Training free Cross-Modal Alignment for Multilingual Image Retrieval

Usman Shahid[†] **and Manan Jain**[†] **and Natalie Parde**
Department of Computer Science
University of Illinois Chicago
{hshahi6, mjain35, parde}@uic.edu

Abstract

We investigate whether multilingual image–text retrieval can be made practical without training a multi-modal model, using frozen multilingual uni-modal encoders and inducing cross-modal alignment through relative representations computed against a large external anchor set. We use English image–caption pairs as anchors and evaluate caption-to-image retrieval on 3,600 images with captions in 36 languages (some low-resource). Across multiple encoder pairings and hyperparameter settings, the method yields usable performance, positioning training-free relative representation as a baseline for multilingual cross-modal alignment when target-language multi-modal training data is unavailable.

1 Introduction

Vision–language retrieval is usually approached by learning a shared embedding space from large collections of paired image-text data, such as in CLIP and LiT (Radford et al., 2021; Zhai et al., 2022). These models are powerful, but their effectiveness depends on large-scale supervised training and the availability of these multi-modal pairs. In multilingual settings, the problem becomes harder: large multi-modal resources remain uneven across languages, and even evaluation has required dedicated benchmarks such as XM3600, which provides 3,600 images with reference captions in 36 languages (Jain et al., 2021; Thapliyal et al., 2022).

To address the scarcity of paired multilingual training data, we investigate a fully training-free alternative based on the ASIF method (Norelli et al., 2023). ASIF successfully aligns frozen uni-modal encoders into a joint multi-modal space using coupled data without additional parameter updates. Instead of learning a new cross-modal projection, we keep both pretrained uni-modal image encoders and uni-modal multilingual text encoders entirely frozen, aligning them instead through a shared set of external anchor pairs. Concretely, we use English image–caption pairs as anchors (Changpinyo et al., 2021), represent each non-English caption and image by a sparse top-k vector of similarities to those anchors, and retrieve images by comparing these modality agnostic relative representations directly. This framing turns cross-modal retrieval into a comparison over shared anchor coordinates rather than over a jointly trained multi-modal space (Norelli et al., 2023).

Our primary objective is not to supersede fully trained multilingual vision-language models. Instead, we seek to determine how far a purely non-parametric formulation can go in a realistic, multilingual retrieval environment in the absence of target-language paired multi-modal data. By evaluating caption-to-image retrieval on the XM3600 benchmark across various encoder pairings and hyperparameter settings, we position training-free relative representation retrieval as a viable baseline for multilingual multi-modal alignment when target-language paired multi-modal training data is otherwise unavailable. Specifically, our contributions are as follows:

- We demonstrate that measurable multilingual caption-to-image retrieval can be achieved without training a new multi-modal model or learning a new multi-modal projection from target-language paired data.
- We evaluate our pipeline on 3,600 images across 36 languages (including low-resource languages) from the XM3600 benchmark. Our experiments span nine encoder combinations, utilizing state-of-the-art multilingual text models and vision backbones.
- We establish a modifiable, modular, and partially interpretable (Norelli et al., 2023) em-

[†]Equal contribution.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4039–4049
November 6-10, 2026 ©2026 Association for Computational Linguistics

pirical baseline for multilingual cross-modal alignment in scenarios where multi-modal training data is limited or absent.

Additionally, we make all of our code and other resources publicly available to facilitate transparency and replicability.[1]

2 Related Work

2.1 Multi-modal Alignment and Relative Representations

Most vision–language retrieval systems learn alignment by fitting a shared embedding space from paired image–text data. Early zero-shot transfer methods learned explicit mappings between modality-specific representations (Socher et al., 2013), and this idea was later scaled via contrastive vision–language pretraining, most prominently CLIP (Radford et al., 2021). A related thread studied performance while reusing strong pretrained uni-modal backbones; for instance, LiT freezes encoders and learns only a lightweight alignment objective (Zhai et al., 2022), while Flamingo keeps large pretrained backbones frozen and introduces additional cross-attention layers for multi-modal reasoning (Alayrac et al., 2022). These approaches still encode alignment in learned parameters and typically rely on substantial paired training.

Our work is closer in spirit to non-parametric views of alignment. Relative representations propose communicating across latent spaces by representing inputs through their similarities to a shared anchor set (Moschella et al., 2023). (Norelli et al., 2023)'s ASIF method adapts this idea to multi-modal alignment by treating alignment as a retrieval-time operation rather than a learned projection. In parallel, multilingual multi-modal systems such as MPM and VisCPM show that multilingual language models can act as pivots to help transfer multi-modal behavior learned from English-rich data to other languages (Hu et al., 2024). Recent analysis also suggests that stronger language models tend to exhibit stronger visual alignment in low-training or frozen-backbone regimes (Ruthardt et al., 2026). Building on these directions, in this work we ask a narrower question: *can we obtain multilingual cross-modal transfer in a fully training-free way, using only frozen uni-modal encoders and a high-resource anchor bank?*

[1] https://github.com/ushahid/npa-multilingual

2.2 Multi-modal Multilingual Retrieval

Multilingual settings add a practical constraint: paired image–text data is far less available outside high-resource languages. XM3600 provides a standardized benchmark for multilingual caption-to-image retrieval (Thapliyal et al., 2022), and models such as MURAL demonstrate that multilingual multi-modal representations can be learned when enough paired data exists (Jain et al., 2021). Multilingual multi-modal corpora, however, remain skewed toward a handful of languages including English, while multilingual text-only resources and multilingual text encoders are comparatively easier to obtain at scale. This mismatch motivates our setup, wherein we keep the anchor bank in a high-resource language (in our case, English) and rely on a multilingual text encoder to connect low-resource captions to those anchors during retrieval.

