Evaluating Perspectival Biases in Cross-Modal Retrieval

Teerapol Saengsukhiran[1]*, Peerawat Chomphooyod[1]*, Narabodee Rodjananant[1]*, Chompakorn Chaksangchaichot[1,2], Patawee Prakrankamanant[1], Witthawin Sripheanpol[1], Pak Lovichit[1], Sarana Nutanong[3], Ekapol Chuangsuwanich[1]†

[1]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University
[2]VISAI.AI
[3]School of Information Science and Technology, VISTEC

Abstract

Multimodal retrieval systems are expected to operate in a semantic space, agnostic to the language or cultural origin of the query. In practice, however, retrieval outcomes systematically reflect perspectival biases: deviations shaped by linguistic **prevalence** and **cultural** associations. We introduce **the Cross-Cultural, Cross-Modal, Cross-lingual Multimodal (3XCM)** benchmark to isolate these effects. Results from our studies indicate that, for image-to-text retrieval, models tend to favor entries from prevalent languages over those that are semantically faithful. For text-to-image retrieval, we observe a consistent *"tugging effect"* in the joint embedding space between semantic alignment and language-conditioned cultural association. When semantic representations are insufficiently resolved, particularly in low-resource languages, similarity is increasingly governed by culturally familiar visual patterns, leading to systematic association bias in retrieval. Our findings suggest that achieving equitable multimodal retrieval necessitates targeted strategies that explicitly decouple language from culture, rather than relying solely on broader data exposure. This work highlights the need to treat linguistic and cultural biases as distinct, measurable challenges in multimodal representation learning.

1 Introduction

As Nietzsche (1887) observed, *there is only a perspective seeing, only a perspective knowing"*; put differently, there is *no view from nowhere*. Large models inherit this perspectival character from training data, where representations depend on frequency and co-occurrence. Consequently, their latent spaces deviate from the expected robust, language-agnostic semantics, skewing retrieval toward linguistic prevalence or cultural associations over true relevance. Figure 1 illustrates this in image-to-text and text-to-image retrieval, highlighting the need to quantify such effects for consistent cross-lingual, cross-cultural performance.

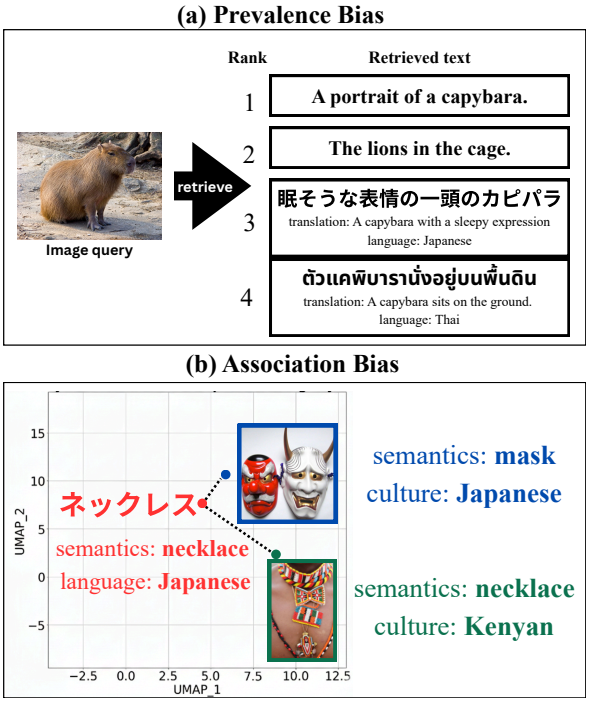

Figure 1: Two Forms of Perspectival Biases. (a) **Prevalence bias**: an image query favors high-resource languages. A model places English results above semantically equivalent Japanese and Thai captions. (b) **Association Bias**: Given a Japanese text query for "necklace", a model places a culturally proximate image (Japanese masks) closer to the query than a semantically correct one (Kenyan necklace).

Multimodal retrieval aligns text and images via paired supervision in early models like CLIP (Radford et al., 2021) or implicit pretraining in recent Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Comanici et al., 2025). Yet, language and cultural biases remain underexplored, especially given English-centric datasets like LAION (Schuhmann et al., 2022) and WebLi (Chen et al., 2022). While culturally specific images may carry

*These authors contributed equally as co-first authors.
†Corresponding Author (Email: ekapolc@cp.eng.chula.ac.th)

Findings of the Association for Computational Linguistics: ACL 2026, pages 36018–36049

native-language alt-text, e.g., Catalan descriptions for "coca de recapte" (Olóndriz et al., 2021). This fosters emergent multilingualism but risks spurious correlations, favoring "expected" languages.

A major barrier is the lack of targeted benchmarks and metrics. To address this, we develop an evaluation framework isolating two perspectival biases across retrieval directions. **For image-to-text** (lacking linguistic cues), we assess *prevalence bias* via the proposed Discounted Language Bias Kullback–Leibler Divergence (DLBKL), extending LBKL (Laosaengpha et al., 2025) to penalize high-resource languages dominating top ranks (Figure 1a). **For text-to-image** (with linguistic/cultural cues), we measure *association bias* using a parallel, cross-cultural dataset to disentangle semantic fidelity from cultural proximity (Figure 1b). Extending this, we test the ability of explicit cultural descriptors (CDs) to override association bias, examining the "tugging effect" where models struggle to prioritize CDs (e.g., "Japanese train" in Thai) over implicit query-language associations.

Using these tools, we compare biases in MLLM-adapted retrievers against those with explicit cross-lingual alignment (Chen et al., 2023; Carlsson et al., 2022). Findings expose biases in both directions, including CDs' limited efficacy: models follow instructions but fall back to language-linked visuals when cultural semantics are unresolved.

Our contributions: (i) DLBKL as a supplementary, rank-sensitive extension to LBKL for measuring prevalence bias in multilingual pools; (ii) the 3XCM benchmark, parallel across cultures/languages for association bias; (iii) analysis of CDs' impact on retrieval, revealing persistent biases despite explicit guidance.

2 Related Works

2.1 Language Bias in Multimodal Retrievers

Language bias in multimodal retrieval refers to performance disparities where high-resource languages dominate rankings for semantically equivalent queries. Osmulsk et al. (2025) report that modern retrievers exhibit significant variation in NDCG scores (Järvelin and Kekäläinen, 2002) across languages, indicating that their effectiveness depends on resources and script types.

To quantify these disparities, fairness-aware metrics like exposure parity (Järvelin and Kekäläinen, 2002) have been adapted, yet Adewumi et al. (2024) emphasizes the lack of dedicated language-focused protocols. Laosaengpha et al. (2025) introduced LBKL, a distributional measure of divergence for text modality bias. While technically extensible to multimodal retrieval, such metrics ignore retrieval rankings in the bias measurement.

2.2 Cultural Benchmarks

While recent benchmarks (Liu et al., 2021; Romero et al., 2024; Nayak et al., 2024; Li et al., 2026) have made progress in evaluating cultural reasoning and knowledge, they primarily focus on culture-specific concepts or broad facets. In contrast, the assessment of bias discussed in Section 2.1 requires a basic-level, parallel corpus of universal concepts to strictly disentangle semantic relevance from cultural association in retrieval rankings. We provide a detailed comparison to other datasets in Appendix A.

3 Methodology

This section outlines the framework developed to investigate perspectival bias in multilingual, multimodal retrieval systems. We first state our guiding research questions and subsequently detail the experiments addressing them in Sections 3.1, 3.2.2, and 3.2.3. Our investigation is organized into three specific research questions as follows:

RQ1 **[Image→Text]:** Effect of *prevalence bias*. To what extent do models favor high-resource languages over semantically equivalent captions in other languages?

RQ2 **[Text→Image]:** Effect of *culture-language association bias*. To what extent do models prioritize culture associated with the query language over semantically faithful results?

RQ3 **[Text→Image]:** Effect of *culture-language association bias under explicit cultural description*. To what extent can models adhere to explicit cultural descriptors (e.g., country mentions) against the *"tugging effect"* of inherent language-cultural association biases?

3.1 Image-to-Text Retrieval (RQ1)

To assess prevalence bias, we measure the divergence between expected and observed language distributions. We adopt the Language Bias Kullback–Leibler (LBKL) divergence (Laosaengpha et al., 2025):

$$\mathrm{LBKL} = \frac{\sum_{i=1}^{q} \left[P_{\mathrm{A}}(x) \log \frac{P_{\mathrm{A}}(x)}{Q_{\mathrm{A}}(x)} + P_{\mathrm{B}}(x) \log \frac{P_{\mathrm{B}}(x)}{Q_{\mathrm{B}}(x)} \right]}{q} \quad (1)$$

Since LBKL is rank-agnostic, we propose **Discounted LBKL (DLBKL)** to prioritize penalizing bias in higher ranks more than lower ranks. We apply a logarithmic discount $w(i) = 1/\log_2(i+1)$ to the observed language proportion $Q'_l(x)$:

$$Q'_l(x) = \frac{\sum_{i=1}^{k} w(i) \cdot \mathbb{I}(\text{doc}_i \text{ is } l)}{\sum_{i=1}^{k} w(i)} \quad (2)$$

DLBKL is computed by substituting $Q'_l(x)$ for $Q_l(x)$ in equation 1. This formulation can be generalized to support cases where the bias assessment involves more than two languages by measuring divergence over the full distribution. *It penalizes models where high-resource languages disproportionately occupy top ranks* relative to the prior P. Crucially, in parallel datasets where semantic relevance is uniform across languages, this formulation isolates prevalence bias without compromising retrieval accuracy (see Figure 2 and Appendix D).

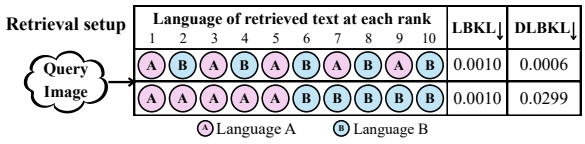

Figure 2: Illustration of how DLBKL, unlike the rank-agnostic LBKL, assigns a higher bias score to lists in which high-resource languages dominate the top ranks, under the setup that the prior P is a uniform distribution

3.2 Text-to-Image Retrieval (RQ2 & RQ3)

To quantify the degree to which models prioritize cultural association over semantic fidelity, a phenomenon we illustrate in Figure 1(b), a benchmark with a parallel structure in its cultural dimension is necessary. To the best of our knowledge, no such benchmark exists, so we make two primary contributions. First, we construct and introduce the Cross-Cultural, Cross-Modal, Cross-lingual Multimodal (3XCM) benchmark, a novel dataset designed specifically for this purpose. Second, we propose the Self-Preference Cultural Bias Score (SP), a new metric for explicitly measuring this form of bias.

3.2.1 The 3XCM Dataset Benchmark

To evaluate association bias, we constructed the 3XCM benchmark, a challenge set designed to elicit culture-language association bias. This evaluation requires strict cultural parallelism at the concept level (e.g., "school" or "wedding" across all cultures), a structural feature that existing datasets

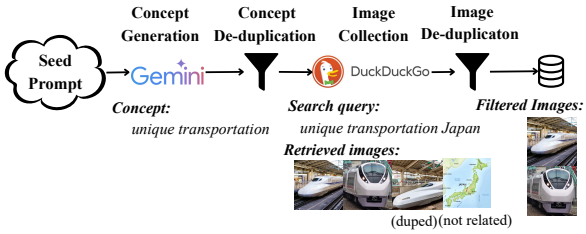

Figure 3: Overview of the XCM dataset creation process, designed to produce a benchmark with parallelism across semantics, cultures, and languages.

(Liu et al., 2021; Romero et al., 2024; Nayak et al., 2024) do not provide.[1] The process involved two primary stages: (i) gathering a corpus of culturally diverse images and (ii) structuring these images into a triplet-based evaluation set.

The image gathering stage, summarized in Figure 3, consisted of three steps:

- **Concept Generation** We used Gemini[2] to generate a large pool of concepts, which we manually curated to a final set of 138 coarse-grained, culturally-inclusive concepts (e.g., "train", "food"). Each concept is an abstract, semantic category that uses shared properties to group a broad, culturally-inclusive range of entities. The prompt for generating concepts can be found in Appendix H.
- **Concept De-duplication:** We use BGE-M3 (Chen et al., 2024) to de-duplicate concepts based on similarity with a threshold of 0.92.
- **Image Collection:** For each concept and a set of 16 diverse countries, we used the DuckDuckGo image search API (Prabhu, 2025) to retrieve the top 10 images using queries in both English (e.g., "train Japan") and the local native language.
- **Image De-duplication:** To ensure visual diversity, we performed two-stage de-duplication within each concept. First, near-exact duplicates were removed automatically using an embedding model. Subsequently, three human annotators, following the guidelines in Appendix J, used a custom tool to manually filter out remaining images that depicted the same scene or object without meaningful variation in viewpoint or time of day.

The final dataset contains 11,724 entries distributed

[1]Research release only (CC BY-NC-SA 4.0). Ethical review required for production use. https://huggingface.co/datasets/Chula-AI/association_bias_benchmark

[2]Version used: `gemini-2.5-flash` (Released June 17, 2025).

across 138 concepts. Further statistics and samples are provided in Appendix K and L respectively. Furthermore, details regarding annotator workloads and dataset generation costs are provided in Appendix X.

3.2.2 RQ2: Effect of Implicit Association Bias

To evaluate implicit association bias, we set up a forced-choice task designed to disambiguate between the model's reliance on semantic understanding (the concept) and its preference for cultural association. As illustrated in Figure 4(a), for a given query (e.g., "train" in Thai), the model is presented with a triplet of image candidates representing three answer categories: (i) **Correct**: The model successfully identifies an image with a semantically relevant object (e.g., a train of any culture). (ii) **Language-biased**: The model prioritizes culture-language association over semantic relevance (e.g., a Thai dance). (iii) **Totally irrelevant**: The model retrieves an image *without* any semantic or cultural relevance (e.g., the Eiffel Tower).

Note that we deliberately exclude images with relevance in terms of object semantics and query language to expose the prioritization of relevance.

With the constructed dataset, we can now measure the discrepancy between ideal retrieval outcomes and observed results, where bias arising from cultural association may intervene. Ideally, the discrepancy should be zero when image retrieval depends solely on semantic relevance, and it should increase as the model's preference tends towards images associated with the culture of the query, rather than semantic accuracy. To quantify the discrepancy, we propose a metric called the **self-preference cultural bias score (SP)**, which can be computed as follows:

$$M_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(S_{k,i} = \max(S_{\text{cr},i}, S_{\text{lb},i}, S_{\text{ti},i})\right) \tag{3}$$

$$\text{SP} = \frac{M_{\text{lb}}}{M_{\text{cr}}} \tag{4}$$

where M_k is the proportion of times a candidate of type k receives the highest similarity score across N total trials. The candidate type k can be **Correct** (`cr`), **Language-biased** (`lb`), or **Totally Irrelevant** (`ti`). The similarity score for candidate type k in trial i is denoted by $S_{k,i}$. The indicator function $\mathbb{I}(\cdot)$ is 1 if the condition is true and 0 otherwise. The SP score (equation 4) is then the ratio of language-biased wins (M_{lb}) to correct answer wins (M_{cr}). In this way, a higher SP score indicates stronger cultural self-preference over semantic faithfulness and thus a greater extent of association bias.

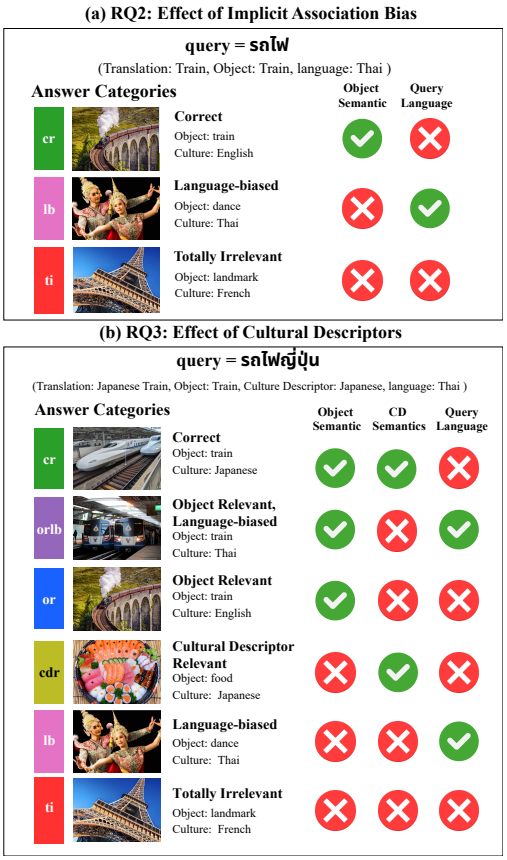

Figure 4: Illustration of *association bias* evaluation. (a) RQ2 evaluates a Thai query for "train" against three candidates to distinguish semantic faithfulness from cultural relevance. (b) RQ3 uses the query "train of Japan" to measure the tension between explicit cultural descriptors and implicit language-cultural associations.

