Multilingual CLIP Alignment via Image Pivoting

Mattia Di Iorio
Supervised by Prof. Uri Hasson

January 17, 2026

Abstract

Multilingual vision–language models have recently gained increasing attention as a means to extend the success of contrastive multimodal learning beyond English-centric settings. While prior work has shown that multilingual alignment can be achieved using parallel text or machine translation, it remains unclear to what extent visual grounding alone is sufficient to induce cross-lingual alignment, and how such alignment manifests in the structure of learned representations.

In this work, we study multilingual vision–language representation learning under an image-pivoted contrastive framework, where images act as language-agnostic anchors and no parallel text is used during training. We adopt a CLIP-style objective and train on multilingual image–caption pairs drawn from diverse languages, encouraging captions that describe the same visual content to align implicitly through shared visual supervision.

Beyond standard retrieval benchmarks, we compare linear and nonlinear text projection heads and focus on a detailed representation-level analysis of multilingual alignment. We examine cross-lingual representational similarity, intrinsic dimensionality, isotropy, and feature activation statistics before and after image-pivot alignment.

We find that multilingual alignment emerges as a gradual and modest convergence of representational structure under linear alignment, while nonlinear heads yield larger retrieval gains with weaker changes in global cross-lingual geometry (e.g., RSA) and no clear signs of degeneration in hubness or activation statistics.

1 Introduction

Learning joint vision–language representations has emerged as a central problem in modern machine learning, enabling progress across a wide range of tasks such as image–text retrieval, visual question answering, and zero-shot image classification. Models trained with contrastive objectives on large-scale image–caption datasets, most notably CLIP-style architectures, have demonstrated that aligning visual and textual modalities in a shared embedding space yields representations that are both flexible and transferable.

Despite these advances, the majority of existing vision–language models remain heavily English-centric. This limitation is particularly problematic given the inherently multilingual nature of visual content on the web and the growing demand for models that operate robustly across languages. Extending multimodal representation learning beyond English is therefore not merely a matter of coverage, but a prerequisite for fairness, accessibility, and global applicability.

A natural approach to multilingual vision–language learning is to rely on parallel text, machine translation, or explicitly aligned multilingual corpora. While effective, these strategies introduce several drawbacks: translation pipelines can be noisy and biased, parallel data is scarce for low-resource languages, and strong textual supervision may obscure the role of visual grounding in cross-lingual alignment. As a result, it remains unclear to what extent multilingual alignment truly emerges from shared semantics, as opposed to being imposed by textual correspondence.

An alternative and conceptually appealing paradigm treats images as language-agnostic pivots. Images provide a universal semantic signal that is independent of linguistic form, allowing captions in different languages that describe the same visual content to be indirectly aligned through contrastive learning. Under this image-pivot assumption, multilingual alignment can, in principle, emerge even in the absence of parallel text, provided that the visual modality is sufficiently informative and consistently grounded.

Rather than proposing a new state-of-the-art multilingual CLIP variant, we intentionally restrict adaptation capacity to study when and how image-pivot supervision induces cross-lingual alignment.

In this work, we explicitly distinguish between alignment capacity and representational structure. To this end, we study both linear residual and nonlinear (MLP-based) text projection heads under the same image-pivoted contrastive objective. This comparison allows us to disentangle gains due to increased model capacity from those arising from genuine structural alignment of multilingual representations.

Several recent works have explored multilingual multimodal learning under this perspective, showing promising results on cross-lingual retrieval and zero-shot transfer. However, most existing studies focus primarily on downstream performance, leaving important questions unanswered. In particular, we still lack a clear understanding of how multilingual alignment manifests in the representation space, how it affects the geometry and dimensionality of embeddings, and whether improved alignment comes at the cost of representational collapse or loss of language-specific nuance.

In this work, we revisit multilingual vision–language learning through the lens of image-pivoted contrastive alignment, with a specific focus on representation analysis. We study a CLIP-style framework trained on multilingual image–caption pairs without using parallel text during training, where the image encoder acts as the sole bridge across languages. Our goal is not only to improve multilingual retrieval performance, but also to characterize what changes in the embedding space as multilingual alignment emerges.