Zero-shot cross-lingual transfer has also been studied in trained multilingual vision–language systems. (Huang et al., 2021) show that multilingual multi-modal pre-training can improve retrieval in non-English languages without additional language-specific annotations, representing a complementary setting to ours. Their approach learns multilingual multi-modal alignment during pre-training, whereas we study whether cross-modal retrieval can be induced at inference time using frozen uni-modal encoders and an English-only paired anchor bank.

2.3 Text and Image Encoders

Since our method does not learn a new cross-modal projection, performance depends heavily on the quality of the frozen encoders. For text, we include (i) *BGE-M3*, a multilingual retrieval-oriented embedding model (Chen et al., 2024), (ii) *NLLB-200-3.3B*, a large multilingual model trained with broad cross-lingual supervision (Costa-jussà et al., 2022), and (iii) *Qwen3-Embed-8B*, a recent high-capacity multilingual embedding model designed for embedding and ranking tasks (Zhang et al., 2025). For vision, we evaluate encoders spanning contrastive and self-supervised paradigms: *OpenCLIP ViT-bigG/14* (trained at scale and supported by reproducible CLIP scaling-law analyses) (Cherti et al., 2023), *DINOv3 ViT-7B/16* (Siméoni et al., 2025), and *C-RADIOv4-H*, a foundation model that aggregates multiple teacher signals into a single strong visual representation (Ranzinger et al., 2026).

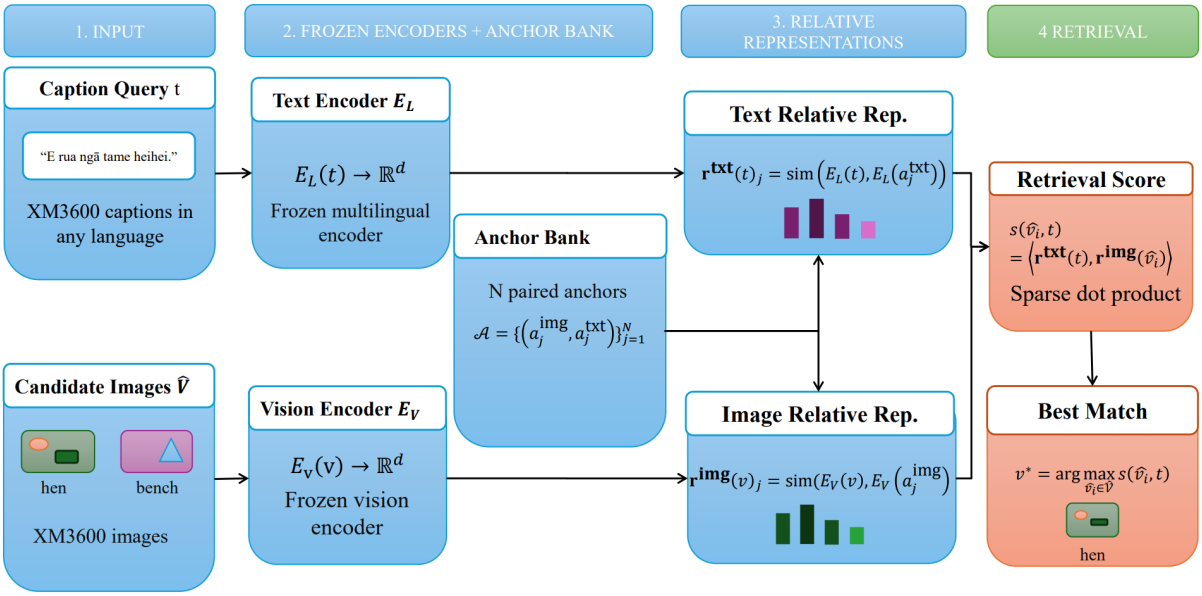


Figure 1: Schematic of the proposed relative-representation retrieval pipeline. Given a caption query in any XM3600 language, we compute its similarity profile over CC12M text anchors and compare it against cached similarity profiles of candidate images over the paired CC12M image anchors. After top-k sparsification and normalization, images are ranked by similarity in the shared relative-representation space, enabling multilingual cross-modal retrieval with frozen encoders only.