3.2.3 RQ3: Effect of Cultural Descriptors

To investigate the model's adherence to explicit instructions against language-driven bias, we extend the forced-choice framework to include specific cultural cues. We construct queries using an "Object & Cultural Descriptor" format (e.g., "Train of Japan" formulated in Thai) to gauge the "tugging effect" between the explicit country mention and the implicit language-cultural association. As illustrated in Figure 4(b), we expand the answer category to six to isolate specific retrieval failures: (i) **Correct**: The model successfully identifies an image with a semantically relevant object and cultural descriptor (e.g., a Shinkansen). (ii) **Object Relevant, Language-biased**: The model successfully

identifies an image with a semantically relevant object, but is distracted by the culture associated with the query language (e.g., a Thai train); (iii) **Object Relevant**: The model successfully identifies an image with a semantically relevant object, but the culture is irrelevant to the query language and the cultural descriptor (e.g., British Train). (iv) **Cultural Descriptor Relevant**: The model successfully identifies an image associated with the cultural descriptor, but misses the object's semantics (e.g., Sushi). (v) **Language-biased**: The model fails to identify an image with a semantically relevant object but matches the query language's culture (e.g., Thai dance). (vi) **Totally Irrelevant**: The model captures neither the relevant object nor the culture associated with the descriptor or language. (e.g., Eiffel Tower).

To quantify the "tugging effect" between implicit language bias and explicit instructions, we calculate the **Similarity Drift** (Δsim). Given a base concept query q_C and a culturally-augmented query q_{C+CD}, the drift for an image I is defined as:

$$\Delta\text{sim}(I) = S_{q_{C+CD},I} - S_{q_C,I} \tag{5}$$

A higher Δsim for the Correct category relative to other categories with mismatch culture indicates a model's ability to prioritize explicit descriptors over inherent culture-language associations.

4 Experimental Setup

To answer our research questions, we conducted three main studies. **For RQ1**, we performed image-to-text retrieval on the Crossmodal-3600 dataset (Thapliyal et al., 2022). The dataset offers a parallel multilingual text pool comprising native captions in 36 languages, making it suitable for auditing cross-lingual behavior in image–text retrieval, without implying any particular pattern of disparities. With this dataset, we compute Language-wise DLBKL by defining the prior $P(x)$ as a uniform distribution (see Appendix E)". We evaluate models using Accuracy@5, NDCG@10, LBKL@10, and our proposed DLBKL@10. **For RQ2**, we performed text-to-image retrieval on our newly created 3XCM benchmark, evaluating models using our proposed SP score. **For RQ3**, we extend text-to-image retrieval of RQ2 by using the extended version of the 3XCM benchmark with six candidates and analyze the similarity drift.

We selected a representative suite of models spanning three distinct architectural paradigms: **Vision-Language Contrastive Models**: foundational models trained with a contrastive objective mainly on English data, **Cross-lingual Alignment Models**: models use knowledge distillation to explicitly align multilingual text encoders to a fixed, pre-trained vision space for multilingual capability, and **MLLM-Based Retrieval Embedders**: utilize MLLM as the backbone and finetune for a contrastive objective for retrieval. We refer readers to Appendix C for detailed architectural descriptions and specific model configurations.

5 Experimental Results

Our experiments are designed to provide empirical examinations of perspectival biases manifested in image-to-text and text-to-image retrievals.

5.1 Image-to-Text Evaluation (RQ1)

All models exhibit some degree of linguistic prevalence bias, as shown in Table 1. This bias is most pronounced in the top-ranked results, as models tend to prioritize high-resource languages over others in the early ranks. Appendix G provides visualizations revealing the dominance of high-resource languages in the top ranks and the overall disparity in retrieval frequency between language resource tiers. Results for additional ranks and an example of a retrieval result can be found in Appendix F.

Model	Acc @5↑	LBKL @10↓	DLBKL @10↓	NDCG @10↑
Vision-Language Contrastive Models				
CLIP-L/14	0.509	15.394	15.398	0.290
CN-CLIP-L/14	0.355	15.287	15.291	0.207
Cross-lingual Alignment Models				
XLM-R-L/14	0.924	12.775	12.793	0.736
XLM-R-B/16plus	0.968	**12.651**	**12.669**	**0.791**
MLLM-Based Retrieval Embedders				
ColQwen2.5-3b-M	0.894	13.654	13.667	0.605
ColQwen2.5-7b-M	0.926	13.439	13.454	0.665
ColQwen2.5-v0.2	0.754	14.228	14.241	0.481
GME-Qwen2-2B	0.967	13.627	13.644	0.717
GME-Qwen2-7B	**0.979**	13.378	13.397	0.770
Jina-E-v4	0.972	13.008	13.025	0.775

Table 1: Image-to-text retrieval on Crossmodal-3600. Bias is measured by LBKL and DLBKL, calculated in a language-wise manner. Explicit alignment models (XLM-R) show substantially lower bias.

Crucially, the explicit alignment models (XLM-R series) achieve the lowest bias scores by a significant margin, with XLM-R-B/16plus demonstrating near-zero linguistic prevalence bias according to both metrics, while maintaining high retrieval ac-

curacy. This provides strong initial evidence that direct alignment is a more effective strategy for enforcing language fairness than relying on emergent capabilities from large-scale pre-training.

Building on these observations, we note that LBKL and DLBKL **quantify distributional bias rather than relevance**, and therefore need *not* correlate with accuracy or NDCG in Table 1. To assess both correctness and fairness, these bias metrics should be interpreted jointly with accuracy (and/or NDCG). Finally, while LBKL/DLBKL capture cross-language imbalance, they do *not* measure model self-preference (e.g., favoring the query language over others); we operationalize and evaluate that phenomenon with our SP score.

5.2 Text-to-Image Evaluation (RQ2)

Using the proposed 3XCM benchmark, we evaluated the association bias of several multimodal retrievers, ranging from CLIP to more recent models. In this evaluation, the semantic win rate (M_{cr}) sserves as a proxy for raw performance, while the SP score quantifies cultural bias. Furthermore, we compared these models against a random baseline (see Appendix R). We observe that the baseline CLIP and CN-CLIP models exhibit a significant cultural bias, **often preferring a culturally associated but semantically incorrect image**, as shown in Table 2.

Model	$M_{\text{cr}}\uparrow$	$M_{\text{lb}}\downarrow$	$M_{\text{ti}}\downarrow$	SP$\downarrow$
Random Baseline	33.5%	33.2%	33.3%	0.99
Vision-Language Contrastive Models				
CLIP-L/14	51.24%	40.78%	7.98%	0.80
CN-CLIP-L/14	56.39%	31.65%	11.95%	0.56
Cross-lingual Alignment Models				
XLM-R-L/14	85.53%	6.84%	7.63%	0.08
XLM-R-B/16plus	87.54%	**6.23%**	6.24%	**0.07**
MLLM-Based Retrieval Embedders				
GME-Qwen2-2B	83.34%	11.64%	5.02%	0.14
GME-Qwen2-7B	84.63%	11.26%	**4.11%**	0.13
ColQwen2.5-v0.2	82.10%	10.93%	6.97%	0.13
ColQwen2.5-3B-M	83.36%	10.65%	6.00%	0.13
ColQwen2.5-7B-M	84.07%	11.40%	4.53%	0.14
Jina-E-v4	**87.56%**	7.20%	5.24%	0.08

Table 2: Results on the XCM benchmark for Self-Preference Cultural Bias.

Our culture-specific analysis reveals that this self-preference is a symptom of missing linguistic knowledge, as shown in Figure 5. The CLIP-L/14 model, lacking a robust understanding of non-Latin scripts, defaults to matching cultural origin as a retrieval heuristic. Training on a large Chinese dataset (CN-CLIP) partially addresses this, improving performance for both Chinese and Japanese queries due to the shared logographic Kanji characters. However, this is a shallow fix that fails to generalize to other non-Latin scripts. In contrast, the text-aligned model (XLM-R-L/14) performs well across most languages, with a notable exception for queries in Yoruba (Nigeria). This challenge with low-resource languages persists even in more advanced architectures. For instance, MLLM-based models employ an LLM as their text encoder, leveraging its pre-training on web-scale multilingual data for a robust understanding of diverse languages. For the vision component, a Vision-Language Model (VLM) is used as the image encoder to improve contextual awareness. However, performance drops for low-resource languages.

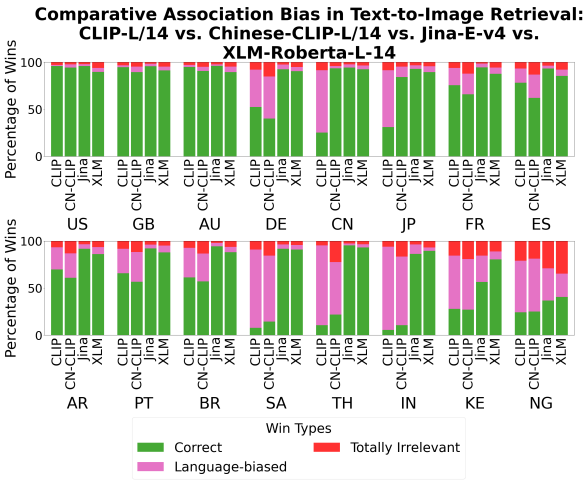

Figure 5: Association bias evaluation across four models reveals the limitations of monolingual training. The baseline CLIP shows significant cultural bias, which is exacerbated by region-specific fine-tuning as seen in CN-CLIP. In contrast, cross-lingual models like XLM-R and particularly Jina-E-v4 prove far more effective at overriding this bias and maintaining high semantic relevance across diverse countries.

This behavior is clearly visualized in the UMAP (McInnes et al., 2018) projections of the text embeddings as shown in Figure 6. The baseline CLIP-L/14 model exhibits a fractured embedding space, with non-Latin languages forming distinct clusters far from the main Latin-script cluster. This demonstrates a lack of shared semantic understanding. In the CN-CLIP model, the Chinese and Japanese embeddings shift closer to the Latin cluster, reflecting the targeted training, but other non-Latin languages remain isolated. In contrast, the explicit alignment model, XLM-R-L/14, successfully unifies the embedding space into a single, language-agnostic clus-

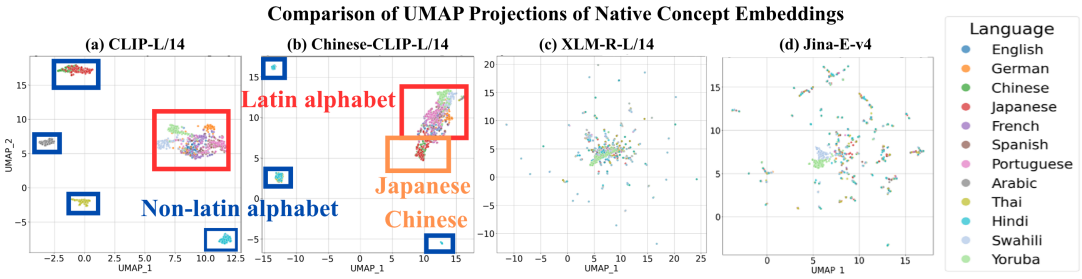

Figure 6: UMAP projection of native concept embeddings across four models: (a) CLIP-L/14 (non-Latin language separation), (b) CN-CLIP-L/14 (language family clustering), (c) XLM-R-L/14 (dense single-cluster unification), and (d) Jina-E-v4 (unified but dispersed cluster).

ter, demonstrating a truly shared semantic representation across scripts. The only notable outlier is Yoruba, which was *not* part of this specific model's alignment training. The MLLM model, Jina-E-v4, exhibits a similar but distinct pattern: it also forms a single, unified cluster, but the embeddings are more widely dispersed. This suggests a more flexible alignment that may capture finer semantic nuances between languages.

To validate these visual findings numerically, we calculated the silhouette score (Shahapure and Nicholas, 2020) for each language's text embeddings. This analysis revealed a strong Pearson correlation (0.68) between a language's silhouette score and its measured SP score as shown in Appendix N. This quantitatively reinforces that poor semantic understanding in the text encoder (as visualized by the disparate UMAP clusters) is a key driver of higher cultural association bias.

Both modern MLLM-based models and explicit alignment models drastically reduce the association bias compared to the baselines, achieving SP scores below 0.16. However, neither paradigm consistently outperforms the other on this specific task.

Model behavior remains consistent regardless of prompt phrasing, such as verbose descriptions (e.g., "a picture of a train for my homework") or culture-agnostic instructions (e.g., "focus on semantics instead of textual language") (Appendices S.1 and U).

5.3 Cultural Descriptor Text-to-Image Evaluation (RQ3)

To investigate whether explicit instruction can override implicit association bias, we evaluate the models on the Extended 3XCM benchmark using queries with explicit Cultural Descriptors (CD) (e.g., "*Train of Japan*" in Thai).

Adherence to Cultural Descriptors. Quantitative results demonstrate that most models are responsive to explicit cultural cues. As shown in Table 3, **the correct answer** (matching both object and cultural descriptor) is the most frequently selected candidate across all high-performing models (XLM series, MLLM-based). This suggests that explicit cultural descriptors (e.g., the word "Japan") exert a stronger influence on the retrieval ranking than the implicit bias of the query language (e.g., Thai script), provided the model has sufficient multilingual capacity.

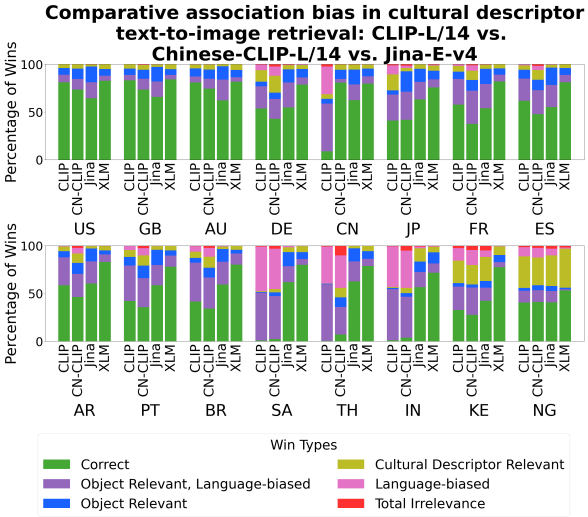

Figure 7: Comparative association bias across CLIP-L/14, CN-CLIP-L/14, Jina-embeddings-v4, and XLM-R-L/14 models in 16 countries. Stacked bars show win percentages for six categories, ranging from Correct to Total Irrelevance, highlighting performance variance across diverse cultural contexts.

Failure Mode Analysis. Let us examine two failure modes presented by the error distribution in Figure 7 First, when the model successfully retrieves the correct semantic object but fails to adhere to the explicit cultural descriptor (i.e., "Semantic Success,

Model	M_{cr} (%)	M_{orlb} (%)	M_{or} (%)	M_{cdr} (%)	M_{lb} (%)	M_{ti} (%)
Vision-Language Contrastive Models						
CLIP-L/14	43.05	29.83	5.30	8.54	12.56	0.72
CN-CLIP-L/14	41.97	24.54	9.81	10.04	11.21	2.42
Cross-lingual Alignment Models						
XLM-R-L/14	77.83	7.21	6.85	7.55	0.26	0.30
XLM-R-B/16plus	72.98	9.88	9.70	6.57	0.40	0.48
MLLM-Based Retrieval Models						
GME-Qwen2-2B	60.01	18.38	11.49	8.05	1.56	0.50
GME-Qwen2-7B	59.75	20.84	10.79	6.81	1.38	0.43
ColQwen2.5-v0.2	54.77	20.15	12.57	9.56	2.13	0.82
ColQwen2.5-3B-M	58.73	20.71	9.26	9.37	1.31	0.62
ColQwen2.5-7B-M	63.19	17.70	7.82	9.80	1.11	0.38
Jina-E-v4	57.76	19.59	13.62	7.14	1.20	0.68

Table 3: Association bias results on the 3XCM benchmark (6 categories). M_{cr}, M_{orlb}, M_{or}, M_{cdr}, M_{lb}, and M_{ti} represent Correct, Object Relevant with Language Bias, Object Relevant, Cultural Descriptor Relevant, Language-biased, and Totally Irrelevant, respectively.

