	Flickr	DBNorm	95.10	99.60	99.90
BEiT3 ft. F Flickr	Flickr	Flickr	NNN	95.30	99.80	99.90
BEiT3 ft. F Flickr	Flickr	Flickr	GLOBE	95.10	99.70	99.90
BEiT3 ft. F Flickr	COCO	COCO	NNN	95.60	99.70	100.00
BEiT3 ft. F Flickr	COCO	COCO	GLOBE	95.70	99.70	100.00
BEiT3 ft. F COCO	-	-	Original	71.02	90.14	94.42
BEiT3 ft. F COCO	Flickr	Flickr	DN	65.64	87.40	93.04
BEiT3 ft. F COCO	Flickr	Flickr	DBNorm	70.86	89.64	94.66
BEiT3 ft. F COCO	Flickr	Flickr	NNN	71.22	89.16	94.16
BEiT3 ft. F COCO	Flickr	Flickr	GLOBE	70.42	88.56	93.80
BEiT3 ft. F COCO	COCO	COCO	NNN	74.90	90.62	94.98
BEiT3 ft. F COCO	COCO	COCO	GLOBE	75.20	90.76	95.04

Table 24: Full I2T results of different algorithms on BEiT3 Large ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-L ft. C Flickr	-	-	Original	87.80	97.50	98.20
BEiT3-L ft. C Flickr	Flickr	Flickr	DN	86.50	95.30	97.70
BEiT3-L ft. C Flickr	Flickr	Flickr	DBNorm	89.00	97.50	98.70
BEiT3-L ft. C Flickr	Flickr	Flickr	NNN	89.00	97.80	98.70
BEiT3-L ft. C Flickr	Flickr	Flickr	GLOBE	89.80	97.70	98.50
BEiT3-L ft. C Flickr	COCO	COCO	NNN	87.80	97.40	98.50
BEiT3-L ft. C Flickr	COCO	COCO	GLOBE	88.90	97.80	98.50
BEiT3-L ft. C COCO	-	-	Original	73.74	90.78	94.92
BEiT3-L ft. C COCO	Flickr	Flickr	DN	59.78	82.88	89.64
BEiT3-L ft. C COCO	Flickr	Flickr	DBNorm	73.30	91.00	95.04
BEiT3-L ft. C COCO	Flickr	Flickr	NNN	74.36	91.22	95.28
BEiT3-L ft. C COCO	Flickr	Flickr	GLOBE	74.54	91.06	95.18
BEiT3-L ft. C COCO	COCO	COCO	NNN	75.30	91.16	95.34
BEiT3-L ft. C COCO	COCO	COCO	GLOBE	75.66	91.16	95.26

Table 26: Full T2I results of different algorithms on CLIP ViT-B/32.

Model	Target	Ref.	Method	R@1	R@5	R@10
CLIP Flickr	-	-	Original	56.52	81.38	87.72
CLIP Flickr	Flickr	Flickr	DN	60.42	84.18	90.30
CLIP Flickr	Flickr	Flickr	DBNorm	61.56	84.60	90.58
CLIP Flickr	Flickr	Flickr	NNN	62.92	86.46	91.94
CLIP Flickr	Flickr	Flickr	GLOBE	64.38	87.06	92.76
CLIP Flickr	COCO	COCO	NNN	61.76	85.16	90.84
CLIP Flickr	COCO	COCO	GLOBE	62.32	85.26	91.10
CLIP COCO	-	-	Original	30.24	54.94	65.72
CLIP COCO	Flickr	Flickr	DN	32.35	58.12	68.84
CLIP COCO	Flickr	Flickr	DBNorm	33.49	59.07	69.53
CLIP COCO	Flickr	Flickr	NNN	32.78	57.98	68.60
CLIP COCO	Flickr	Flickr	GLOBE	32.34	57.41	68.27
CLIP COCO	COCO	COCO	NNN	36.80	62.25	72.42
CLIP COCO	COCO	COCO	GLOBE	37.92	63.21	73.33

Table 27: Full T2I results of different algorithms on SigLIP ViT-B/16.