3 Methodology

We adapt the ASIF relative-representation framework (Norelli et al., 2023) to multilingual caption-to-image retrieval. The central idea is to avoid learning a new multi-modal projection and instead induce cross-modal alignment at retrieval time through a shared bank of paired anchors. This design is motivated by the challenges described in §2: multilingual image–text pairs are scarce and unevenly distributed across languages, while strong uni-modal encoders and high-resource paired data are easier to obtain. Rather than learning a new shared multi-modal space, we keep a pretrained vision encoder E_V and a pretrained multilingual text encoder E_L fixed, and induce alignment at retrieval time through an ordered set of paired anchors

$$\mathcal{A} = \{(a_j^{\text{img}}, a_j^{\text{txt}})\}_{j=1}^{N}.$$

For an encoded input x' and an ordered anchor set $\{x_1, \ldots, x_N\}$, the relative representation is

$$x_{\text{rel}} = [\text{sim}(x', x_1), \text{sim}(x', x_2), \ldots, \text{sim}(x', x_N)]$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity. We instantiate this separately for images and captions while preserving anchor-pair order:

$$\mathbf{r}^{\text{img}}(v)_j = \text{sim}(E_V(v), E_V(a_j^{\text{img}})),$$

$$\mathbf{r}^{\text{txt}}(t)_j = \text{sim}(E_L(t), E_L(a_j^{\text{txt}})).$$

Since the j-th coordinate in both vectors refers to the same image–text pair, the two representations become comparable across modalities even though they originate from different encoder spaces.

Given a caption query t and a set of candidate images $\hat{\mathcal{V}} = \{\hat{v}_1, \ldots, \hat{v}_c\}$, we follow the ASIF assumption and treat the relative representation of the query caption as if it were the relative representation of its ideal image. This corresponds to defining the ideal image profile by the query itself, i.e., $y^* := x^*$, and retrieving the candidate whose relative representation is most similar to it (Norelli et al., 2023). In our implementation this reduces to

$$s(\hat{v}_i, t) = \left\langle \mathbf{r}^{\text{txt}}(t), \mathbf{r}^{\text{img}}(\hat{v}_i) \right\rangle,$$

$$v^* = \arg\max_{\hat{v}_i \in \hat{\mathcal{V}}} s(\hat{v}_i, t).$$

Here $s(\hat{v}_i, t) \in [-1, 1]$ represents the retrieval score for image $\hat{v}_i$ given the text query t. This makes the method fully training-free. Once the uni-modal embeddings and anchor bank are fixed, retrieval is performed entirely by comparing anchor-based similarity vectors in the modality agnostic relative representation space.

As in ASIF, to reduce noise in these high-dimensional vectors, we used sparsification and exponentiation. Let $\text{TopK}(\mathbf{r}, k)$ denote the set of

indices of the top-k entries of $\mathbf{r}$. We keep only those coordinates,

$$\tilde{\mathbf{r}}_i = \begin{cases} \mathbf{r}_i & \text{if } i \in \text{TopK}(\mathbf{r}, k), \\ 0 & \text{otherwise,} \end{cases}$$

and then sharpen the remaining values by raising them to a power $p > 1$ and applying the L2 norm:

$$\hat{\mathbf{r}} = \frac{\tilde{\mathbf{r}}^{\,p}}{\|\tilde{\mathbf{r}}^{\,p}\|_2}.$$

The final retrieval score is therefore

$$s(\hat{v}_i, t) = \left\langle \hat{\mathbf{r}}^{\text{txt}}(t), \hat{\mathbf{r}}^{\text{img}}(\hat{v}_i) \right\rangle.$$

Here, k controls how many anchors remain active, while p determines the spread of similarity score distribution. We summarize our overall method in Figure 1.

In practice, we use CC12M (Changpinyo et al., 2021) as the fixed anchor bank and XM3600 (Thapliyal et al., 2022) as the evaluation set. Image relative representations are computed once and cached, caption relative representations are computed on the fly, and retrieval is implemented as a sparse dot product between caption and image relative representations (RRs). This design keeps the method modular and interpretable: each active coordinate corresponds to a specific anchor pair. Additionally, any of the vision encoder, text encoder, or anchor bank can be swapped without retraining a multimodal alignment layer. Anchor pairs can likewise be added or removed without retraining, by adding or removing the corresponding RR dimensions.

4 Experimental Setup

4.1 Datasets and Anchor Bank

We evaluate multilingual caption-to-image retrieval on XM3600 (Thapliyal et al., 2022), a benchmark containing 3,600 images with captions in 36 languages. Following the formulation in §3, we construct the shared cross-modal reference system from an ordered bank of paired anchors drawn from a separate large English image–caption collection, CC12M (Changpinyo et al., 2021). Each anchor pair consists of an image anchor and its corresponding caption anchor, and the shared ordering of these pairs defines the coordinates of the relative-representation space. CC12M is not used for fine-tuning or parameter updates; it is used only to induce alignment non-parametrically at retrieval time. In the final experiments, we use the first 1.6M CC12M image–caption pairs as anchors, following the experimental setup of Norelli et al. (2023).

4.2 Baselines

We compare relative representations to strong baselines in the multilingual caption-to-image retrieval setting. We emphasize that the baselines are intended to set a performance ceiling rather than a performance floor. To assess feasibility of relative representations for settings with limited multimodal paired data availability, experiments are thus designed to reveal how the performance of the approach compares with strong proprietary systems (e.g., Google Translate) or native multi-modal embedding spaces (e.g., Qwen3-VL).