Cultural Failure"), we observe that the selection rate of the Object Relevant Language-biased category (which matches the object and the query language's culture, represented by purple color) is higher than that of the Object Relevant category (which matches the object with an unrelated culture, represented by blue color). This difference is significant (p-value < 0.05, Chi-squared test) for most models except the XLM family (see Appendix P.2). The trend is most pronounced in non-Latin and low-resource languages, suggesting that when the explicit cultural constraint is violated, the model falls back on implicit language-driven associations.

Second, when the model fails to retrieve the correct semantic object (i.e., "Semantic Failure"), the model is more likely to select the Cultural Descriptor Relevant category (which matches the explicit cultural descriptor only, represented by yellow color) than the Language-biased category (which matches the query language's culture only, represented by pink color). This difference is significant (p-value < 0.05, Chi-squared test) for most models except the Vision-Language Contrastive group (CLIP and CN-CLIP), which lack robust non-Latin script understanding (see Appendix P.2). This indicates that when the model fails to ground the visual object, it prioritizes the explicit cultural token (e.g., the country name) over the implicit association of the script. (See Appendices P and S.2).

Similarity Drift Analysis. The Similarity Drift (Δsim) metric measures the change in cosine similarity for a target image when a cultural descriptor is added to the query. On average, adding a cultural descriptor successfully shifts the similarity distribution in the positive direction (toward the **Correct** category) across most models and languages. This confirms that the models generally recognize and attend to the explicit instruction. However, a notable exceptions emerge. For the baseline CLIP model, queries in non-Latin scripts exhibit near-zero drift, indicating that the text encoder fails to process the cultural descriptor in these scripts, rendering the explicit instruction ineffective (see Appendix Q.1).

6 Discussion

Perspectival Biases and the Power of Explicit Alignment. Our investigation establishes that perspectival bias manifests in two correlated forms: prevalence bias (RQ1) from distributional imbalances in training data, favoring high-resource languages, and association bias (RQ2) from spurious correlations between language scripts and cultural visual features. While both arise from training data and correlate empirically (high DLBKL models tend toward high SP), explicit cross-lingual alignment effectively reduces both, as seen in XLM-R models' low bias scores and high accuracy. This alignment not only mitigates prevalence and association biases but also enhances the model's ability to follow explicit cultural descriptors (CDs).

Guidance via Cultural Descriptors (CDs) and the Tugging Effect. CDs provide guidance to steer culture. The positive Similarity Drift (Δsim) in RQ3 shows models can adhere to cultural descriptors, steering retrieval toward target cultures in most cases despite the tugging effect of query-language associations. However, failure mode analysis reveals the tugging effect: when matching the object but not the CD, models revert to culture-language association over neutral alternatives. This indicates association bias as a latent prior suppressed but not eliminated by guidance. Modern MLLM-based models still exhibit issues with both biases and inconsistent compliance with CD guidance, especially for low-resource or non-Latin languages, where drifts are erratic or near-zero without robust multilingual support.

SP score and joint interpretation. SP score only illustrates language bias, therefore, overall retrieval accuracy is still needed to be validated as well. If M_{cr} is near zero (retrieval failure), prioritize standard metrics over SP. The SP score is then interpreted as a conditional language-bias ratio within the correct predictions along with evaluation score

like accuracy. With this joint interpretation, we can distinguish the group of models with high retrieval accuracy (MLLM-based models) from high performers with much less language bias (the cross-lingual aligned group) as shown in Table 1.

7 Conclusion

We distinguish prevalence bias, a distributional skew, from association bias, a semantic entanglement between language and culture. Even with explicit cultural descriptors, models exhibit a systematic "tugging effect" to language-culture association bias when semantic grounding succeeds but cultural disambiguation fails. These findings demonstrate that achieving cross-cultural equity requires more than naive scaling. Future work must prioritize training paradigms that actively decorrelate language from cultural features, enforce cultural constraints for steerability, or explicit cross-lingual alignment. Only by compiling these can we achieve a retrieval system that truly operates across global contexts with cultural controllability.

8 Limitations

Our work has several limitations. First, our DLBKL metric **measures fairness via distributional parity, *not* semantic correctness.** It therefore *cannot* distinguish between retrieving irrelevant documents and over-representing a language with relevant ones.

Second, the 3XCM benchmark **simplifies culture by using country as a proxy**, a necessary choice for tractability that does *not* capture hybrid, diasporic, transnational, or sub-national cultural expressions. The benchmark's universal basic-level semantics (e.g., "food") and lack of accounting for polysemy also limit its representation of real-world query complexity, as its usage is intended to be used as a **challenging, controlled environment** used to uncover biases.

Third, the Self-Preference (SP) score is intentionally undefined (mathematically infinite) when $M_{\text{cr}} = 0$, correctly signaling *complete retrieval failure*. Moreover, M_{ti} (totally irrelevant) is omitted from SP because it captures a confusion mode rather than language-cultural association bias.

Acknowledgments

This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [grant number B13F680099]. We also would like to thank Chulalongkorn University's Artificial Intelligence Institute for providing compute and infrastructure support.

References

Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. 2024. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and 10 others. 2022. Pali: A jointly-scaled multilingual language-image model. *CoRR*, abs/2209.06794.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Common Crawl. 2025. Distribution of languages in common crawl monthly archives. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, Suzhuo, China. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Napat Laosaengpha, Thanit Tativannarat, Attapol Rutherford, and Ekapol Chuangsuwanich. 2025. Mitigating language bias in cross-lingual job retrieval: A recruitment platform perspective. *CoRR*, abs/2502.03220.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. 2026. RAVENEA: A benchmark for multimodal retrieval-augmented visual culture understanding. In *The Fourteenth International Conference on Learning Representations*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Friedrich Nietzsche. 1887. *On the Genealogy of Morals*. Vintage, New York. Part III, Section 12.

David Amat Olóndriz, Ponç Palau Puigdevall, and Adrià Salvador Palau. 2021. Foodi-ml: a large multilanguage dataset of food, drinks and groceries images and descriptions. *arXiv preprint arXiv:2110.02035*.

Radek Osmulsk, Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. 2025. Miracl-vision: A large, multilingual, visual document retrieval benchmark. *Preprint*, arXiv:2505.11651.

Deepan (deepanprabhu) Prabhu. 2025. duckduckgo-images-api. GitHub repository. Accessed: 3 August 2025.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chi-

nese CLIP: contrastive vision-language pretraining in chinese. *CoRR*, abs/2211.01335.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285.

A Comparison with Other Datasets

We compare 3XCM with existing VQA benchmarks in Table 4. The 3XCM dataset was developed to address the lack of cross-cultural parallelism in current resources.

Feature	MaRVL (Liu et al., 2021)	CulturalVQA (Nayak et al., 2024)
Primary Task	Visual Reasoning	Visual QA
Languages	5	11
Granularity	Specific indigenous	Broad cultural facets
Parallelism	Low (Culture-specific)	Low (Culture-specific)
Bias Target	Reasoning Failure	Knowledge Gaps
Data Availability	Public	Public

Feature	CVQA (Romero et al., 2024)	RAVENEA (Li et al., 2026)
Primary Task	Visual QA	VQA & Image Cap.
Languages	31	8 (Countries)
Granularity	Broad categories	Varies (Specific/Broad)
Parallelism	Low (Region-specific)	Low (VQA: Cult-Spec. IC: Too Broad)
Bias Target	Knowledge Gaps	Cultural Nuances
Data Availability	Public	To be released*

Feature	3XCM (Ours)
Primary Task	**Cross-Modal Retrieval**
Languages	**16**
Granularity	**Universal basic-level**
Parallelism	**High (Cross-cultural)**
Bias Target	**Association Bias**
Data Availability	**To be released***

Table 4: Comparison of 3XCM with existing cultural benchmarks. *Data to be released upon acceptance.

B Language Resources

To estimate language resource availability, we utilized the Distribution of Languages from the Common Crawl dataset (CC-MAIN-2025-18) (Common Crawl, 2025) as an approximation. Table 5 and 6 present the resulting language composition for RQ1 and RQ2, respectively.

Type	Language	ID	Distribution (%)
High	English	en	43.9499
	Russian	ru	5.7614
	German	de	5.5691
Medium	Japanese	ja	4.9152
	Chinese-Simpl.	zh	4.8778
	Spanish	es	4.5422
	French	fr	4.3271
	Italian	it	2.4060
	Portuguese	pt	2.3369
	Polish	pl	1.8744
	Dutch	nl	1.8083
	Indonesian	id	1.1759
	Turkish	tr	1.1274
	Czech	cs	1.0479
	Vietnamese	vi	1.0213
Low	Korean	ko	0.7865
	Farsi	fa	0.7087
	Swedish	sv	0.6736
	Arabic	ar	0.6722
	Romanian	ro	0.6374
	Ukrainian	uk	0.6079
	Greek	el	0.5651
	Hungarian	hu	0.5082
	Danish	da	0.4792
	Thai	th	0.4269
	Finnish	fi	0.3649
	Norwegian	no	0.3135
	Hebrew	he	0.2654
	Croatian	hr	0.2339
	Hindi	hi	0.2004
	Bengali	bn	0.1064
	Telugu	te	0.0213
	Swahili	sw	0.0102
	Filipino	fil	0.0084
	Maori	mi	0.0014
	Cusco Quechua	quz	0.0005

Table 5: Composition of Language Resources in the CommonCrawl Dataset (CC-MAIN-2025-18) for the language experimented in RQ1

C Evaluated Models and Details

In this work, we assess three distinct architectural paradigms for multimodal retrieval. The specific characteristics of these families are as follows:

- **Vision-Language Contrastive Models:** These are foundational models trained

Code	Country	Language	Resources (%)
US GB AU	America Great Britain Australia	English	43.950
DE CN JP	Germany China Japan	German Chinese Japanese	5.569 4.878 4.915
ES AR	Spain Argentina	Spanish	4.542
FR	France	French	4.327
PT BR	Portugal Brazil	Portuguese	2.337
SA TH IN KE NG	Saudi Arabia Thailand India Kenya Nigeria	Arabic Thai Hindi Swahili Yoruba	0.672 0.427 0.200 0.010 0.001

Table 6: Composition of Language Resources in the CommonCrawl Dataset (CC-MAIN-2025-18) for the language experimented in RQ2

primarily on English data. We include the original CLIP-L/14 as a powerful baseline, and CN-CLIP-L/14 to observe the effect of monolingual fine-tuning on a non-English corpus.

- **Cross-lingual Alignment Models:** These models use knowledge distillation to explicitly align multilingual text encoders to a fixed, pre-trained vision space for multilingual capability. We evaluate two variants of m-CLIP, which use XLM-RoBERTa as the text encoder (XLM-R-L/14 and XLM-R-B/16plus).
- **MLLM-Based Retrieval Embedders:** This modern paradigm adapts large, pre-trained Multimodal Language Models for retrieval. We evaluate several state-of-the-art models, including the ColQwen series (v0.2, 3b-M, 7b-M), GME models (Qwen2-2B, Qwen2-7B), and Jina-E-v4.

Table 7 provides the precise aliases, full model names, parameter counts, and citations used throughout the paper.

D Details on LBKL and DLBKL

In Section 3.1, we utilize LBKL and DLBKL to measure prevalence bias. Here, we elaborate on the rationale behind these metrics.

In an ideal, bias-free retrieval scenario, the distribution of languages in the retrieved list should match the expected "fair" distribution (e.g., the ground truth distribution). LBKL measures the Kullback–Leibler divergence between these two distributions. In equation 1, $P(x)$ represents the proportion of a specific language in the ground truth, and $Q(x)$ represents the proportion in the retrieved set.

However, standard LBKL treats a deviation at rank 1 the same as a deviation at rank 100. In multimodal retrieval, users rarely scroll far, making top-rank bias significantly more harmful. Our proposed DLBKL in equation 2 addresses this by applying a discount factor derived from NDCG (Järvelin and Kekäläinen, 2002). By weighting the observed proportion $Q(x)$ by the inverse log of the rank, we ensure that high-resource language dominance at the top of the list results in a significantly higher divergence score than if that same dominance occurred lower in the list.

E Formulation of Language-wise LBKL and DLBKL

In this section, we detail the calculation of two bias metrics: Language-wise LBKL (Country-LBKL) and Language-wise DLBKL (Country-DLBKL). Both metrics quantify the divergence between the retrieved country distribution and an ideal uniform distribution. The key distinction lies in how the predicted distribution is constructed: Country-LBKL treats the retrieved documents as an unweighted set, while Country-DLBKL applies rank-based discounting to prioritize higher-ranked results.

E.1 Ground Truth Distribution (Shared)

Let $C = \{c_1, c_2, ..., c_N\}$ be the set of target countries, where $N = |C|$. We define the ground truth distribution P as:

$$P(c_j) = \frac{1}{N}, \quad \forall c_j \in C \tag{6}$$

E.2 Country-LBKL (Unweighted)

Country-LBKL evaluates bias based on the frequency of countries in the top-k retrieved documents $D = [d_1, ..., d_k]$, treating all rank positions equally.

E.2.1 Predicted Distribution (Unweighted)

The unweighted proportion $Q(c_j)$ is calculated as:

$$Q(c_j) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{c_j}(d_i) \tag{7}$$

Alias Used in Paper	Full Model Name	Parameter
Vision-Language Contrastive Models		
CLIP-L/14	`clip-vit-large-patch14`[1]	427.6M
CN-CLIP-L/14	`Chinese-clip-vit-large-patch14`[2]	406.2M
Cross-lingual Alignment Models		
XLM-R-L/14	`XLM-Roberta-Large-Vit-L-14`[6]	998.3M
XLM-R-L/16plus	`XLM-Roberta-Large-Vit-B-16Plus`[6]	768.9M
MLLM-Based Retrieval Embedders		
ColQwen2.5-v0.2	`ColQwen2.5-v0.2`[3]	3814.8M
ColQwen2.5-3B-M	`ColQwen2.5-3b-multilingual-v1.0`[3]	3994.6M
ColQwen2.5-7B-M	`ColQwen2.5-7b-multilingual-v1.0`[3]	8071.1M
GME-Qwen2-2B	`gme-Qwen2-VL-2B-Instruct`[4]	2209.0M
GME-Qwen2-7B	`gme-Qwen2-VL-7B-Instruct`[4]	7070.6M
Jina-E-v4	`jina-embeddings-v4`[5]	3934.7M

The models are based on the following works: 1) Radford et al. (2021) for CLIP-L/14; 2) Yang et al. (2022) for CN-CLIP-L/14; 3) Faysse et al. (2025) for ColQwen2 models; 4) Zhang et al. (2025) for GME-Qwen2 models; 5) Günther et al. (2025) for Jina-E-v4; and 6) Carlsson et al. (2022) for XLM-R-VL models.

Table 7: Aliases Used in Paper and Corresponding Full Model Names and Parameters

where the indicator function $\mathbb{1}_{c_j}(d_i)$ is:

$$\mathbb{1}_{c_j}(d_i) = \begin{cases} 1 & \text{if } d_i \text{ is from } c_j \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

E.2.2 Score Calculation

The Country-LBKL score is the KL divergence between P and Q. We simplify the substitution of $P(c_j)$ as follows:

$$\begin{aligned} \text{Country-LBKL} &= \sum_{c_j \in C} P(c_j) \log\left(\frac{P(c_j)}{Q(c_j)}\right) \\ &= \frac{1}{N} \sum_{c_j \in C} \log\left(\frac{1}{N \cdot Q(c_j)}\right) \end{aligned} \tag{9}$$

E.3 Country-DLBKL (Rank-Weighted)

Country-DLBKL assigns higher importance to documents appearing earlier in the list using logarithmic discounting.

E.3.1 Rank-Based Weighting

For rank position i, the weight w_i is:

$$w_i = \frac{1}{\log_2(i+1)} \tag{10}$$

The total weight normalization factor is:

$$W_{total} = \sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{1}{\log_2(i+1)} \tag{11}$$

E.3.2 Predicted Distribution (Weighted)

The weighted distribution $Q'(c_j)$ is calculated as:

$$Q'(c_j) = \frac{\sum_{i=1}^{k} w_i \cdot \mathbb{1}_{c_j}(d_i)}{W_{total}} \tag{12}$$

E.3.3 Score Calculation

The Country-DLBKL score is the KL divergence between P and the weighted distribution Q':

$$\begin{aligned} \text{Country-DLBKL} &= \sum_{c_j \in C} P(c_j) \log\left(\frac{P(c_j)}{Q'(c_j)}\right) \\ &= \frac{1}{N} \sum_{c_j \in C} \log\left(\frac{1}{N \cdot Q'(c_j)}\right) \end{aligned} \tag{13}$$

E.4 Implementation Note

For both metrics, a small smoothing value is applied to $Q(c_j)$ and $Q'(c_j)$ in implementation to ensure numerical stability when a country is not represented in the retrieval results.