Model	Target	Ref.	Method	R@1	R@5	R@10
SigLIP	Flickr	-	Original	73.94	91.88	95.46
SigLIP	Flickr	Flickr	DN	73.18	91.50	94.70
SigLIP	Flickr	Flickr	DBNorm	75.96	92.88	96.06
SigLIP	Flickr	Flickr	NNN	76.10	93.22	96.18
SigLIP	Flickr	Flickr	GLOBE	76.68	93.44	96.14
SigLIP	Flickr	COCO	NNN	75.44	92.66	95.92
SigLIP	Flickr	COCO	GLOBE	75.24	92.78	95.70
SigLIP	COCO	-	Original	47.78	72.43	81.03
SigLIP	COCO	COCO	DN	47.76	72.47	80.83
SigLIP	COCO	COCO	DBNorm	49.67	73.97	82.17
SigLIP	COCO	Flickr	NNN	49.26	73.44	81.84
SigLIP	COCO	Flickr	GLOBE	48.34	72.57	81.26
SigLIP	COCO	COCO	NNN	51.05	74.90	83.10
SigLIP	COCO	COCO	GLOBE	51.50	75.32	83.42

Table 29: Full T2I results of different algorithms on BLIP ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
BLIP ft. F	Flickr	-	Original	93.40	99.40	99.80
BLIP ft. F	Flickr	Flickr	DN	93.60	99.50	99.90
BLIP ft. F	Flickr	Flickr	DBNorm	95.00	99.80	100.00
BLIP ft. F	Flickr	Flickr	NNN	95.80	99.80	100.00
BLIP ft. F	Flickr	Flickr	GLOBE	95.70	99.50	99.90
BLIP ft. F	Flickr	COCO	NNN	95.40	99.50	99.90
BLIP ft. F	Flickr	COCO	GLOBE	95.40	99.60	100.00
BLIP ft. F	COCO	-	Original	69.90	90.50	94.88
BLIP ft. F	COCO	COCO	DN	71.84	91.08	95.64
BLIP ft. F	COCO	COCO	DBNorm	73.10	91.88	95.84
BLIP ft. F	COCO	Flickr	NNN	73.14	91.22	95.36
BLIP ft. F	COCO	Flickr	GLOBE	73.06	91.00	95.28
BLIP ft. F	COCO	COCO	NNN	76.40	92.86	96.12
BLIP ft. F	COCO	COCO	GLOBE	76.80	92.94	96.18

Table 31: Full T2I results of different algorithms on ALBEF ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
ALBEF ft. C	Flickr	-	Original	72.74	92.10	96.14
ALBEF ft. C	Flickr	Flickr	DN	72.34	92.20	95.78
ALBEF ft. C	Flickr	Flickr	DBNorm	74.34	93.32	96.46
ALBEF ft. C	Flickr	Flickr	NNN	75.34	93.44	96.40
ALBEF ft. C	Flickr	Flickr	GLOBE	75.66	93.86	96.70
ALBEF ft. C	Flickr	COCO	NNN	74.52	93.40	96.32
ALBEF ft. C	Flickr	COCO	GLOBE	74.52	93.36	96.52
ALBEF ft. C	COCO	-	Original	53.06	79.32	87.32
ALBEF ft. C	COCO	COCO	DN	53.45	79.34	87.32
ALBEF ft. C	COCO	COCO	DBNorm	54.36	79.98	87.70
ALBEF ft. C	COCO	Flickr	NNN	53.73	79.59	87.52
ALBEF ft. C	COCO	Flickr	GLOBE	52.91	78.95	87.09
ALBEF ft. C	COCO	COCO	NNN	55.75	80.75	88.36
ALBEF ft. C	COCO	COCO	GLOBE	55.74	81.22	88.61

Table 28: Full T2I results of different algorithms on BLIP pretrained.

Model	Target	Ref.	Method	R@1	R@5	R@10
BLIP pre.	Flickr	-	Original	83.00	96.30	98.60
BLIP pre.	Flickr	Flickr	DN	86.90	96.90	99.00
BLIP pre.	Flickr	Flickr	DBNorm	88.40	97.30	99.00
BLIP pre.	Flickr	Flickr	NNN	89.10	97.90	99.20
BLIP pre.	Flickr	Flickr	GLOBE	90.20	97.90	98.90
BLIP pre.	Flickr	COCO	NNN	85.00	97.00	98.90
BLIP pre.	Flickr	COCO	GLOBE	86.20	97.00	98.80
BLIP pre.	COCO	-	Original	61.88	85.44	92.08
BLIP pre.	COCO	COCO	DN	67.24	88.90	94.58
BLIP pre.	COCO	COCO	DBNorm	68.74	89.52	94.50
BLIP pre.	COCO	Flickr	NNN	65.06	87.60	93.40
BLIP pre.	COCO	Flickr	GLOBE	65.36	87.58	93.10
BLIP pre.	COCO	COCO	NNN	71.68	90.68	95.30
BLIP pre.	COCO	COCO	GLOBE	72.18	90.96	95.08