Two baselines, *CLIP + NMT translation* and *CLIP + TLLM translation*, test the former point. For each XM3600 caption, we generate English translations using Google's Neural Machine Translation (NMT) model or translation large language model (TLLM) system; NMT is a popular, baseline machine translation model and TLLM is the current state-of-the-art proprietary machine translation system.[2] We embed the translated captions using OpenCLIP ViT-bigG/14, and we embed the corresponding XM3600 images using the aligned CLIP image encoder. Retrieval is performed directly in the CLIP embedding space using cosine similarity between caption and image embeddings. Since the image set is unchanged across translation variants, image embeddings are computed once and reused across both translated-caption baselines.

Another baseline, *Qwen3-VL-Embedding-8B*, leverages a native multi-modal embedding model. In this setting, the original XM3600 source captions in their native languages and the corresponding images are embedded directly by the same multi-modal model, and caption-to-image retrieval is again performed by cosine similarity in the shared embedding space. This baseline is complementary to the translation-based CLIP setup. Rather than translating queries into English before retrieval, it uses an off-the-shelf multilingual multimodal embedding model to solve the task directly without an anchor-based alignment step.

4.3 Encoder Combinations

We evaluate nine encoder combinations formed by pairing three multilingual text encoders with three vision encoders, as discussed in §2. On the text side, we use `bge-m3`, `nllb-200-3.3B`, and `qwen3-embed-8b`. These models were cho-

[2] https://docs.cloud.google.com/translate/docs/translation-llm

sen to cover different kinds of multilingual representations: `bge-m3` is a retrieval-oriented multilingual embedding model (Chen et al., 2024), `nllb-200-3.3B` is a large translation-oriented multilingual model whose encoder representations are shaped by broad cross-lingual supervision (Costa-jussà et al., 2022), and `qwen3-embed-8b` is a recent high-capacity embedding model designed for embedding and ranking tasks (Zhang et al., 2025). Together, they represent a wide variety of model sizes and pre-training objectives.

On the vision side, we use `C-RADIOv4-H`, `CLIP ViT-bigG/14`, and `DINOv3 ViT-7B/16`. These models span different visual representation-learning paradigms: `CLIP ViT-bigG/14` is a large contrastive vision encoder (Costa-jussà et al., 2022), `DINOv3 ViT-7B/16` is a self-supervised vision backbone (Siméoni et al., 2025), and `C-RADIOv4-H` is a recent vision foundation model (Ranzinger et al., 2026). We reiterate that because our method does not learn a new multi-modal projection layer, retrieval quality depends directly on the capability of the frozen uni-modal model; comparing vision models drawn from different representation-learning paradigms allows us to isolate which model type(s) are best suited for relative representation settings.

For each text–image encoder pair, both encoders are kept frozen. Images and captions are embedded independently, then mapped to the shared anchor-relative space via cosine similarities to the corresponding image and text anchors. This enables us to isolate the effect of encoder choice for evaluation.

4.4 Relative-Representation Processing

For every encoder pair, we compute relative representations for captions and images using the formulation from §3. We then apply the top-k sparsification and p exponentiation step. Specifically, we sweep

$$k \in \{200, 400, 800, 1200, 1600\}$$

for sparsification; and

$$p \in \{2, 4, 8\}$$

for the elementwise exponent and pick the best performing hyperparameters based on our evaluation metric. After this processing, caption queries and image candidates are compared by the dot product between their normalized relative representations. This produces a fully training-free retrieval pipeline in which the only changes across runs are the encoder pair and the relative-representation hyperparameters.

4.5 Evaluation Protocol

We evaluate our approach within the framework of caption-to-image retrieval task. Given a query caption in one of the XM3600 languages, the system ranks all $3,600$ candidate images based on the retrieval score descibed in §3 and the results are considered correct if the corresponding image is ranked within the top-K results. We report Recall@K for

$$K \in \{1, 5, 10, 25, 50\},$$

For evaluations, we use all available XM3600 captions as queries, yielding 261,375 captions in total. This setup allows us to measure not only overall retrieval quality, but also how well the anchor-based transfer mechanism holds across a diverse multilingual benchmark.

4.6 Compute Environment

All experiments were run on a single NVIDIA H200 NVL GPU with 143GB device memory. The host machine provided 251GiB RAM and approximately 14TB of local NVMe storage. Since the method is training-free, computation is dominated by embedding inference and anchor-relative retrieval rather than by multi-modal optimization. The experiments required approximately one week of GPU time. Translation baselines were implemented using the Google Cloud Platform API; generating the necessary captions using the NMT and TLLM models costing us approximately $579 USD.

5 Results

5.1 Overall Comparison

Table 1 compares the best run for each relative-representation (RR) encoder pair against the three dense baselines. Within the proposed RR setting, the strongest configuration is `Qwen3-Embed-8B + C-RADIOv4-H`, which reaches **57.5** R@10 at $(k, p) = (800, 4)$. It is followed closely by `Qwen3-Embed-8B + CLIP ViT-bigG/14` at **56.9** R@10, while `Qwen3-Embed-8B + DINOv3 ViT-7B/16` reaches **49.9** R@10. The BGE-M3 pairs form a clear middle tier, with best R@10 values between **38.9** and **46.5**, and the NLLB-200-3.3B experiments show substantially weaker performance, with best R@10 values between **17.5** and **20.4**.