F Example of Prevalence Bias Evaluation

To further elaborate on the result of research question 1, we provide the example of retrieval from image to text from CLIP and M-CLIP in Figure 8. We also provide the results of LBKL and DLBKL scores at other ranks in Table 8.

G Language and Rank Frequency Diagram

To illustrate bias in image-to-text retrieval, we present a visualization of language groups categorized by resource level as shown in Appendix B, showing both their overall retrieval frequency as shown in Figure 9 and their frequency distribution across ranks as shown in Figure 10.

Image Query:

ID: c4c286b83715da59

Caption (English):

"A night view of an ancient building with lights.",

"A well lit huge stone gateway with an arch at night."

CLIP (clip-vit-large-patch14) Retrieval Results LBKL@5: 15.508, **DLBKL@5: 15.518**

Language counts: {'fr': 1, 'nl': 2, 'pt': 2} → Proportions: [1, 2, 2]

Rank	Similarity score	Correct	Language (Resource)	Caption
1	0.2728	Yes	Dutch (Med)	Arc the Triomphe.
2	0.2725	Yes	Dutch (Med)	Verlichte arc de triomphe 's nachts
3	0.2688	Yes	Portuguese (Med)	Arco do triunfo em paris
4	0.2687	Yes	Portuguese (Med)	O Arco do Triunfo de Paris iluminado a noite.
5	0.2570	Yes	French (Med)	Arc de triomphe à Paris, illuminé, de nuit

M-CLIP (XLM-Roberta-Large-Vit-B-16Plus) Retrieval result LBKL@5: 15.508, **DLBKL@5: 15.514**

Language counts: {'da': 2, 'he': 1, 'sv': 2} → Proportions: [1, 2, 2]

Rank	Similarity score	Correct	Language	Caption
1	0.3738	Yes	Danish (Low)	Triumfbuen i paris om natten
2	0.3687	Yes	Swedish (Low)	Triumfbågen i Paris upplyst på kvällen.
3	0.3664	Yes	Danish (Low)	Triumfbuen i Paris lyst op om aftenen
4	0.3655	Yes	Hebrew (Low)	שער הניצחון בפריז בלילה. מסביב יש נורות רחוב דולקות ועצים ואין אנשים.
5	0.3646	Yes	Swedish (Low)	Triumfbågen i Paris upplyst på natten.

Figure 8: Qualitative Comparison of Image-to-Text Retrieval in Crossmodal-3600. Two retrieval results share identical semantic proportions and LBKL scores. Despite this semantic equality, the DLBKL score differs, capturing the model's implicit preference for specific linguistic groups in the ranking process.

Model	@5		@25		@50		@99	
	LBKL↓	DLBKL↓	LBKL↓	DLBKL↓	LBKL↓	DLBKL↓	LBKL↓	DLBKL↓
Vision-Language Contrastive Models								
CLIP-L/14	15.846	15.848	14.706	14.711	14.158	14.162	13.629	13.631
CN-CLIP-L/14	15.784	15.787	14.419	14.427	13.680	13.688	12.968	12.974
Cross-lingual Alignment Models								
XLM-R-L/14	14.710	14.718	8.547	8.583	4.759	4.800	2.386	2.413
XLM-R-B/16plus	14.654	14.662	8.248	8.285	4.345	4.388	2.007	2.036
MLLM-Based Retrieval Embedders								
ColQwen2.5-3B-M	15.010	15.016	11.231	11.253	9.326	9.347	7.839	7.846
ColQwen2.5-7B-M	14.947	14.954	10.585	10.610	8.316	8.339	6.750	6.755
ColQwen2.5-v0.2	15.340	15.346	12.176	12.199	10.561	10.586	9.315	9.331
GME-Qwen2-2B	15.144	15.151	10.457	10.494	7.664	7.710	5.322	5.368
GME-Qwen2-7B	15.035	15.042	9.766	9.805	6.503	6.552	4.024	4.068
Jina-E-v4	14.781	14.789	9.320	9.354	6.079	6.120	3.734	3.769

Table 8: Image-to-text retrieval bias on Crossmodal-3600, measured by LBKL and DLBKL at various retrieval depths (k).

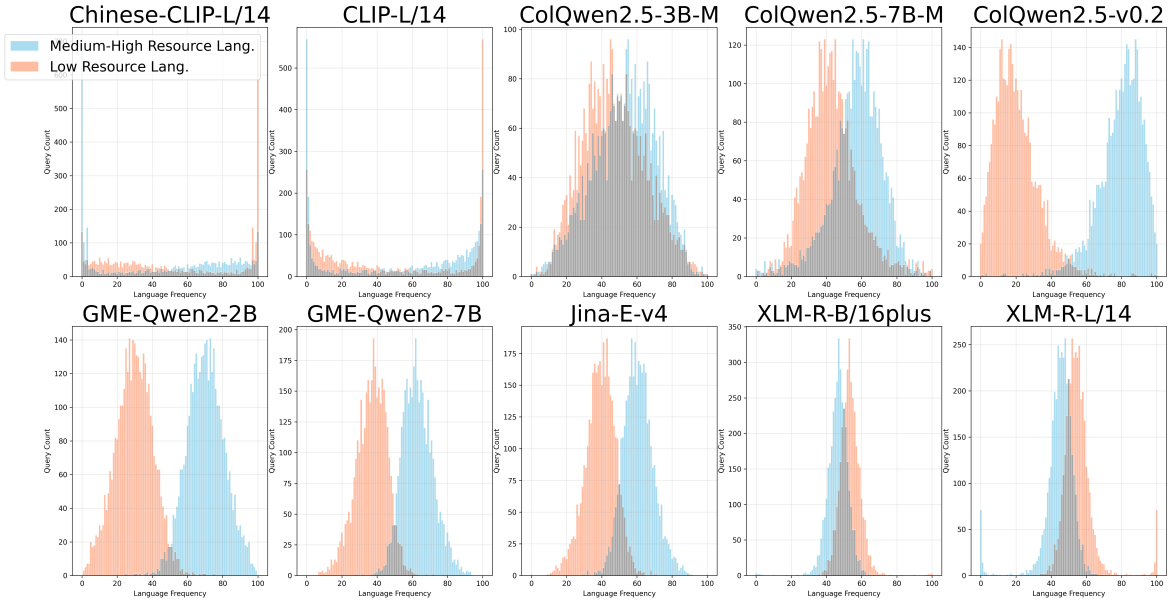

Figure 9: A language frequency histogram of each language group (see Appendix B) for all models

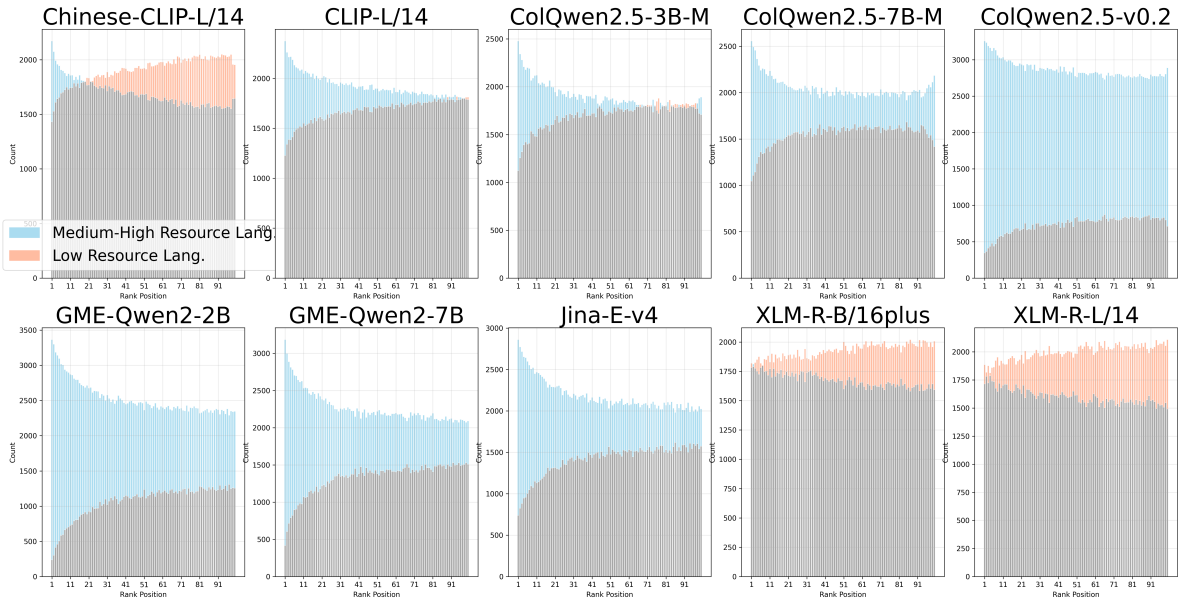

Figure 10: A frequency of language group (see Appendix B) at each rank for all model

H Culturally Relevant Concept Identification

To identify image concepts unique to each country, we first employed the Gemini[2] as a tool for generating culturally relevant suggestions prior to data collection. The prompt used in this process is shown in Figure 11.

```
list 100 concepts that unique and vary in these country
including China, India, Japan, saudi arabia, France,
German, Brazil, Kenya, Thailand, USA

like this

{"food":{"China":"Mala Xiang Guo", ..., "Thailand":
"Padthai", "USA":"hamburger"}, "costume":{"China":
"Hanfu", ..., "Thailand":"Sabai", "USA":"cowboy"}}
```

Figure 11: The prompt given to Gemini to generate unique country-specific image concepts.

I Multi-Concept Detection in Images

To establish the self-preference cultural bias, the culturally relevant and non-relevant candidate images must *not* share concepts with the text label. We utilized Gemini[2] with the following prompt in Figure 12 to identify all labels associated with an image.

> Please classify the following image by assigning them to one or more of the following cultural categories:
> {category}.
>
> **Comprehensive Output Format (JSON):** output as JSON for example:
>
> ```
> {{
>  "<INDEX>": ["<CATEGORY>", "<CATEGORY>", ...],
> }}
> ```
>
> in the categories please order by priority (high to low).

Figure 12: The prompt given to Gemini for multi-label image classification, where the {category} placeholder is dynamically populated with the full list of categories.

J Annotator Guideline

The guideline we provide to the annotators is to remove duplicates across multiple views. If an image depicts the same scene or object with no meaningful change, keep only one copy. Keep images if there is a significant variation. Allowed differences include time of day (e.g., day vs. night), viewpoint or angle (if the perspective changes enough that visual elements in the image are noticeably different). Minor or trivial variations are *not* allowed as they would be too similar. This 397 includes slight shifts, crops, or zooms of the same scene.

K Dataset Statistics

The distribution of cultural concepts in the 3XCM dataset is shown in Table 10, with each concept being represented by approximately 85 images on average.

Country	Samples	Country	Samples
Argentina	771	Kenya	600
Australia	721	Nigeria	773
Brazil	724	Portugal	824
China	727	Saudi Arabia	619
France	760	Spain	841
Germany	744	Thailand	649
India	774	UK	644
Japan	944	USA	609
Average		**732.75**	

Table 9: Dataset Statistics of 3XCM per Country

Concepts (A-G)	
airlines	24
airport	65
alcohol drink	26
ancient city	112
ancient craft	111
ancient painting	139
animal	100
architecture	107
art	89
artwork	26
bag	20
bakery	47
banknotes	69
bathroom	30
bedroom	77
boat	99
bracelet	62
building	196
bus	92
bus station	56
capital	147
celebrity	66
child	69

Concepts (C-M)	
cinema	54
coin	182
combination food	72
congress	127
costume	119
craft	98
dance	145
deep fried food	30
department store	21
dessert	66
devil	43
diningroom	49
doll	131
drink	31
dry heat food	21
embroidery style	133
fashion	57
festival	145
fire station	162
folk tale	38
folklore character	90
food	52
football player	104

Concepts (F-M)	
formal uniform	117
fountain style	91
funeral	141
game	76
gas station	63
gathering place	92
ghost	19
graduated uniform	78
hat	95
headwear	106
historical event	89
historical figure	81
historical image	120
hot pot concept	66
hotel	70
house	86
instrument	58
lottery tickets	48
mailbox	81
major mountain range	52
major religious site	97
major river	57
map	32
market	74

Concepts (M-P)	
marriage ceremony	95
martial art	96
mask	104
military parade	189
moist heat food	34
museum	99
music band	95
mythical creature	56
mythological figure	68
native inhabitants	163
natural landmark	152
necklace	63
night view	110
older	38
painting	128
pants	30
people	136
poaching food	25
police station	158
popular street food	98
pottery style	150
priest	68
prime minister	86

Concepts (R-S)	
religious building	123
restaurant	38
ritual	108
rural dwelling	127
sacred object	101
school	120
series	40
shirt	64
shopping mall	91
singer	56
snack	56
social custom	147
soldier	167
sport	79
stageplay	108
statue	106
street entertainment	111
street sign	81
street vendor cart	32
street view	86
symbolic bird	84
symbolic plant	72

Concepts (T-Z)	
tattoo style	59
taxi	102
tea culture	85
textile pattern	56
tourist attraction	130
toy	66
train	116
train station	128
tree	94
tv program	43
unique art form	119
unique cuisine trait	43
unique food ingredient	22
unique natural phenomenon	102
unique transportation	75
university	136
wall painting	65
warrior	79
weapon	75
wedding	108
writing character	15
zoo	79

Table 10: Distribution of Concepts and Image Counts

L 3XCM Dataset Benchmark Samples

This research provides an association evaluation benchmark and image metadata. Examples of the benchmark for RQ2, RQ3, and image metadata are shown in Figure 13, 14, and 15 respectively.

Figure 13: Benchmark examples from the 3XCM dataset featuring three retrieval candidates.

Figure 14: Benchmark examples from the 3XCM dataset featuring six retrieval candidates.

M UMAP Analysis for Self-Preference Cultural Bias

To visualize cultural bias, the UMAP projections of text and image embeddings of all models as shown in Figure 16 and 17. The text embeddings cluster strongly by language, a proximity that supersedes semantic content. Conversely, the image embeddings do *not* exhibit strong country-based clustering, suggesting lower cultural bias. While other models show a similar, albeit less severe, tendency for text embeddings to be more biased than image embeddings, this effect is diminished in modern models. The GME-Qwen2 and Jina-E-v4 models only cluster very low-resource languages (Swahili, Yoruba), and the XLM-R models demonstrate superior alignment, forming a single central cluster. This discrepancy challenges retrieval systems: a query's text embedding is biased by its language, leading the system to favor images from the same cultural context over potentially more visually relevant content from others.