Table 30: Full T2I results of different algorithms on BLIP ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
BLIP ft. C	Flickr	-	Original	90.80	98.70	99.60
BLIP ft. C	Flickr	Flickr	DN	90.80	99.20	99.70
BLIP ft. C	Flickr	Flickr	DBNorm	92.10	99.30	99.80
BLIP ft. C	Flickr	Flickr	NNN	92.80	99.40	99.90
BLIP ft. C	Flickr	Flickr	GLOBE	93.00	99.30	99.80
BLIP ft. C	Flickr	COCO	NNN	93.80	99.00	99.90
BLIP ft. C	Flickr	COCO	GLOBE	94.00	99.00	99.80
BLIP ft. C	COCO	-	Original	75.76	93.80	96.62
BLIP ft. C	COCO	COCO	DN	77.34	93.68	96.88
BLIP ft. C	COCO	COCO	DBNorm	78.00	94.02	96.88
BLIP ft. C	COCO	Flickr	NNN	77.50	93.58	96.70
BLIP ft. C	COCO	Flickr	GLOBE	77.44	93.68	96.66
BLIP ft. C	COCO	COCO	NNN	79.10	94.00	97.02
BLIP ft. C	COCO	COCO	GLOBE	79.06	94.20	97.10

Table 32: Full T2I results of different algorithms on ALBEF ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
ALBEF ft. F	Flickr	-	Original	78.62	95.34	97.36
ALBEF ft. F	Flickr	Flickr	DN	76.74	94.26	97.00
ALBEF ft. F	Flickr	Flickr	DBNorm	80.08	95.46	97.58
ALBEF ft. F	Flickr	Flickr	NNN	79.74	95.38	97.64
ALBEF ft. F	Flickr	Flickr	GLOBE	80.08	95.38	97.42
ALBEF ft. F	Flickr	COCO	NNN	78.98	95.36	97.48
ALBEF ft. F	Flickr	COCO	GLOBE	79.08	95.32	97.46
ALBEF ft. F	COCO	-	Original	49.52	76.56	85.03
ALBEF ft. F	COCO	Flickr	DN	44.64	72.10	81.67
ALBEF ft. F	COCO	Flickr	DBNorm	50.41	76.81	85.18
ALBEF ft. F	COCO	Flickr	NNN	50.20	76.50	84.73
ALBEF ft. F	COCO	Flickr	GLOBE	50.15	76.89	85.20
ALBEF ft. F	COCO	COCO	NNN	53.07	78.94	87.11
ALBEF ft. F	COCO	COCO	GLOBE	53.42	78.99	87.04

didates, approximate nearest-neighbor indices such

as IVF(Johnson et al., 2019) or HNSW(Malkov

Table 33: Full T2I results of different algorithms on BEiT3-base-its.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-base	Flickr	Flickr	Original	74.36	92.04	95.76
BEiT3-base	Flickr	Flickr	DN	72.86	91.26	95.12
BEiT3-base	Flickr	Flickr	DBNorm	75.42	92.98	96.16
BEiT3-base	Flickr	Flickr	NNN	76.52	93.86	96.54
BEiT3-base	Flickr	Flickr	GLOBE	77.54	93.90	96.76
BEiT3-base	Flickr	COCO	NNN	76.28	93.18	96.28
BEiT3-base	Flickr	COCO	GLOBE	76.64	93.40	96.30
BEiT3-base	COCO	-	Original	47.52	72.03	80.45
BEiT3-base	COCO	COCO	DN	45.13	70.40	79.62
BEiT3-base	COCO	COCO	DBNorm	47.96	72.75	81.08
BEiT3-base	COCO	Flickr	NNN	48.04	72.48	81.00
BEiT3-base	COCO	Flickr	GLOBE	47.61	72.45	80.92
BEiT3-base	COCO	COCO	NNN	50.25	74.75	82.51
BEiT3-base	COCO	COCO	GLOBE	50.79	75.37	82.93