Method	k	p	R@1	R@5	R@10	R@25	R@50
CLIP + NMT translation	–	–	64.0	85.4	90.6	95.2	97.3
CLIP + TLLM translation	–	–	65.8	87.1	92.1	96.4	98.2
Qwen3-VL-Embedding-8B	–	–	63.8	83.2	88.0	92.3	94.4
Qwen3-Embed-8B + C-RADIOv4-H	800	4	21.2	46.1	57.5	71.2	79.8
Qwen3-Embed-8B + CLIP ViT-bigG/14	1200	8	21.0	45.5	56.9	70.6	79.2
Qwen3-Embed-8B + DINOv3 ViT-7B/16	800	2	17.2	39.0	49.9	63.3	72.4
BGE-M3 + C-RADIOv4-H	1200	4	15.0	35.4	46.5	61.5	72.4
BGE-M3 + CLIP ViT-bigG/14	800	4	14.5	34.6	45.6	60.7	71.3
BGE-M3 + DINOv3 ViT-7B/16	800	2	12.3	29.4	38.9	52.8	63.3
NLLB-200-3.3B + C-RADIOv4-H	1600	4	5.6	14.7	20.4	29.4	37.1
NLLB-200-3.3B + CLIP ViT-bigG/14	1600	4	5.0	13.6	19.1	27.9	35.8
NLLB-200-3.3B + DINOv3 ViT-7B/16	1600	2	4.7	12.5	17.5	25.5	32.6

Table 1: Overall caption-to-image retrieval results on XM3600. For the proposed relative representation method, each row reports the configuration with the highest R@10 for that text–image encoder pair over the (k, p) sweep. The three top rows are dense retrieval baselines that do not use anchor-relative representations. Scores are percentages.

%	Pairs	Emb. size (GB)	RR cost (TFLOP)	Avg. R@10 $\pm\sigma$
1	16K	0.40	0.27	42.42 ± 0.53
2.5	40K	0.99	0.67	46.65 ± 0.18
5	80K	1.98	1.34	49.85 ± 0.39
7.5	120K	2.98	2.01	51.44 ± 0.31
10	160K	3.97	2.68	52.56 ± 0.23
15	240K	5.95	4.03	53.80 ± 0.15
20	320K	7.93	5.37	54.48 ± 0.12
25	400K	9.92	6.71	55.00 ± 0.10
30	480K	11.90	8.05	55.38 ± 0.11
40	640K	15.87	10.73	55.93 ± 0.16
50	800K	19.84	13.42	56.32 ± 0.13
75	1.2M	29.75	20.13	57.07 ± 0.05
100	1.6M	39.67	26.84	$\mathbf{57.49 \pm 0.02}$

Table 2: Anchor-bank size ablation for the strongest RR encoder pair i.e. `Qwen3-Embed-8B + C-RADIOv4-H`, using random subsets of the CC12M anchor bank. Embedding size reports the stored anchor-bank embedding footprint, and RR cost reports the estimated anchor-relative computation cost for constructing relative representations. Scores are R@10 percentages averaged across repeated anchor samples; σ denotes standard deviation.

The dense baselines are strongest in absolute terms, as expected. The best overall system is the CLIP baseline using LLM-translated captions, which reaches **92.1** R@10 and **98.2** R@50. The CLIP baseline using NMT-translated captions is nearly as strong at **90.6** R@10, and the direct `Qwen3-VL-Embedding-8B` baseline reaches **88.0** R@10. Although unsurprisingly within the present setting these off-the-shelf dense retrieval baselines remain stronger than the training-free RR formulation, the RR results isolate the extent to which multilingual cross-modal transfer can be induced *without* learning a new multi-modal alignment layer.

5.2 Effect of Encoder Choice in the RR Model

The RR experiments reveal a clear and stable ranking of encoding methods, with the text encoder being the main factor determining R@10 performance. Across all three image encoders, `Qwen3-Embed-8B` is consistently strongest, `BGE-M3` is intermediate, and `NLLB-200-3.3B` trails by a large margin. For example, with `C-RADIOv4-H` fixed, best R@10 rises from **20.4** with NLLB-200-3.3B to **46.5** with BGE-M3 and then to **57.5** with Qwen3-Embed-8B. The same ordering is preserved with both `CLIP ViT-bigG/14` and `DINOv3 ViT-7B/16`. This indicates that, in a non-parametric cross-modal transfer setup, multilingual text encoder quality is the central determinant of success.

The results suggest that the image encoder still matters, but less than the text encoder. Within the strongest Qwen3-based RR expeirments, `C-RADIOv4-H` performs best, `CLIP ViT-bigG/14` is a very close second, and `DINOv3 ViT-7B/16` is consistently lower. The spread is modest between the top two image encoders (about 0.6 R@10 points between `Qwen3-Embed-8B + C-RADIOv4-H` and `Qwen3-Embed-8B + CLIP ViT-bigG/14`) but larger when moving to DINOv3. This suggests that stronger visual features help, but cannot compensate for weak multilingual text representations.