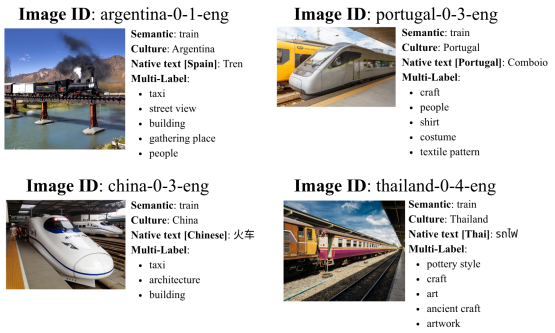

Figure 15: Metadata for image of 3XCM dataset benchmark

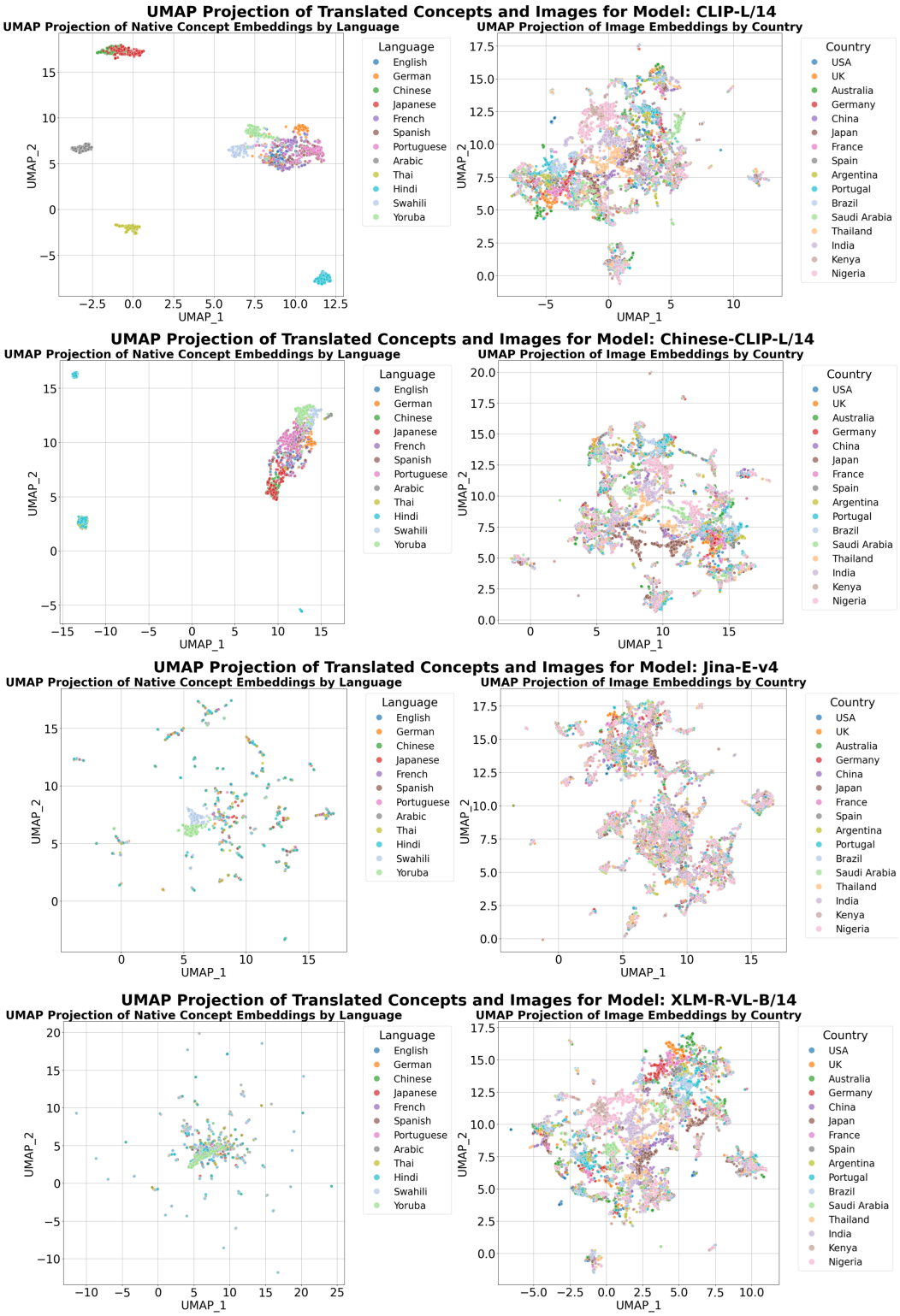

Figure 16: The UMAP visualizations of the caption embeddings (left) and image embeddings (right) from the CLIP-L/14 model applied to our dataset.

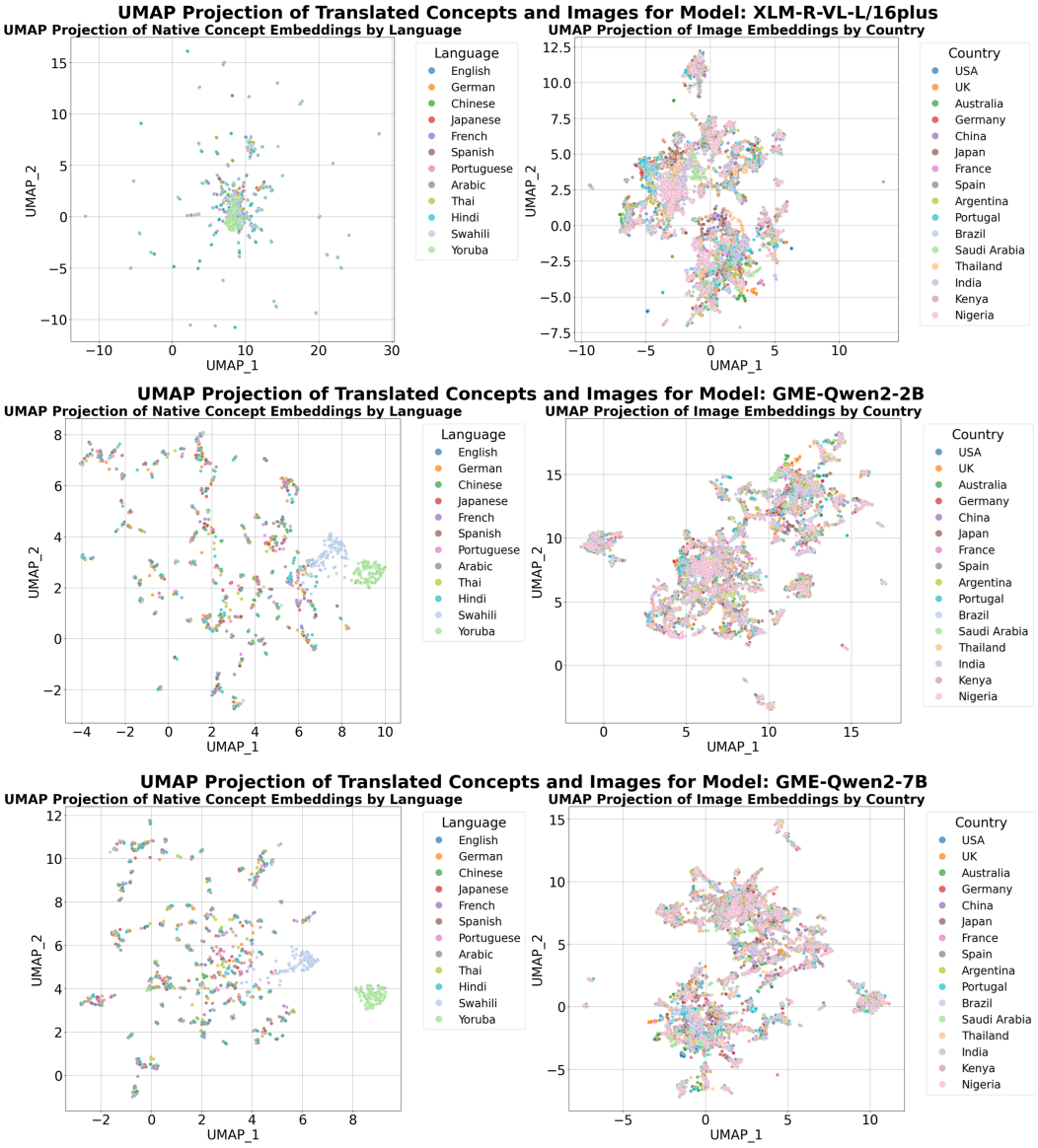

Figure 17: The UMAP visualizations of the caption embeddings (left) and image embeddings (right) from the CN-CLIP-L/14 model applied to our dataset.

N Correlation Analysis for Self-Preference Cultural Bias

We investigate how unimodal bias, which our UMAP analysis shows is more severe in the text modality as shown in Appendix M, impacts cross-modal retrieval. To quantify this, we use the Silhouette score and find that high scores in low-resource languages correlate with the self-preference cultural bias score (SP), as shown in Table 12. We confirm this relationship by calculating the Pearson correlation between SP and the Silhouette scores. For example, CLIP-L/14's Text Silhouette score correlates strongly with SP (0.827), while its Image Silhouette correlation is only moderate (0.550), as shown in Figure 18. Across all tested models, the average correlations reveal that SP is predominantly driven by the text encoder, as shown in Table 11.

Model	TSC	ISC
CLIP-L/14	0.83	0.55
CN-CLIP-L/14	-0.26	0.58
XLM-R-VL-B/16	0.98	-0.27
XLM-R-VL-L/14	0.94	0.05
Jina-E-v4	0.86	0.16
GME-Qwen2-2B	0.80	0.09
GME-Qwen2-7B	0.64	0.07
Average	**0.68**	**0.18**

Table 11: This table presents a Pearson correlation analysis between model performance and bias. We measure the correlation between association and the quality of data clusters (via Silhouette Score) in both the text embedding space (TSC) and the image embedding space (ISC).

O Detailed Results

The full details of the RQ2 and RQ3 experiments, including all the win rates of all models, are illustrated in the Table 13 and 14, respectively.

P Analysis of Comparative Association Bias in Text-to-Image Retrieval

To rigorously validate the failure modes and perspectival biases identified in RQ2 and RQ3, we employed a statistical evaluation targeting retrieval performance, embedding confusion, and the impact of query augmentation.

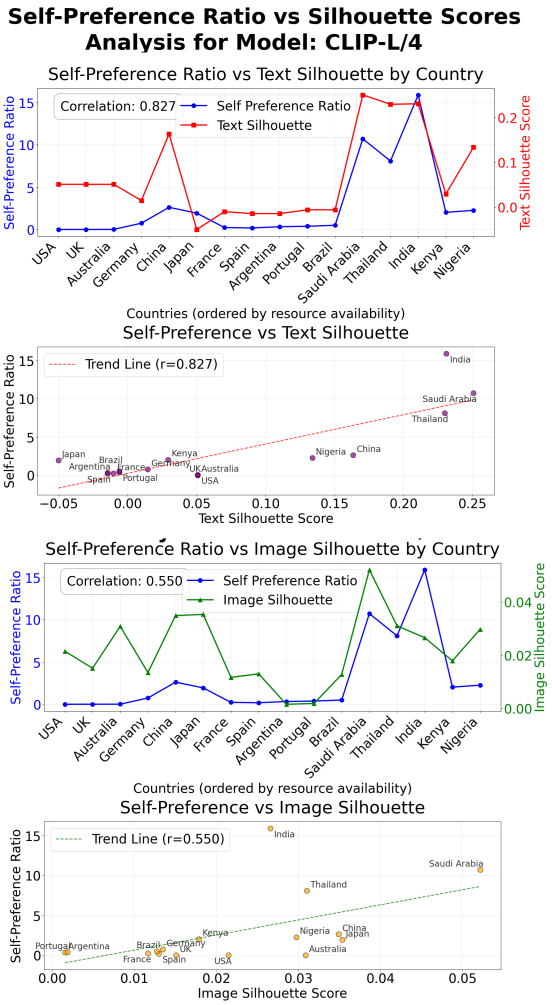

Figure 18: Comparison of Self-Preference Cultural Bias with Text and Image Silhouette

P.1 Rank-1 Win Rate Analysis

Complementing the continuous score analyses, we evaluated the model's explicit retrieval preference at the top rank using a *Chi-Square (χ^2) Goodness of Fit Test*. This approach determines whether the model exhibits a statistically significant tendency to select one specific failure mode over another when the ground truth is not prioritized. We defined the Null Hypothesis as the condition where the model is equally likely to retrieve either of the two competing distractor categories ($P(A) = P(B)$). To compute the statistic, the expected counts (E) for the two competing groups were derived by averaging their combined observed wins ($E_A = E_B = \frac{O_A + O_B}{2}$), while the count of unrelated or ground truth retrievals remained fixed as observed values. The resulting χ^2 statistic, calculated with 1 degree of freedom, identifies whether the observed disparity in win rates is a result of systematic algorithmic bias.

Model	Metrics	Country															
		US	GB	AU	DE	CN	JP	FR	ES	AR	PT	BR	SA	TH	IN	KE	NG
CLIP-L/14	SP ↓	0.01	0.02	0.02	0.76	2.63	1.94	0.24	0.19	0.34	0.39	0.51	10.71	8.09	15.88	2.04	2.27
	TS ↓	0.05	0.05	0.05	0.02	0.16	-0.05	-0.01	-0.02	-0.02	-0.01	-0.01	0.25	0.23	0.23	0.03	0.13
	IS ↓	0.02	0.02	0.03	0.01	0.03	0.04	0.01	0.01	0.00	0.00	0.01	0.05	0.03	0.03	0.02	0.03
CN-CLIP-L/14	SP ↓	0.03	0.06	0.05	1.11	0.03	0.13	0.33	0.40	0.43	0.56	0.52	4.88	2.55	6.98	1.99	2.23
	TS ↓	-0.05	-0.05	-0.05	0.00	0.13	-0.07	-0.03	-0.03	-0.03	-0.01	-0.01	-0.40	0.93	-0.37	0.00	0.03
	IS ↓	0.02	0.02	0.03	0.01	0.03	0.03	0.01	0.01	0.00	0.00	0.00	0.04	0.05	0.04	0.02	0.01
Jina-E-v4	SP ↓	0.02	0.03	0.02	0.05	0.04	0.04	0.04	0.03	0.05	0.04	0.04	0.05	0.02	0.12	0.49	0.93
	TS ↓	-0.01	-0.01	-0.01	-0.02	0.01	0.00	-0.01	-0.02	-0.02	-0.02	-0.02	0.02	0.01	0.00	0.00	0.13
	IS ↓	-0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	0.00	-0.01	-0.01	-0.01	0.02	0.00	0.01	0.01	0.00
XLM-R-L/14	SP ↓	0.05	0.04	0.07	0.05	0.04	0.07	0.08	0.08	0.09	0.08	0.07	0.05	0.04	0.04	0.11	0.61
	TS ↓	-0.18	-0.18	-0.18	-0.11	-0.07	-0.10	-0.16	-0.16	-0.16	-0.12	-0.12	-0.07	-0.06	-0.17	-0.14	0.44
	IS ↓	0.01	0.02	0.03	0.01	0.03	0.03	0.01	0.01	0.01	0.01	0.01	0.06	0.04	0.03	0.02	0.03
XLM-R-B/16plus	SP ↓	0.05	0.04	0.06	0.04	0.04	0.05	0.05	0.06	0.07	0.06	0.06	0.04	0.02	0.05	0.10	0.66
	TS ↓	-0.24	-0.24	-0.24	-0.21	-0.19	-0.19	-0.24	-0.23	-0.23	-0.23	-0.23	-0.19	-0.19	-0.23	-0.23	0.49
	IS ↓	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	-0.01	0.00	-0.01	0.03	0.02	0.03	0.02	0.00
GME-Qwen2-2B	SP ↓	0.02	0.03	0.02	0.10	0.04	0.04	0.05	0.05	0.08	0.08	0.09	0.14	0.10	0.24	1.24	2.02
	TS ↓	0.02	0.02	0.02	0.04	0.02	0.00	0.04	0.01	0.01	0.03	0.03	0.01	0.03	0.00	0.03	0.15
	IS ↓	0.00	0.00	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.02	0.01	0.01
GME-Qwen2-7B	SP ↓	0.03	0.02	0.01	0.14	0.04	0.05	0.07	0.07	0.07	0.09	0.13	0.14	0.08	0.12	0.62	1.88
	TS ↓	0.04	0.04	0.04	0.05	0.02	0.02	0.07	0.02	0.02	0.04	0.04	0.09	0.11	0.03	0.04	0.14
	IS ↓	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.01	0.01	0.01

Table 12: Cross-Country and Cross-Model Comparison of Language and Cultural Bias Metrics. This table presents the results for the Self-Preference Cultural Bias score (SP), Text Silhouette (TS), and Image Silhouette (IS) scores across various multimodal retrievers for a selection of countries.

P.2 Test Pairs

We conducted statistical analysis on specific pairs of candidate images, identified by the color codes in Figure 4, to isolate distinct levels of bias. The pairs were selected to evaluate how the model resolves conflicts between semantic content, explicit cultural descriptors, and implicit language priors.

- **Query Language Association Bias Test (Object-Relevant Language Bias (Purple) vs. Object Relevant (Blue))**
  This comparison investigates association bias in scenarios where the model successfully retrieves the correct semantic object (e.g., a train) but fails to adhere to the explicit cultural descriptor. By comparing a candidate that aligns with the query's language culture (Purple) against a random candidate (Blue), we test whether the script of the query exerts a stronger influence on visual selection than the explicit textual description.

- **Cultural Descriptor Relevance (Yellow) vs. Language Bias (Pink):**
  This comparison analyzes the model's prioritization mechanism during *Semantic Failure*. In cases where the model fails to retrieve the correct object entirely, we determine whether the retrieval is driven by a specific token match (e.g., the word "Japanese" leading to Japanese food; Yellow) or by the latent prior of the query language itself (e.g., Thai script leading to Thai cultural imagery; Pink). This reveals whether the error stems from specific keyword fixation or broad language-modality leakage.