Table 35: Full T2I results of different algorithms on BEiT3 ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3 ft. C	Flickr	-	Original	82.16	96.64	98.48
BEiT3 ft. C	Flickr	Flickr	DN	78.80	95.42	97.46
BEiT3 ft. C	Flickr	Flickr	DBNorm	81.68	96.44	98.30
BEiT3 ft. C	Flickr	Flickr	NNN	82.32	96.74	98.54
BEiT3 ft. C	Flickr	Flickr	GLOBE	83.54	96.80	98.68
BEiT3 ft. C	Flickr	COCO	NNN	82.18	96.58	98.50
BEiT3 ft. C	Flickr	COCO	GLOBE	81.72	96.52	98.30
BEiT3 ft. C	COCO	-	Original	61.38	84.61	90.76
BEiT3 ft. C	COCO	COCO	DN	58.76	82.40	89.30
BEiT3 ft. C	COCO	COCO	DBNorm	60.87	84.03	90.50
BEiT3 ft. C	COCO	Flickr	NNN	61.39	84.39	90.73
BEiT3 ft. C	COCO	Flickr	GLOBE	60.06	83.59	90.19
BEiT3 ft. C	COCO	COCO	NNN	61.58	84.56	90.80
BEiT3 ft. C	COCO	COCO	GLOBE	60.91	84.30	90.71

Table 37: Full T2I results of different algorithms on BEiT3 Large ft. COCO.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-L ft. C	Flickr	-	Original	76.46	92.88	95.88
BEiT3-L ft. C	Flickr	Flickr	DN	74.00	91.18	94.76
BEiT3-L ft. C	Flickr	Flickr	DBNorm	77.40	93.38	96.14
BEiT3-L ft. C	Flickr	Flickr	NNN	77.52	93.62	96.30
BEiT3-L ft. C	Flickr	Flickr	GLOBE	77.58	93.18	96.32
BEiT3-L ft. C	Flickr	COCO	NNN	76.22	92.66	95.88
BEiT3-L ft. C	Flickr	COCO	GLOBE	76.08	92.54	95.74
BEiT3-L ft. C	COCO	-	Original	56.35	80.79	88.24
BEiT3-L ft. C	COCO	Flickr	DN	48.31	73.74	82.19
BEiT3-L ft. C	COCO	Flickr	DBNorm	54.37	79.16	86.71
BEiT3-L ft. C	COCO	Flickr	NNN	56.19	80.55	87.94
BEiT3-L ft. C	COCO	Flickr	GLOBE	56.17	80.44	87.97
BEiT3-L ft. C	COCO	COCO	NNN	56.71	81.05	88.44
BEiT3-L ft. C	COCO	COCO	GLOBE	56.91	81.00	88.44

Table 34: Full T2I results of different algorithms on BEiT3 ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3 ft. F	Flickr	-	Original	85.52	97.36	98.74
BEiT3 ft. F	Flickr	Flickr	DN	81.48	95.90	97.98
BEiT3 ft. F	Flickr	Flickr	DBNorm	84.52	97.04	98.74
BEiT3 ft. F	Flickr	Flickr	NNN	85.90	97.44	98.70
BEiT3 ft. F	Flickr	Flickr	GLOBE	85.66	97.38	98.80
BEiT3 ft. F	Flickr	COCO	NNN	85.58	97.32	98.80
BEiT3 ft. F	Flickr	COCO	GLOBE	85.26	97.22	98.72
BEiT3 ft. F	COCO	-	Original	52.78	77.12	85.07
BEiT3 ft. F	COCO	COCO	DN	49.33	74.78	83.37
BEiT3 ft. F	COCO	COCO	DBNorm	52.57	77.33	85.25
BEiT3 ft. F	COCO	Flickr	NNN	53.15	77.36	85.15
BEiT3 ft. F	COCO	Flickr	GLOBE	52.12	76.43	84.47
BEiT3 ft. F	COCO	COCO	NNN	55.49	79.15	86.40
BEiT3 ft. F	COCO	COCO	GLOBE	56.02	79.34	86.75