5.3 Anchor-Bank Scaling and Cost–Performance Trade-off

To better understand the compute–accuracy trade-off of the anchor-relative formulation, we additionally evaluate how performance changes with the size of our anchor bank (CC12M). This ex-

periment uses the strongest RR encoder pair, `Qwen3-Embed-8B + C-RADIOv4-H`, and reports mean R@10 and standard deviation over 10 independently sampled random subsets at each anchor-bank size. Since the method does not update model parameters, increasing the anchor bank does not introduce training cost, but it does increase embedding storage and the number of anchor-similarity computations required to construct relative representations. The storage and computation values reported in Table 2 quantify RR-specific overhead. They account for storing the paired anchor embeddings and computing anchor similarities, while excluding encoder inference, which is independent of anchor-bank size; dense retrieval baselines do not require this storage of anchor set embeddings or computation of relative representations.

Table 2 shows that retrieval quality improves monotonically with anchor-bank size, but the gains quickly begin to saturate. Using only 10% of the anchor bank already reaches 52.56 R@10, compared to 57.49 R@10 using the full 1.6M-anchor bank. Increasing the anchor bank from 10% to 25% improves R@10 by 2.44 points, whereas increasing it from 50% to 100% improves R@10 by only 1.17 points while doubling both storage and RR computation cost. This suggests that most of the cross-modal alignment can be achieved with a relatively small subset of anchors, while further increasing the anchor bank size provides diminishing returns.

5.4 Sensitivity to k and p

In our experiments, based on the R@10 results, we have observed that the strongest RR encoder pairs generally benefit from moderately large values of k. For `Qwen3-Embed-8B + C-RADIOv4-H`, R@10 improves steadily from $k = 200$ to $k = 800$, peaks at $k = 800$, and then largely plateaus for larger k. For `Qwen3-Embed-8B + CLIP ViT-bigG/14`, the same upward trend continues through $k = 1200$, after which performance is essentially unchanged.

Second, exponentiation still helps the stronger RR configurations, but the best value of p is more encoder pair-dependent under R@10. `Qwen3-Embed-8B + C-RADIOv4-H` reaches its best R@10 at $p = 4$, while `Qwen3-Embed-8B + CLIP ViT-bigG/14` peaks at $p = 8$. The strongest BGE-M3 pairs also prefer moderate-to-strong exponentiation, while the DINOv3-based RR systems remain more sensitive to over-exponentiation and achieve their best R@10 at $p = 2$. The NLLB-based runs remain low across all settings, though they still show modest gains as k increases.

L	Q3T + C-RADIO	Q3T + CLIP	Q3T + DINO	CLIP NMT	CLIP TLLM	Q3-VL
ar	55.7	54.5	47.3	90.8	90.8	92.0
bn	47.6	45.4	40.1	85.1	86.1	87.8
cs	48.5	47.1	41.3	89.7	90.1	88.8
da	54.4	53.2	46.6	94.9	94.9	94.1
de	66.3	63.9	57.6	96.5	96.6	97.7
el	53.8	51.8	46.3	91.2	91.4	89.7
en	61.1	59.8	53.9	86.6	86.6	89.7
es	60.2	58.6	52.1	92.4	92.5	94.7
fa	55.6	54.7	48.1	92.7	92.8	91.9
fi	49.5	48.4	41.5	95.6	95.7	89.8
fil	46.3	45.1	39.9	86.0	85.7	79.5
fr	61.4	59.5	53.1	96.3	96.3	97.4
he	48.0	46.7	39.6	94.5	94.3	92.8
hi	40.9	38.7	34.1	74.5	75.5	78.7
hr	53.1	51.5	44.9	93.7	93.6	91.1
hu	58.3	55.7	49.1	95.6	95.9	93.6
id	64.9	61.7	57.1	96.0	96.1	97.7
it	59.2	57.3	50.5	95.7	95.7	97.2
ja	60.3	58.0	52.2	95.8	95.9	97.0
ko	58.6	57.0	49.7	93.5	93.8	95.5
mi	9.7	9.6	7.9	68.9	–	26.1
nl	54.1	51.8	46.4	91.6	91.6	91.9
no	52.3	50.0	45.3	94.8	94.6	93.4
pl	59.2	57.1	50.9	94.5	94.5	94.9
pt	60.5	59.5	52.0	93.6	93.7	94.4
quz	12.3	11.7	11.1	56.2	–	18.8
ro	56.8	55.8	48.6	96.3	96.4	94.3
ru	66.6	64.5	57.4	96.5	96.8	97.8
sv	55.8	53.6	47.4	93.1	92.9	92.3
sw	18.6	18.5	16.4	81.5	80.1	48.6
te	47.8	46.6	40.8	79.9	79.8	75.7
th	60.8	59.6	51.7	93.9	94.1	95.2
tr	53.7	52.6	45.7	92.1	92.1	91.2
uk	62.7	61.3	53.5	95.1	95.5	95.8
vi	61.6	59.9	52.2	94.3	94.6	96.8
zh	64.1	62.4	56.3	92.3	92.6	95.7
avg	52.8	51.2	45.2	90.3	92.0	87.2

Table 3: Per-language (L) R@10 comparison under the same experimental conditions. The three Q3T embedding columns correspond to the Qwen3-Embed-8B text encoder paired with C-RADIOv4-H, CLIP ViT-bigG/14, and DINOv3 ViT-7B/16, all evaluated at $k = 200$, $p = 2$. The final three columns report the dense baselines; CLIP NMT and CLIP TLLM correspond to the translated captions and Q3-VL (Qwen3-VL-Embedding-8B) is the multimodal model applied directly to the original multilingual captions and images. Scores are percentages. "–" indicates that a language was not available in the translated-caption baseline output.