P.3 Statistical Test Results

Table 15 summarizes the overall results of the statistical significance tests. For a breakdown of each test, Tables 16 and 17 provide the detailed p-values for each model across all evaluated countries. In the **Query Language Association Bias Test**, the majority of architectures, specifically the Vision-Language Contrastive models and LLM-Based Retrieval Embedding models, exhibited a statistically significant preference for the *Object-Relevant Language Bias* candidate (Purple) over the neutral alternative (Blue). This consistent rejection of the null hypothesis for these models confirms the presence of a strong association bias, demonstrating that the script of the input query affects the retrieved image result. However, the Cross-lingual Alignment

Model	Metrics	Country															
		US	GB	AU	DE	CN	JP	FR	ES	AR	PT	BR	SA	TH	IN	KE	NG
CLIP-L/14	$M_{cr}(\%) \uparrow$	95.73	94.57	94.73	52.42	25.17	31.07	75.53	78.00	69.65	65.78	61.33	7.75	10.48	5.56	27.83	24.19
	$M_{lb}(\%) \downarrow$	1.31	2.02	2.22	39.65	66.16	60.34	18.29	15.10	23.61	25.85	31.35	83.04	84.75	88.24	56.67	54.85
	$M_{ti}(\%) \downarrow$	2.96	3.42	3.05	7.93	8.67	8.59	6.18	6.90	6.74	8.37	7.32	9.21	4.78	6.20	15.50	20.96
CN-CLIP -L/14	$M_{cr}(\%) \uparrow$	94.25	89.44	90.57	40.05	93.40	84.20	65.79	61.95	60.83	56.55	57.04	14.38	21.88	10.47	27.00	25.10
	$M_{lb}(\%) \downarrow$	3.28	5.75	4.44	44.62	2.48	10.92	21.97	24.73	26.07	31.80	29.56	70.11	55.78	73.00	53.83	56.02
	$M_{ti}(\%) \downarrow$	2.46	4.81	4.99	15.32	4.13	4.88	12.24	13.32	13.10	11.65	13.40	15.51	22.34	16.54	19.17	18.89
XLM-R-L/14	$M_{cr}(\%) \uparrow$	89.66	91.15	89.46	90.32	92.30	89.40	87.63	85.37	86.12	87.86	87.98	90.95	93.07	89.41	80.33	40.49
	$M_{lb}(\%) \downarrow$	4.11	4.04	5.83	4.44	4.13	6.26	6.58	6.54	7.39	7.16	5.80	4.85	3.54	3.75	8.50	24.71
	$M_{ti}(\%) \downarrow$	6.24	4.81	4.72	5.24	3.58	4.35	5.79	8.09	6.49	4.98	6.22	4.20	3.39	6.85	11.17	34.80
XLM-R-B /16Plus	$M_{cr}(\%) \uparrow$	91.95	91.30	90.71	93.95	94.50	91.62	90.66	88.47	87.81	90.05	89.36	92.57	95.22	91.09	83.17	40.88
	$M_{lb}(\%) \downarrow$	4.11	3.88	5.13	4.03	3.58	4.56	4.34	5.47	5.97	5.70	5.52	3.88	2.00	4.91	8.33	26.78
	$M_{ti}(\%) \downarrow$	3.94	4.81	4.16	2.02	1.93	3.82	5.00	6.06	6.23	4.25	5.11	3.55	2.77	4.01	8.50	32.34
ColQwen2.5 -3B-Multi	$M_{cr}(\%) \uparrow$	95.24	93.79	95.01	83.20	94.91	90.99	89.61	89.06	87.94	89.44	89.09	87.40	90.60	81.65	45.67	27.30
	$M_{lb}(\%) \downarrow$	3.28	2.95	2.77	11.69	2.75	5.20	6.18	6.54	7.13	6.07	6.49	8.72	7.40	13.44	32.83	48.64
	$M_{ti}(\%) \downarrow$	1.48	3.26	2.22	5.11	2.34	3.82	4.21	4.40	4.93	4.49	4.42	3.88	2.00	4.91	21.50	24.06
ColQwen2.5 -7B-Multi	$M_{cr}(\%) \uparrow$	96.55	96.74	96.81	84.27	95.32	91.73	90.13	88.35	84.82	88.71	88.67	86.75	91.99	86.82	42.17	32.08
	$M_{lb}(\%) \downarrow$	1.31	1.55	1.66	11.56	2.34	6.47	7.76	9.39	10.25	8.01	8.29	12.12	6.78	8.91	37.83	49.68
	$M_{ti}(\%) \downarrow$	2.13	1.71	1.53	4.17	2.34	1.80	2.11	2.26	4.93	3.28	3.04	1.13	1.23	4.26	20.00	18.24
ColQwen2.5 -v0.2	$M_{cr}(\%) \uparrow$	93.10	89.60	92.09	83.60	94.09	88.55	90.79	87.51	86.64	90.29	90.61	85.14	88.60	79.59	38.83	29.88
	$M_{lb}(\%) \downarrow$	4.11	4.97	3.47	9.95	3.16	6.68	5.00	7.37	6.87	5.22	4.83	8.72	8.94	13.82	42.83	42.95
	$M_{ti}(\%) \downarrow$	2.79	5.43	4.44	6.45	2.75	4.77	4.21	5.11	6.49	4.49	4.56	6.14	2.47	6.59	18.33	27.17
GME-Qwen2 -2B-Instruct	$M_{cr}(\%) \uparrow$	96.06	95.34	96.95	87.77	94.09	93.43	92.76	90.49	88.59	90.05	88.95	84.49	89.06	76.87	37.17	25.87
	$M_{lb}(\%) \downarrow$	2.30	2.48	1.66	8.47	3.58	3.92	4.74	4.64	7.26	7.04	7.87	11.63	8.63	18.48	46.00	52.26
	$M_{ti}(\%) \downarrow$	1.64	2.17	1.39	3.76	2.34	2.65	2.50	4.88	4.15	2.91	3.18	3.88	2.31	4.65	16.83	21.86
GME-Qwen2 -7B-Instruct	$M_{cr}(\%) \uparrow$	95.40	95.34	96.53	84.68	95.05	92.47	90.39	90.73	90.27	88.35	85.50	85.46	90.91	85.92	55.17	29.62
	$M_{lb}(\%) \downarrow$	2.46	2.33	0.97	11.69	3.44	4.77	6.45	5.95	6.10	7.77	11.46	11.63	7.09	10.34	34.00	55.76
	$M_{ti}(\%) \downarrow$	2.13	2.33	2.50	3.63	1.51	2.76	3.16	3.33	3.63	3.88	3.04	2.91	2.00	3.75	10.83	14.62
Jina-E-v4	$M_{cr}(\%) \uparrow$	95.89	95.65	95.98	92.47	94.22	92.79	94.34	93.22	91.70	91.99	94.20	91.60	95.38	86.30	56.50	36.74
	$M_{lb}(\%) \downarrow$	1.97	2.48	1.66	4.97	3.44	3.92	4.08	3.09	4.80	4.13	3.73	4.68	2.00	9.95	27.83	34.15
	$M_{ti}(\%) \downarrow$	2.13	1.86	2.36	2.55	2.34	3.29	1.58	3.69	3.50	3.88	2.07	3.72	2.62	3.75	15.67	29.11

Table 13: Cross-Country and Cross-Model Comparison of Win Percentages. This table presents the results for the Correct, Language-biased, and Totally Irrelevant Win Percentages across various multimodal retrievers for a selection of countries.

models (XLM-R) served as an exception, showing no statistically significant difference ($p > 0.05$) between the candidates, suggesting these models are less susceptible to implicit script bias when object relevance remains constant. Conversely, the **Cultural Descriptor vs. Language Bias** comparison reveals a divergence in model behavior during semantic failure. The LLM-Based Retrieval Embeddings models and Cross-lingual Alignment models displayed a statistically significant preference for the *Cultural Descriptor Relevance* candidate (Yellow), indicating a general tendency to prioritize specific explicit tokens (e.g., "Japanese") over the implicit language prior. Notably, standard CLIP-L/14 exhibited a statistically significant preference for the *Language-biased category* (Pink) over the Cultural Descriptor (Diff < 0), suggesting that within CLIP's embedding space, the implicit bias derived from the query language script outweighs the explicit semantic cultural tokens.

Q Similarity Drift

Q.1 Similarity Drift Significance

Table 18 presents the detailed quantitative breakdown of the Similarity Drift (Δsim) analysis across all evaluated languages. In general, adding a cultural descriptor results in a positive similarity shift across most models, validating that explicit cultural context helps narrow the retrieval scope toward the target culture. The data reveals a critical failure mode in the baseline *CLIP-L/14* model regarding non-Latin scripts. While CLIP performs robustly for Western languages (e.g., English, German), it exhibits negligible drift for Chinese and Thai. This negative value implies that for certain script-heavy or linguistically complex contexts in

Model	Metrics	Country															
		US	**GB**	**AU**	**DE**	**CN**	**JP**	**FR**	**ES**	**AR**	**PT**	**BR**	**SA**	**TH**	**IN**	**KE**	**NG**
CLIP-L/14	$M_{cr}(\%)$	80.39	82.17	79.94	43.36	4.12	35.23	50.33	48.28	40.86	37.14	36.80	0.81	0.76	1.03	13.79	17.16
	$M_{orlb}(\%)$	8.33	6.05	7.47	33.56	51.92	31.12	28.52	32.94	42.28	38.83	40.36	51.05	58.38	53.81	42.86	23.35
	$M_{or}(\%)$	7.68	8.06	8.71	3.62	3.71	4.64	8.28	11.30	8.04	8.86	7.52	1.13	0.76	0.26	2.49	5.03
	$M_{cdr}(\%)$	3.10	3.72	3.60	10.07	3.16	12.76	7.88	4.64	5.06	6.80	6.84	0.81	0.00	0.90	8.97	20.65
	$M_{lb}(\%)$	0.49	0.00	0.28	8.72	36.13	15.08	3.94	2.38	2.98	7.77	7.52	46.05	38.87	43.61	30.73	27.48
	$M_{ti}(\%)$	0.00	0.00	0.00	0.67	0.96	1.16	1.05	0.48	0.78	0.61	0.96	0.16	1.22	0.39	1.16	6.32
CN-CLIP -L/14	$M_{cr}(\%)$	62.91	64.65	63.21	22.82	78.43	25.00	25.10	26.99	26.72	22.45	22.71	0.97	12.35	5.55	14.12	19.35
	$M_{orlb}(\%)$	18.63	16.59	16.74	34.77	2.47	48.73	41.13	31.03	33.72	34.59	33.93	53.62	17.99	36.77	43.69	32.77
	$M_{or}(\%)$	14.38	12.71	11.89	7.38	14.29	14.98	10.12	12.84	10.51	12.74	12.04	1.29	14.63	5.55	4.65	5.68
	$M_{cdr}(\%)$	3.59	4.65	6.92	12.89	4.67	2.43	7.88	11.77	11.80	11.53	10.40	0.64	18.29	10.19	8.47	16.00
	$M_{lb}(\%)$	0.33	0.62	0.55	17.85	0.00	7.81	11.70	12.60	13.10	14.08	14.36	42.83	19.82	34.06	25.58	20.00
	$M_{ti}(\%)$	0.16	0.78	0.69	4.30	0.14	1.05	4.07	4.76	4.15	4.61	6.57	0.64	16.92	7.87	3.49	6.19
XLM-R-L/14	$M_{cr}(\%)$	80.56	82.64	79.81	78.93	75.69	73.84	78.98	77.41	80.16	75.97	76.74	74.24	76.52	69.16	69.93	32.90
	$M_{orlb}(\%)$	8.17	6.05	6.50	9.53	9.48	8.76	8.28	10.34	9.73	11.89	13.27	7.25	8.54	9.03	4.65	6.19
	$M_{or}(\%)$	6.70	6.36	8.99	8.32	9.34	9.39	7.62	7.73	5.06	8.74	5.88	10.95	9.76	14.19	12.46	9.94
	$M_{cdr}(\%)$	4.58	4.96	4.01	2.55	4.67	7.49	4.60	4.40	4.93	3.40	3.56	6.44	4.88	5.29	9.47	33.29
	$M_{lb}(\%)$	0.00	0.00	0.55	0.54	0.41	0.32	0.53	0.00	0.13	0.00	0.14	0.00	0.30	0.77	1.00	7.10
	$M_{ti}(\%)$	0.00	0.00	0.14	0.13	0.41	0.21	0.00	0.12	0.00	0.00	0.41	1.13	0.00	1.55	2.49	10.58
XLM-R-B /16Plus	$M_{cr}(\%)$	77.29	78.45	75.24	72.89	71.98	71.84	73.19	72.29	74.19	68.57	73.46	68.28	69.51	64.00	64.45	32.77
	$M_{orlb}(\%)$	7.03	7.91	8.71	11.01	10.85	9.39	11.04	14.51	13.10	17.48	16.28	10.31	11.74	13.16	8.31	7.61
	$M_{or}(\%)$	12.25	9.46	12.03	12.62	12.23	12.03	9.99	9.51	7.52	10.19	7.80	12.88	14.33	14.97	13.79	9.16
	$M_{cdr}(\%)$	3.27	3.57	3.73	3.09	4.40	6.22	5.12	3.21	4.54	3.64	2.33	7.57	4.27	5.68	10.47	30.58
	$M_{lb}(\%)$	0.16	0.31	0.28	0.40	0.27	0.11	0.26	0.24	0.39	0.12	0.14	0.48	0.15	1.16	1.50	8.65
	$M_{ti}(\%)$	0.00	0.31	0.00	0.00	0.27	0.42	0.39	0.24	0.26	0.00	0.00	0.48	0.00	1.03	1.50	11.23
ColQwen2.5 -3B-Multi	$M_{cr}(\%)$	70.42	72.25	70.95	54.23	69.51	57.81	54.27	56.48	54.22	54.25	55.95	62.96	49.85	39.74	26.08	23.35
	$M_{orlb}(\%)$	13.24	11.32	13.00	28.32	15.11	27.95	32.19	28.42	29.57	27.79	29.69	17.71	36.59	35.74	34.72	28.39
	$M_{or}(\%)$	15.03	13.02	14.11	14.77	11.40	11.71	10.78	13.20	13.10	15.66	11.35	14.01	9.45	15.10	6.98	3.35
	$M_{cdr}(\%)$	0.98	2.95	1.52	1.61	3.30	1.90	2.10	1.31	1.69	1.46	2.19	4.03	1.68	4.39	14.95	21.29
	$M_{lb}(\%)$	0.33	0.00	0.14	0.67	0.41	0.21	0.53	0.36	0.91	0.73	0.27	0.97	1.98	3.61	13.79	21.16
	$M_{ti}(\%)$	0.00	0.47	0.28	0.40	0.27	0.42	0.13	0.24	0.52	0.12	0.55	0.32	0.46	1.42	3.49	2.45
ColQwen2.5 -7B-Multi	$M_{cr}(\%)$	69.44	71.94	69.71	49.26	64.56	56.65	55.19	49.11	51.62	54.37	56.09	63.45	49.09	49.94	31.56	27.10
	$M_{orlb}(\%)$	14.05	11.47	11.20	38.93	20.05	30.49	34.30	35.67	34.50	29.85	33.38	19.81	39.18	28.13	27.08	22.06
	$M_{or}(\%)$	14.71	13.33	16.32	9.93	11.68	10.86	7.62	9.99	10.51	12.14	8.21	9.82	6.86	13.55	6.98	5.29
	$M_{cdr}(\%)$	1.63	2.79	2.63	1.21	3.43	1.90	2.50	3.45	2.08	2.79	2.05	4.99	2.29	5.55	18.60	24.77
	$M_{lb}(\%)$	0.00	0.00	0.00	0.54	0.14	0.11	0.39	1.31	1.30	0.36	0.00	1.77	2.29	1.68	13.12	17.81
	$M_{ti}(\%)$	0.16	0.47	0.14	0.13	0.14	0.00	0.00	0.48	0.00	0.49	0.27	0.16	0.30	1.16	2.66	2.97
ColQwen2.5 -v0.2	$M_{cr}(\%)$	59.64	65.89	56.85	56.64	58.10	46.41	57.42	59.93	62.52	54.85	56.50	54.11	31.71	28.13	16.11	25.03
	$M_{orlb}(\%)$	20.42	17.21	19.64	25.77	22.94	33.97	24.44	25.09	19.33	24.51	27.50	21.42	38.72	43.23	44.85	18.58
	$M_{or}(\%)$	18.63	15.50	21.16	14.77	14.29	11.81	14.85	11.77	13.36	17.23	12.86	14.01	9.60	8.39	5.98	7.61
	$M_{cdr}(\%)$	1.31	1.40	1.94	2.42	3.98	5.80	3.15	2.85	3.63	2.79	2.60	6.28	1.83	5.42	11.63	26.32
	$M_{lb}(\%)$	0.00	0.00	0.00	0.27	0.41	1.79	0.13	0.24	0.52	0.24	0.41	3.54	17.68	13.42	19.27	15.87
	$M_{ti}(\%)$	0.00	0.00	0.41	0.13	0.27	0.21	0.00	0.12	0.65	0.36	0.14	0.64	0.46	1.42	2.16	6.58
GME-Qwen2 -2B-Instruct	$M_{cr}(\%)$	70.75	71.32	67.77	61.61	69.51	63.92	62.16	60.88	61.09	57.89	59.37	60.87	55.79	53.68	31.73	31.35
	$M_{orlb}(\%)$	12.25	12.71	15.35	20.13	11.95	19.20	17.87	22.95	21.92	22.69	24.21	19.00	25.61	18.19	20.60	13.81
	$M_{or}(\%)$	14.05	13.80	15.21	12.48	13.74	13.61	13.80	12.25	11.67	15.90	12.45	11.92	10.98	14.32	6.98	6.32
	$M_{cdr}(\%)$	2.94	1.86	1.52	3.89	4.26	2.85	4.73	2.97	3.63	2.67	2.05	5.64	4.27	10.19	24.42	28.90
	$M_{lb}(\%)$	0.00	0.00	0.00	0.94	0.41	0.21	1.05	0.48	1.17	0.49	1.50	1.29	2.13	2.45	11.79	13.29
	$M_{ti}(\%)$	0.00	0.31	0.14	0.94	0.14	0.21	0.39	0.48	0.52	0.36	0.41	1.29	1.22	1.16	4.49	6.32
GME-Qwen2 -7B-Instruct	$M_{cr}(\%)$	67.81	68.68	68.88	61.34	67.58	62.97	63.34	58.86	59.66	54.61	55.54	60.39	56.71	60.26	37.87	34.84
	$M_{orlb}(\%)$	11.60	11.63	12.31	22.01	16.35	23.00	20.24	25.68	26.98	28.03	30.37	21.74	28.51	17.42	23.92	17.81
	$M_{or}(\%)$	17.81	16.90	16.04	11.54	13.19	10.55	12.75	11.65	10.89	14.20	11.63	10.79	9.76	12.39	8.47	5.94
	$M_{cdr}(\%)$	2.78	2.48	2.63	4.03	2.75	2.95	3.15	2.85	1.82	2.79	1.92	5.48	3.81	8.65	18.94	25.03
	$M_{lb}(\%)$	0.00	0.16	0.00	0.67	0.14	0.42	0.39	0.71	0.52	0.36	0.55	1.13	0.91	0.90	7.81	11.48
	$M_{ti}(\%)$	0.00	0.16	0.14	0.40	0.00	0.11	0.13	0.24	0.13	0.00	0.00	0.48	0.30	0.39	2.99	4.90
Jina-E-v4	$M_{cr}(\%)$	61.76	64.81	61.13	60.67	58.65	57.38	56.64	54.10	58.88	57.28	59.92	60.06	57.16	55.74	38.54	36.39
	$M_{orlb}(\%)$	16.67	16.12	19.92	23.76	22.25	22.57	23.13	25.56	24.51	20.75	24.62	16.91	23.48	18.84	17.44	11.35
	$M_{or}(\%)$	19.77	17.21	17.15	13.69	15.66	18.04	17.08	17.12	14.79	20.02	14.36	18.52	16.01	17.55	9.97	8.77
	$M_{cdr}(\%)$	1.63	1.71	1.38	1.34	3.16	1.79	2.63	2.73	1.43	1.70	0.96	3.22	2.59	5.68	20.10	28.39
	$M_{lb}(\%)$	0.00	0.00	0.00	0.27	0.14	0.11	0.13	0.24	0.39	0.12	0.14	0.64	0.46	1.55	8.64	9.55
	$M_{ti}(\%)$	0.16	0.16	0.41	0.27	0.14	0.11	0.39	0.24	0.00	0.12	0.00	0.64	0.30	0.65	5.32	5.55