Table 36: Full T2I results of different algorithms on BEiT3 Large.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-L	Flickr	-	Original	69.82	89.12	93.44
BEiT3-L	Flickr	Flickr	DN	68.40	88.14	92.34
BEiT3-L	Flickr	Flickr	DBNorm	71.94	90.40	94.38
BEiT3-L	Flickr	Flickr	NNN	71.74	90.20	94.36
BEiT3-L	Flickr	Flickr	GLOBE	72.12	90.42	94.26
BEiT3-L	Flickr	COCO	NNN	71.50	89.76	93.76
BEiT3-L	Flickr	COCO	GLOBE	71.42	89.94	93.78
BEiT3-L	COCO	-	Original	43.07	67.80	76.76
BEiT3-L	COCO	Flickr	DN	35.65	60.20	70.58
BEiT3-L	COCO	Flickr	DBNorm	42.98	67.72	76.94
BEiT3-L	COCO	Flickr	NNN	43.43	68.34	77.22
BEiT3-L	COCO	Flickr	GLOBE	43.42	68.34	77.07
BEiT3-L	COCO	COCO	NNN	45.65	70.80	79.52
BEiT3-L	COCO	COCO	GLOBE	45.86	70.72	79.61

Table 38: Full T2I results of different algorithms on BEiT3 Large ft. Flickr.

Model	Target	Ref.	Method	R@1	R@5	R@10
BEiT3-L ft. F	Flickr	-	Original	77.74	93.66	96.48
BEiT3-L ft. F	Flickr	Flickr	DN	74.16	91.22	94.68
BEiT3-L ft. F	Flickr	Flickr	DBNorm	76.80	93.46	96.16
BEiT3-L ft. F	Flickr	Flickr	NNN	77.84	93.86	96.54
BEiT3-L ft. F	Flickr	Flickr	GLOBE	77.90	93.88	96.48
BEiT3-L ft. F	Flickr	COCO	NNN	78.16	93.50	96.32
BEiT3-L ft. F	Flickr	COCO	GLOBE	78.20	93.54	96.26
BEiT3-L ft. F	COCO	-	Original	47.98	72.77	81.13
BEiT3-L ft. F	COCO	Flickr	DN	39.80	64.37	74.31
BEiT3-L ft. F	COCO	Flickr	DBNorm	45.55	70.21	79.07
BEiT3-L ft. F	COCO	Flickr	NNN	47.81	72.65	81.08
BEiT3-L ft. F	COCO	Flickr	GLOBE	47.51	72.37	80.84
BEiT3-L ft. F	COCO	COCO	NNN	50.00	74.43	82.79
BEiT3-L ft. F	COCO	COCO	GLOBE	50.77	74.93	83.17

and Yashunin, 2020) can be used to generate the required shortlists, with the standard speed–recall

trade-off.

Table 39 reports estimated post-hoc calibration

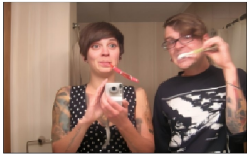
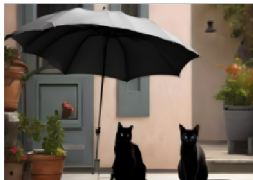
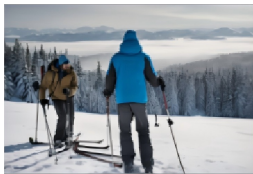
Query caption	Real COCO GT	SDXL near-positive distractor
People with a camera have toothbrushes in their mouths.		
A white toilet in a bathroom next to a sink.		
A parade on the street with a person riding on elephant.		
Two cows outside one laying down and the other standing near a building.		
A black cat sitting under an umbrella sitting on a table.		
A man standing on a snow covered field holding skis.		

Figure 4: Representative examples from IRB-COCO-500. Each row shows a query, the corresponding COCO ground-truth image, and an AI-generated near-positive distractor generated from the same caption.

runtime on a single NVIDIA RTX A4000 GPU with cached embeddings. The reported time corresponds to one fixed hyperparameter configuration and does not include feature extraction, base feature encoding, or offline hyperparameter search. The results show that Faiss substantially reduces the GLOBE calibration time and makes the reranking stage practical on COCO-scale retrieval.

Table 39: Post-hoc GLOBE calibration runtime on COCO using cached embeddings on a single NVIDIA RTX A4000 GPU. The columns “w/o Faiss” and “w/ Faiss” report the calibration time without and with Faiss-accelerated, respectively. Speedup is computed on the GLOBE post-hoc calibration stage.

Model	Data	w/o Faiss	w/ Faiss	Speedup
CLIP ViT-B/32	COCO	45 s	1.5 s	30.0×
SigLIP ViT-B/16	COCO	61 s	1.8 s	33.9×
BLIP ft. COCO	COCO	56 s	1.6 s	35.0×
ALBEF ft. F	COCO	45 s	1.4 s	32.1×
BEiT3-base	COCO	93 s	3.3 s	28.2×
BEiT3 Large	COCO	132 s	3.5 s	37.7×