5.5 Per-Language Comparison

Table 3 shows the per-language comparison of the baselines and the three Qwen3-based RR encoder pairs, revealing several patterns. First, the ranking of image encoders within the Qwen3-based RR experiments is stable across most languages:

C-RADIOv4-H is usually best, CLIP ViT-bigG/14 is usually second, and DINOv3 ViT-7B/16 is typically third. Second, the dense baselines outperform the RR systems across nearly every language as was our expectation. This is true for both higher-resource (e.g., German, Russian, French, and Chinese) and lower-resource (e.g., Hindi, Telugu, and Swahili) languages. Third, the lowest-performing languages remain low-performing for all systems. In particular, mi (Maori) and quz (Cusco Quechua) still exhibit substantially weaker performance than the rest of the benchmark under both the RR setup and the dense baselines.

Although the baselines are broadly competitive, their strengths vary by language. Qwen3-VL exhibits the strongest performance for several high-resource languages such as German, French, Indonesian, Russian, and Chinese, whereas the translation-based systems remain competitive or stronger for some others. The TLLM CLIP baseline is the best overall method, but it could not produce translations for mi and quz. Thus, the RR method shows consistent internal structure and consistent multilingual transfer patterns while still revealing plenty of room for performance growth compared to proprietary translation approaches or native multi-modal embedding spaces.

6 Discussion

Overall, our results validate that the proposed relative representations (RR) are usable for multilingual caption-to-image retrieval without end-to-end multi-modal training. As opposed to the multi-modal encoders, the RR pipeline must first encode each multilingual caption, convert it into a modality agnostic relative representation using an anchor set, and finally match it against the relative representations of images built from the paired image anchors. In our implementation, this matching is carried out as a sparse dot product over top-k anchor coordinates, so the method remains training-free and interpretable. As expected, while the RR pipeline does not require paired data or pre-training, it underperforms baselines that employ supervised pre-training with paired data. The CLIP baselines encode captions translated to English and then retrieve most similar images directly from the multimodal CLIP space using cosine similarity. The Qwen3-VL baseline encodes the original multilingual captions and images using the same pre-trained multi-modal encoder and performs retrieval without requiring an anchor set. Both sets of baselines therefore avoid the extra step of computing the relative representations using the anchor set, which suggests that the inference time computational cost of going through a high-resource language anchor bank is still substantial in the present setup even though there is no training cost.

Our follow-up analyses reveal that text encodings are more influential than image encodings to overall RR performance. Qwen3-Embed-8B is consistently the strongest text model, BGE-M3 is reliably second, and NLLB-200-3.3B trails far behind. That pattern is stable across all three image backbones. This lines up with the basic assumption that the method depends on the multilingual text encoder already having cross-lingual connections in the semantic space, because RR never learns to repair weak text-side alignment after the fact. If that assumption is only partially true, the error propagates through the entire retrieval chain. The image encoder still affects performance, but its role is secondary; better vision features help, yet they do not compensate for weak multilingual text representations.

The hyperparameter sweeps show that for the strongest RR runs, moderately large k and aggressive exponentiation are helpful. Since the relative-representation vectors are high-dimensional and noisy by construction, sparsification and exponentiation appear to positively influence overall performance. The fact that weaker encoder pairs remain weak even after tuning k and p also tells us that hyperparameter search alone will not yield substantial performance boosts.

The per-language results show that the lowest-performing languages remain challenging for every system, offering opportunity for future improvement. For the current RR pipeline, the anchor bank is English-only, so the method depends entirely on the multilingual text encoder to bridge from a low-resource caption into the anchor space before any image-side matching can help. The translation-based CLIP baselines bypass that issue by moving the caption into English first, and the Qwen3-VL baseline bypasses it by using a single multi-modal embedding model for both modalities. Thus, the translation and Qwen3-VL baselines are tasked with solving an easier alignment problem than RR, shedding additional light on the performance gap.

Overall, the relative-representation remains attractive because it is modular, interpretable, and does not require training a new multi-modal model.

The bottleneck is the quality of the encoders, especially text encoders and especially for low-resource languages.

7 Conclusion

Our work examined whether multilingual caption-to-image retrieval can be carried out without training a new multi-modal model, by combining frozen uni-modal encoders with relative representations over a large English anchor bank. The results confirm that the method holds promise. The strongest configuration of our proposed method, `Qwen3-Embed-8B + C-RADIOv4-H`, reached R@10 = 57.5, with `Qwen3-Embed-8B + CLIP ViT-bigG/14` close behind at 56.9. Across our experiments Qwen3-based text representations consistently yielded the strongest performance, followed by BGE-M3 and then NLLB-200-3.3B. Our results also suggest that in the current training-free setup, success depends more on the quality of the multilingual text space than on the choice of vision backbone.