Table 14: This table presents the results for the Correct, Object Relevant Language-biased, Object Relevant, Cultural Descriptor Relevant, and Totally Irrelevant Win Percentages across various multimodal retrievers for a selection of countries.

Model	QL Association (Blue vs Purple)					CD vs QL (Yellow vs Pink)				
	Blue	**Purple**	**Diff**	p**-val**	**Sig.**	**Yellow**	**Pink**	**Diff**	p**-val**	**Sig.**
Vision-Language Contrastive Models										
CLIP-L/14	0.053	0.298	-0.245	0.000	True	0.085	0.126	-0.040	0.000	True
CN-CLIP-L/14	0.098	0.245	-0.147	0.000	True	0.100	0.112	-0.012	0.006	True
Cross-lingual Alignment Models										
XLM-R-L/14	0.068	0.072	-0.004	0.290	False	0.076	0.003	0.073	0.000	True
XLM-R-B/16plus	0.097	0.099	-0.002	0.647	False	0.066	0.004	0.062	0.000	True
LLM-Based Retrieval Embedders										
ColQwen2.5-3B-M	0.093	0.207	-0.114	0.000	True	0.094	0.013	0.081	0.000	True
ColQwen2.5-7B-M	0.078	0.177	-0.099	0.000	True	0.098	0.011	0.087	0.000	True
ColQwen2.5-v0.2	0.126	0.201	-0.076	0.000	True	0.096	0.021	0.074	0.000	True
GME-Qwen2-2B	0.115	0.184	-0.069	0.000	True	0.081	0.016	0.065	0.000	True
GME-Qwen2-7B	0.108	0.208	-0.100	0.000	True	0.068	0.014	0.054	0.000	True
Jina-E-v4	0.136	0.196	-0.060	0.000	True	0.071	0.012	0.059	0.000	True

Table 15: Win rates, differences, and statistical significance for the Object Relevance and Cultural Descriptor studies. Results are significant if $p < 0.05$. (Blue is the Object Relevant category, matching only the object. Purple represents the Object Relevant Language-biased category, which matches both the object and the culture of the language of the text query. Yellow represents the Cultural Descriptor Relevant category, which matches the CD's culture, and Pink is the Language-biased category, which matches only the culture of the language of the text query.)

Model	Metrics	Country															
		US	**GB**	**AU**	**DE**	**CN**	**JP**	**FR**	**ES**	**AR**	**PT**	**BR**	**SA**	**TH**	**IN**	**KE**	**NG**
CLIP-L/14	Diff	-0.01	0.01	0.02	-0.18	-0.45	-0.23	-0.19	-0.14	-0.23	-0.28	-0.36	-0.49	-0.58	-0.53	-0.21	-0.09
	p-val	**0.60**	**0.43**	**0.10**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CN-CLIP-L/14	Diff	0.00	-0.02	-0.01	-0.14	0.06	-0.08	-0.23	-0.14	-0.13	-0.17	-0.22	-0.41	-0.19	-0.39	-0.25	-0.07
	p-val	**1.00**	**0.30**	**0.74**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jina-E-v4	Diff	0.00	-0.01	-0.08	-0.12	-0.01	-0.05	-0.09	-0.05	-0.10	-0.05	-0.10	-0.02	-0.07	-0.05	-0.06	-0.07
	p-val	**0.94**	**0.73**	0.00	0.00	**0.65**	0.01	0.00	0.03	0.00	0.02	0.00	**0.39**	0.00	0.01	0.00	0.00
XLM-R-L/14	Diff	0.02	0.01	0.03	0.01	0.00	0.01	0.00	0.00	-0.04	-0.06	-0.08	0.01	0.00	0.02	0.02	0.01
	p-val	**0.20**	**0.33**	0.03	**0.32**	**0.85**	**0.70**	**0.77**	**0.72**	0.01	0.00	0.00	**0.38**	**0.77**	**0.21**	**0.17**	**0.39**
XLM-R-B/16plus	Diff	0.02	0.04	0.02	0.02	0.03	0.02	-0.01	-0.04	-0.07	-0.08	-0.04	0.04	0.00	0.03	0.01	0.01
	p-val	**0.24**	0.02	**0.26**	**0.27**	0.06	**0.28**	**0.42**	0.02	0.00	0.00	0.03	0.02	**0.80**	**0.14**	**0.48**	**0.17**
ColQwen2.5-3B-M	Diff	-0.03	0.01	0.00	-0.16	-0.03	-0.16	-0.22	-0.15	-0.15	-0.13	-0.20	-0.06	-0.18	-0.18	-0.10	-0.01
	p-val	0.07	**0.69**	**0.79**	0.00	**0.14**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.51**
ColQwen2.5-7B-M	Diff	0.02	0.00	0.03	-0.13	-0.06	-0.11	-0.23	-0.17	-0.15	-0.11	-0.16	-0.05	-0.17	-0.12	-0.09	-0.01
	p-val	**0.21**	**0.92**	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	**0.45**
ColQwen2.5-v0.2	Diff	-0.05	-0.03	-0.02	-0.05	-0.02	-0.10	-0.13	-0.05	-0.06	-0.03	-0.13	-0.01	-0.16	-0.15	-0.18	-0.04
	p-val	0.05	**0.13**	**0.28**	0.01	**0.33**	0.00	0.00	0.01	0.00	**0.12**	0.00	**0.76**	0.00	0.00	0.00	0.01
GME-Qwen2-2B	Diff	0.02	0.01	0.00	-0.07	0.00	-0.09	-0.09	-0.11	-0.09	-0.07	-0.18	-0.10	-0.08	-0.06	-0.13	-0.03
	p-val	**0.41**	**0.64**	**0.89**	0.00	**0.94**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
GME-Qwen2-7B	Diff	0.05	0.03	0.02	-0.12	-0.01	-0.11	-0.14	-0.14	-0.11	-0.12	-0.24	-0.15	-0.14	-0.09	-0.16	-0.13
	p-val	0.03	**0.13**	**0.29**	0.00	**0.50**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 16: Win rate differences and p-values for each model across different countries in the QL Association Test.

Model	Metrics	Country															
		US	GB	AU	DE	CN	JP	FR	ES	AR	PT	BR	SA	TH	IN	KE	NG
CLIP-L/14	Diff	0.04	0.04	0.04	0.06	-0.24	0.07	0.05	0.04	0.04	0.03	0.01	-0.46	-0.38	-0.41	0.10	0.23
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.67**	0.00	0.00	0.00	0.00	0.00
CN-CLIP-L/14	Diff	0.05	0.05	0.05	0.08	0.05	0.01	0.04	0.06	0.04	0.03	0.02	-0.39	-0.24	-0.34	0.05	0.19
	p-val	0.00	0.00	0.00	0.00	0.00	**0.22**	0.00	0.00	0.00	0.06	**0.26**	0.00	0.00	0.00	0.05	0.00
Jina-E-v4	Diff	0.02	0.02	0.02	0.04	0.05	0.05	0.04	0.03	0.02	0.03	0.02	0.04	0.02	0.12	0.18	0.25
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
XLM-R-L/14	Diff	0.05	0.05	0.06	0.04	0.04	0.05	0.04	0.04	0.05	0.05	0.04	0.06	0.06	0.05	0.09	0.40
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XLM-R-B/16plus	Diff	0.03	0.04	0.05	0.04	0.04	0.05	0.04	0.03	0.03	0.03	0.03	0.08	0.03	0.05	0.08	0.35
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ColQwen2.5-3B-M	Diff	0.08	0.10	0.10	0.07	0.03	0.02	0.06	0.03	0.04	0.04	0.04	0.09	0.02	0.07	0.20	0.34
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.12**	0.00	0.00	0.00
ColQwen2.5-7B-M	Diff	0.04	0.07	0.07	0.08	0.04	0.02	0.07	0.05	0.05	0.07	0.07	0.07	0.02	0.11	0.23	0.35
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
ColQwen2.5-v0.2	Diff	0.05	0.04	0.05	0.12	0.06	0.03	0.08	0.05	0.06	0.07	0.05	0.11	0.02	0.10	0.07	0.22
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00
GME-Qwen2-2B	Diff	0.02	0.04	0.03	0.06	0.04	0.03	0.05	0.03	0.03	0.06	0.01	0.07	0.04	0.11	0.14	0.28
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.11**	0.00	0.00	0.00	0.00	0.00
GME-Qwen2-7B	Diff	0.03	0.04	0.04	0.05	0.03	0.02	0.05	0.03	0.04	0.06	0.02	0.05	0.02	0.07	0.13	0.20
	p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00

Table 17: Win rate differences and p-values for each model across different countries in the CD vs QL test.

standard CLIP, adding a cultural descriptor introduces embedding noise rather than semantic clarity, effectively pushing the retrieved result further away from the ground truth. In contrast, multilingual models such as *XLM-R* and *GME-Qwen2* maintain consistent positive drift across all tested regions, confirming their superior cross-cultural alignment capabilities.

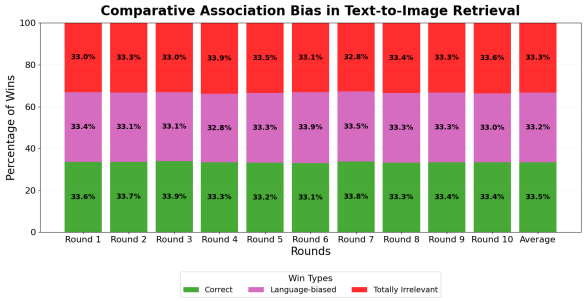

Figure 19: Association Bias Evaluation in Random Baseline.

R SP score for Random Baseline

We conducted a text-to-image experiment for the three candidates using a setup similar to the one described in Section 3.2.2. However, instead of selecting the image with the highest similarity, we selected images uniformly at random. Theoretically, a perfectly unbiased random baseline should yield an SP score of 1.0. We repeated this process for ten rounds (illustrated in Figure 19). Across all iterations, the observed SP scores ranged narrowly between 0.98 and 1.02, with an overall average of 0.99. These results closely align with the theoretical expectation and confirm the absence of association bias in the random baseline.

S Association Bias Evaluation in Verbose Text Query.

We extended the association bias evaluation to encompass scenarios both in the absence of and incorporating explicit cultural descriptors (in both RQ2 and RQ3) in the case where we *increase the verbosity of the prompt without giving more object or cultural cues*.

S.1 Association Bias in Verbose Queries without Cultural Descriptors

Our framework evaluates association bias in detailed text queries using two primary formats: task-based and casual. The former specifies the intended use of the image (e.g., "a picture of a train *for*

Model	Difference by Query Language Culture															
	US	**GB**	**AU**	**DE**	**CN**	**JP**	**FR**	**ES**	**AR**	**PT**	**BR**	**SA**	**TH**	**IN**	**KE**	**NG**
CLIP-L/14	4.89	4.54	4.68	4.93	**0.64**	3.43	3.55	4.34	4.51	3.14	3.02	**0.21**	**0.18**	**-0.16**	2.81	3.74
CN-CLIP-L/14	5.04	5.09	5.36	4.40	5.27	2.80	3.58	4.44	4.52	3.52	3.59	**-0.51**	**0.01**	**0.02**	3.20	5.61
GME-Qwen2-2B	7.47	6.41	7.51	7.21	7.78	7.08	7.30	8.31	9.36	7.32	7.84	7.26	8.03	6.96	6.14	8.15
GME-Qwen2-7B	3.91	3.62	3.98	3.48	7.81	6.29	1.60	3.41	4.01	2.10	2.03	3.77	4.96	2.17	1.69	5.36
Jina-E-v4	2.82	2.52	2.57	1.71	**0.73**	2.05	1.67	2.15	2.66	1.72	1.28	2.29	2.69	1.90	2.76	1.62
XLM-R-L/14	5.71	5.28	5.31	4.13	3.70	3.06	4.96	4.92	5.15	4.93	4.87	3.86	3.97	2.82	4.64	5.65
XLM-R-B/16Plus	9.52	9.10	9.34	6.57	5.55	4.40	8.82	8.05	8.20	8.18	7.98	5.41	5.91	3.92	8.42	11.77

Table 18: Mean Similarity Drift ($\Delta Sim \times 100$) by Model and Query Language Culture. The table measures the change in cosine similarity to the *Correct* when a specific cultural descriptor is added to the query. Positive values indicate successful instruction following, while near-zero or negative values (bolded) indicate that the model fails to process the cultural descriptor or treats it as noise.

my homework"), whereas the latter captures the user's situational intent (e.g., "*I am looking for* a bicycle"). Query generation was facilitated by Gemini[3] using the prompt templates illustrated in Figure 22. The evaluation result is shown in Figure 20. The results indicate that the accuracy for verbose text queries across both task-based and casual formats decreases by 6–9% compared to short queries (RQ2) for CLIP-L/14 and CN-CLIP-L/14. In contrast, Cross-lingual Alignment Models and MLLM-based retrieval models remain stable, with deviations of less than 4%, except for ColQwen2.5-7B-M, which decreases by 6% in the task-based format. The drop in accuracy occurs because the additional words in the text can distract the model from the core concept of the query. This issue is more severe in CLIP-L/14 and CN-CLIP-L/14, as these models appear to struggle with non-Latin script languages (with the exception of Chinese and Japanese for CN-CLIP-L/14). Consequently, as the query length increases, the model becomes more biased toward the culture associated with the language rather than the object.