Relative representations are practically usable when paired data is unavailable for training and they provide a decent baseline but they do not provide comparable performance to systems that employ supervised pre-training; which was expected. The best overall result came from the CLIP baseline with LLM-translated captions at R@10 = 92.1, then CLIP with NMT-translated captions at 90.6 followed by the multimodal `Qwen3-VL-Embedding-8B` model at 88.0. These results show that when feasible (i.e. paired multimodal data or competitive translation models are available), translating into a high resource language like English or using a native multi-modal embedding model is a more effective approach.

Taken together, these results lead to the conclusion that relative representations are viable as a non-parametric mechanism for multilingual cross-modal alignment in the absence of paired data, but their success is limited by the quality of the anchor space and multilingual text encoder. In our case, this is weakest for low-resource languages and when the multilingual text encoder does not accurately capture the cross-lingual relationships in the semantic space. Our experiments show that (1) relative representations can provide alignment across modalities and languages, (2) the choice of text encoder dominates performance, and (3) ASIF-style (Norelli et al., 2023) sparsification and exponentiation remain important for making the method practical.

8 Limitations & Future Work

The main limitation of our training free method is the comparative performance gap with respect to the baselines that employ supervised training. Our best configuration, `Qwen3-Embed-8B + C-RADIOv4-H`, remains well below the strongest baseline. CLIP with LLM-translated captions had the best performance for retrieval and other baselines had similar results. This means that the current RR pipeline is informative as a training-free alignment method and can be employed in the absence of paired data, but it cannot compete with supervised methods for caption-image retreival.

Since we have only experimented with an anchor bank containing English captions, multilingual transfer depends heavily on the encoder's ability to capture multilingual relationships in the semantic space with respect to the English captions. The results also show that performance is highly sensitive to the text encoder itself: `Qwen3-Embed-8B` which clearly outperforms `BGE-M3` and both of these encoders outperform `NLLB-200-3.3B` by a large margin. Future work should examine multilingual and domain-matched anchor banks, more text encoders as well as informed anchor-selection strategies beyond random sampling.

Additionally, the interpretability of relative representations is limited to tracing retrieval decisions back to (image, caption) pairs in the anchor set; it does not explain the cross-lingual geometry learned by the multilingual text encoder. Since alignment is achieved through relative representations rather than learned cross-modal interactions, RR may be less effective at capturing fine-grained relationships between modalities.

Finally, we had to limit the scope of our experimental evaluation to one pair of modalities (image, text), one task (caption-image retrieval) using one dataset (XM3600) for evaluations. The proposed relative representations can potentially be used with other modality pairs e.g. (audio, text) and other tasks such as zero-shot image classification (Norelli et al., 2023).

Our method is training-free but still requires non-trivial computational resources; we trade-off the cost of training a multimodal model with the cost of inference. Our relative representations incur a one-time cost of computing and caching embed-

dings for the anchor set and calculating the relative representations requires non-trivial computational resources at inference time. We do however quantify these limitations in §5.

9 Ethics Statement

This work studies multilingual caption-to-image retrieval using frozen pretrained encoders and a fixed anchor bank derived from existing datasets. We do not collect new human-subject data, and all experiments are conducted on previously released resources, namely XM3600 for evaluation and CC12M for anchor construction (Thapliyal et al., 2022; Changpinyo et al., 2021). By relying on large-scale pretrained models and a high-resource English-centered anchor set, the method may inherit biases in language coverage, cultural representation, and semantic associations present in those resources. Retrieval quality may therefore vary across languages and concepts. The system does not generate new text or images, but multilingual retrieval methods can still be misused in applications such as surveillance, profiling, or unfair filtering of visual content. We do not develop or evaluate the method for such uses, nor do we claim that it is suitable for high-stakes deployment.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions. Compute resources used for this work were partially supported by the National Science Foundation through Grant No. 2408401. U. Shahid was partially supported via Grant R61DA057629 through the National Institutes of Health. N. Parde was partially supported by Grant R61DA057629 as well, and was also partially supported by Grant R01AG091762. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568.

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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jinyi Hu, Yuan Yao, Chongyi Wang, SHAN WANG, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. Large multilingual models pivot zero-shot multimodal learning across languages. In *The Twelfth International Conference on Learning Representations*.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online. Association for Computational Linguistics.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In *The Eleventh*

International Conference on Learning Representations.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. 2023. Asif: Coupled data turns unimodal models to multimodal without training. In *Advances in Neural Information Processing Systems*, volume 36, pages 15303–15319. Curran Associates, Inc.

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.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. 2026. C-radiov4 (tech report). *Preprint*, arXiv:2601.17237.

Jona Ruthardt, Gertjan J. Burghouts, Serge Belongie, and Yuki M Asano. 2026. Better language models exhibit higher visual alignment. *Transactions on Machine Learning Research*.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. Dinov3. *Preprint*, arXiv:2508.10104.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.