S.2 Association Bias in Verbose Queries with a Cultural Descriptor

To evaluate association bias in verbose queries with a cultural descriptor, a country descriptor is added to the neutral text queries from subsection S.1 using Gemini[2], following the prompt shown in Figure 21. An example of a resulting query is "A picture of a Japanese train for my homework". The evaluation result is shown in Figure 21, short queries consistently achieve the highest correct answer across almost all models. The additional words in query cause models to drift away from the target concept or culture, especially in CLIP and CN-CLIP models, where accuracy drops by approximately 9-12% in verbose formats. This performance loss is primarily driven by an increase in the "Object Relevant, Language-biased" category, indicating that verbose queries cause these models to overrely on the language's cultural priors. Furthermore, a distinct "Language Bias" failure mode (Pink) is observed in CLIP models, which worsens with verbosity, confirming a tugging effect toward pure language heuristics. In contrast, the "Cultural Diversity Relevant" category (Yellow) remains stable (5-10%) for most architectures but diminishes in CLIP models, suggesting that the additional text dilutes the model's focus on explicit cultural descriptors.

T Analysis of Tugging Effect per Concept

We calculated the mean similarity difference scores across all countries for the *gme-Qwen2-VL-7B-Instruct* model, with the results detailed in Table 19. Our analysis reveals that concepts where cultural descriptors inadvertently increase bias (indicated by a negative drift for the target category) are typically general concepts with minimal cross-cultural variation (e.g., taxi, shopping mall, bakery). While these concepts are included to ensure comprehensive coverage of everyday items, they represent only a small minority ($\sim$20%) of the benchmark. In contrast, the concepts that successfully decrease bias (exhibiting a positive drift) are those with highly distinct, easily definable visual characteristics (e.g., map, coin, banknotes). Ultimately, these findings highlight the wide diversity of concepts encapsulated within the 3XCM benchmark.

[3]Version used: `gemini-2.5-flash` (Released September 25, 2025).

Concepts (1–23)		Concepts (24–46)		Concepts (47–69)	
map	0.2387	capital	0.0753	graduated uniform	0.0473
coin	0.1595	headwear	0.0720	unique art form	0.0467
banknotes	0.1516	folklore character	0.0708	military parade	0.0463
prime minister	0.1357	warrior	0.0698	ancient city	0.0459
football player	0.1268	soldier	0.0692	unique cuisine trait	0.0455
sport	0.1164	tourist attraction	0.0688	major religious site	0.0447
major river	0.1130	artwork	0.0683	building	0.0446
folk tale	0.1048	street sign	0.0672	unique food ingredient	0.0445
writing character	0.0976	ancient painting	0.0671	celebrity	0.0432
tattoo style	0.0975	congress	0.0614	natural landmark	0.0426
night view	0.0917	formal uniform	0.0610	priest	0.0421
ghost	0.0850	architecture	0.0571	child	0.0405
lottery tickets	0.0847	painting	0.0567	storytelling	0.0396
costume	0.0847	older	0.0538	doll	0.0389
devil	0.0830	museum	0.0537	fashion	0.0389
police station	0.0812	street view	0.0527	sacred object	0.0380
university	0.0809	shirt	0.0525	tv program	0.0359
gathering place	0.0807	art	0.0513	statue	0.0352
social custom	0.0805	pants	0.0506	textile pattern	0.0344
historical image	0.0801	mythological figure	0.0499	funeral	0.0344
people	0.0792	wall painting	0.0492	historical figure	0.0340
embroidery style	0.0781	rural dwelling	0.0487	fountain style	0.0330
religious building	0.0771	historical event	0.0475	series	0.0326

Concepts (70–92)		Concepts (93–115)		Concepts (116–138)	
mailbox	0.0292	department store	0.0103	train	-0.0038
stageplay	0.0285	mythical creature	0.0099	necklace	-0.0041
school	0.0282	airlines	0.0098	instrument	-0.0047
train station	0.0281	tea culture	0.0097	bus	-0.0047
combination food	0.0267	house	0.0096	mask	-0.0108
fire station	0.0246	unique transportation	0.0085	weapon	-0.0112
ritual	0.0230	airport	0.0081	poaching food	-0.0112
dance	0.0230	drink	0.0077	dessert	-0.0144
snack	0.0223	festival	0.0075	animal	-0.0153
street entertainment	0.0221	tree	0.0063	zoo	-0.0181
pottery style	0.0205	marriage ceremony	0.0058	cinema	-0.0214
street vendor cart	0.0199	wedding	0.0057	hot pot concept	-0.0221
singer	0.0190	gas station	0.0042	diningroom	-0.0225
moist heat food	0.0175	bedroom	0.0036	martial art	-0.0231
craft	0.0172	popular street food	0.0032	bathroom	-0.0231
ancient craft	0.0166	toy	0.0030	boat	-0.0285
unique natural phenomenon	0.0158	bracelet	0.0018	restaurant	-0.0292
native inhabitants	0.0144	hat	0.0013	bag	-0.0346
food	0.0141	music band	0.0008	deep fried food	-0.0363
major mountain range	0.0139	symbolic bird	-0.0001	alcohol drink	-0.0397
bus station	0.0129	dry heat food	-0.0006	bakery	-0.0443
game	0.0128	hotel	-0.0025	shopping mall	-0.0607
market	0.0110	symbolic plant	-0.0036	taxi	-0.0806

Table 19: Distribution of Concepts and Mean Similarity Scores

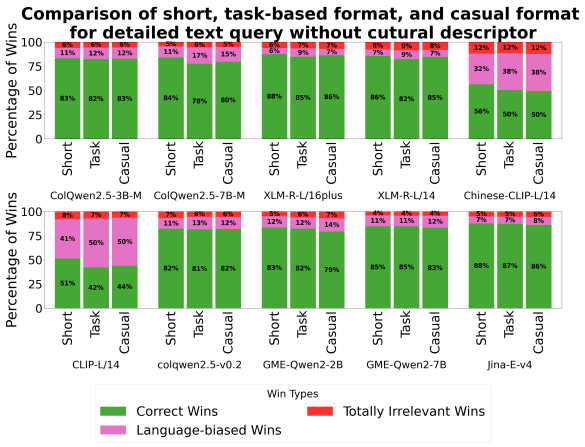

Figure 20: Comparative association bias across verbose text query in task-based and casual format without cultural descriptors.

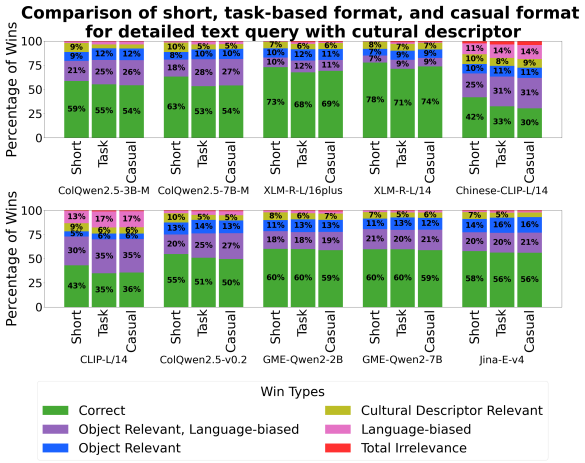

Figure 21: Comparative association bias across verbose text query in task-based and casual format with cultural descriptors.

U Semantically Focused Instruction Prompt Effects

This section investigates the impact of system prompts on MLLM retriever performance. We evaluate three distinct prompts: (1) a baseline default prompt, (2) a standard text-to-image retriever prompt ("Find an image that matches the given caption"), and (3) a culturally agnostic prompt ("Find an image that matches the given caption. Focus on the semantics instead of the textual language"). The results are presented in Figure 24. As shown, the associative bias remains relatively consistent across different prompt configurations. This suggests that varying the system prompt does not significantly alter the model's retrieval behavior.

V Computational Resource

The experiment is performed with a single A100 GPU for approximately 3 gpu hours for each model or 54 hours in total with library version of colpali-engine 0.3.13.dev1+g9bee9b2b7, transformers 4.53.3 for most experiments except GME models are utilized under transformers 4.51.3

W Authoring and Implementation Tools

In preparing this manuscript, we utilized several generative large language models. For language editing and stylistic refinement, we employed Google's Gemini, along with models from xAI's Grok family (e.g., Grok Expert and Fast variants). For assistance with code implementation, scripting, and debugging, we used a model from Anthropic's Claude series (e.g., Claude Sonnet).

X Annotator Workloads and Dataset Generation Costs

The dataset was annotated by three members of the research team over approximately 60 total hours, with an average annotation rate of roughly 200 images per hour. In terms of cost, the culturally relevant images were collected via the DuckDuckGo API entirely for free, and internal manual annotation resulted in no external labor costs. The only financial expense incurred was approximately $40 USD for Gemini API usage.

```
You are an expert Computational Linguist and Native Speaker for the following countries: 'China', 'India', 'Japan',
'Saudi Arabia', 'France', 'Germany', 'Brazil', 'Kenya', 'Thailand', 'USA', 'Spain', 'Argentina', 'UK', 'Australia',
'Nigeria', 'Portugal'
 Your task is to generate natural search queries for a Cross-Modal Information Retrieval dataset. I will provide you
with a list of target languages and the specific "Native Concept Word" (the search object) for each.

 INPUT DATA:
              {
        'bus':
                    {
                     'China': '公交车',
                      'India': 'बस',
                      'Japan': 'バス',
                      'Saudi Arabia': 'حافلة',
                      'France': 'Bus',
                      'Germany': 'Bus',
                      'Brazil': 'Ônibus',
                      'Kenya': 'Basi',
                      'Thailand': 'รถบัส',
                      'USA': 'Bus',
                      'Spain': 'Autobús',
                      'Argentina': 'Colectivo',
                      'UK': 'Bus',
                      'Australia': 'Bus',
                      'Nigeria': 'Basi',
                      'Portugal': 'Ônibus'
                    },
                }

RULES:
                STRICT CONSTRAINT: You MUST include the provided "Native Concept Word" exactly as written in the
input.  Do not change it, do not use synonyms, and do not translate it yourself.
                Grammar & Fluency: Ensure the sentence is grammatically natural for a native speaker.  Pay attention
to classifiers (e.g., Thai 'khan'), gender agreement, and polite particles.
                Parallel: meaning across language is the same.
                Culture Agnostic: No country/culture cue.
                Output Format: Return ONLY a valid JSON object mapping the concept to the generated prompts.

 Target Language Mapping:
                China: Chinese
                India: Hindi
                Japan: Japanese
                Saudi Arabia: Arabic
                France: French
                Germany: German
                Brazil: Portuguese (Brazilian)
                Portugal: Portuguese (European)
                Kenya: Swahili
                Thailand: Thai
                Spain: Spanish (Peninsular)
                Argentina: Spanish (Rioplatense)
                Nigeria: Yoruba
                USA, UK, Australia: Localized English variants.

Example of expected Output Format:
                {
                  "<concept_key>": {
                    "format_task": {
                      "China": "Value representing a task-based search query...",
                      "India": "Value...",
                      "Japan": "Value..."
                    },
                    "format_casual": {
                      "China": "Value representing a casual/conversational search query...",
                      "India": "Value...",
                      "Japan": "Value..."
                    }
                  }
                }
The Variable Task Definitions (One at a time)
                TASK_DEFINITIONS = {
                "format_task": "The user needs the image for a generic, realistic project. Use contexts like 'for a
presentation,' 'for a school project,' or 'for a blog post.' Ensure the task makes sense for the object.",
                "format_casual": "Create a casual, chatty, or verbose sentence describing user intent. (e.g., concept
is 'Train' output should be I'm looking for a train because I like them' or'Find me a cool [Concept] to look at')."
                }
```

Figure 22: The prompt given to Gemini to generate unique country-specific image concepts.

````
**Role:** You are a Senior Computational Linguist and Polyglot with native-level fluency in the primary languages of the
following countries:
China (ZH), India (HI), Japan (JA), Saudi Arabia (AR), France (FR), Germany (DE), Brazil (PT-BR), Kenya (SW), Thailand (TH),
USA (EN-US), Spain (ES-ES), Argentina (ES-AR), UK (EN-GB), Australia (EN-AU), Nigeria (EN-NG), and Portugal (PT-PT).

**Task:**
1. **Modify:** Take the English sentences provided in `format_task` and `format_casual`. Insert a country-specific
reference (e.g., concept: train, mention country: China, original sentence: I need images of a Train for a school project.
The modified sentence: I need images of a chinese train for a school project.) into the English sentence.
2. **Translate:** Translate the modified sentences into the native language of the "Target Country."

**Strict Constraints:**
1. **Grammar & Fluency:** Adjust the sentence structure, classifiers (e.g., Thai 'khan'), and gender markers to ensure
the sentence is natural for a native speaker in that country.
2. **No Stereotypes:** Keep the sentences "Culture Agnostic." Do not add descriptors like "high-speed" for China or
"vintage" for India. Only add the country name/adjective.
3. **Target Language Mapping:**
    *   Nigeria: Yoruba

**Input Data:**
```json
{
  "concept": "bus",
  "format_task": "I need an image of a Bus for my school project.",
  "format_casual": "I'm just looking for some cool Bus pictures to check out."
}
```

**Output Format:**
Return ONLY a valid JSON object following this exact structure:

```json
{
  "bus": {
    "format_task": {
      "Nigeria": {
        "mention_China": "Translated string here...",
        "mention_India": "Translated string here...",
        "mention_Japan": "...",
        "mention_Saudi Arabia": "...",
        "mention_France": "...",
        "mention_Germany": "...",
        "mention_Brazil": "...",
        "mention_Kenya": "...",
        "mention_Thailand": "...",
        "mention_USA": "...",
        "mention_Spain": "...",
        "mention_Argentina": "...",
        "mention_UK": "...",
        "mention_Australia": "...",
        "mention_Nigeria": "...",
        "mention_Portugal": "..."
      },

    },
    "format_casual": {
      "Nigeria": {
        "mention_China": "Translated casual string here...",
        "...": "..."
      },

    }
  }
}
````

Figure 23: The prompt given to Gemini to generate unique country-specific image concepts.
````

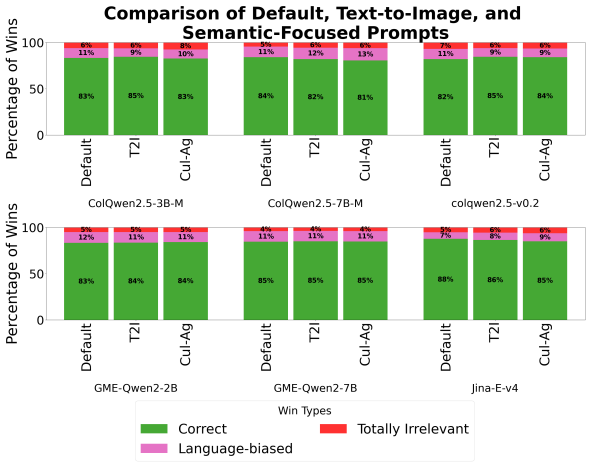

Figure 24: Comparative association bias across default prompt, standard text-to-image retriever prompt (T2I), and a culturally agnostic prompt (Cul-Ag).