To this end, we conduct a systematic analysis of multilingual text embeddings before and after image-pivot alignment. We examine cross-lingual representational similarity via Gram matrices, intrinsic dimensionality and isotropy, and feature activation statistics to characterize how alignment reshapes the embedding space. This analysis allows us to disentangle genuine semantic alignment from superficial improvements driven by hubness or over-regularization.

Our contributions can be summarized as follows:

- We demonstrate that image-pivoted contrastive learning can significantly improve multilingual image–text retrieval across a diverse set of languages, without relying on parallel text or machine translation.

- We show that image-pivot alignment increases cross-lingual Gram-matrix correlation and improves isotropy (lower mean pairwise cosine similarity), while moderately reducing effective rank without collapse indicators (PoZ↓, entropy↑).

- We show that effective multilingual alignment can be achieved while preserving healthy representation statistics, avoiding collapse and excessive loss of language-specific information.

- We release a reproducible training and analysis pipeline to facilitate further research on multilingual multimodal representation learning at `https://github.com/mattiadri/crosslingual-contrastive`.

Overall, our findings support the view that visual grounding alone constitutes a powerful and underexplored signal for multilingual alignment, and that careful analysis of representation geometry is essential for understanding the strengths and limitations of image-pivoted approaches.

2 Related Work

Multilingual multimodal representation learning aims to build joint embedding spaces that align visual and textual semantics across multiple languages. Early work in vision–language modeling focused primarily on English-centric datasets and models, but recent years have seen a growing interest in extending these approaches to multilingual settings. In this section, we review prior work on multilingual vision–language representation learning, with particular emphasis on methods that leverage images as alignment pivots, contrastive learning objectives, and large-scale pre-training.

2.1 Multilingual Multimodal Representation Learning

One of the earliest directions in multilingual multimodal learning investigates how to align representations across languages by leveraging shared semantic signals. Kim et al. introduce MULE (Multimodal Universal Language Embedding), a framework designed to learn language-agnostic embeddings by jointly modeling text and images across multiple languages [5]. MULE combines multilingual textual supervision with visual grounding to encourage the emergence of a shared semantic space, demonstrating improved cross-lingual transfer on downstream tasks such as image–text retrieval and classification.

Ni et al. propose M3P, a large-scale pre-training framework that unifies multilingual and multimodal learning via multitask objectives [7]. M3P combines masked language modeling, masked region modeling, and cross-modal alignment objectives across multiple languages, showing that large-scale pre-training can substantially improve zero-shot and cross-lingual performance. However, M3P relies on parallel or weakly-aligned multilingual corpora and requires substantial computational resources.

Zhou et al. introduce UC^2 (Universal Cross-Lingual Cross-Modal Vision-and-Language Pre-Training), which explicitly targets universality across both languages and modalities [8]. UC^2 integrates cross-lingual and cross-modal objectives within a unified transformer architecture, demonstrating strong performance on multilingual vision–language benchmarks. While effective, the approach again assumes access to multilingual supervision that may include parallel data.

2.2 Image-Pivoted Alignment Across Languages

A particularly relevant line of work explores the idea of using images as language-agnostic pivots to align text representations across languages. Gella et al. propose image pivoting as a mechanism to induce multilingual multimodal representations without requiring direct parallel text [3]. In this paradigm, captions in different languages that describe the same image are implicitly encouraged to align through their shared visual grounding. This insight is central to many subsequent approaches and motivates methods that avoid explicit cross-lingual textual supervision.

Difference from Gella et al. and positioning. While Gella et al. introduce the image-pivot intuition to induce multilingual multimodal representations, our focus is deliberately different: we study *lightweight* image-pivot alignment on top of *frozen* pretrained CLIP-compatible encoders, by training only a small text-side projection head (linear residual or shallow MLP). This controlled setting isolates how much alignment can be induced purely from visual grounding without end-to-end adaptation, and enables detailed geometric diagnostics (RSA, intrinsic dimensionality, isotropy, activation statistics). Practically, this makes our approach well-suited to small-resource or domain-specific scenarios where end-to-end training is unstable, expensive, or data-limited (e.g., narrow-domain artwork or medical-image caption retrieval).

Mohammadshahi et al. study the problem of aligning multilingual word embeddings for cross-modal retrieval, showing that visual information can serve as an effective bridge between languages [6]. Their work highlights how cross-modal objectives can induce cross-lingual structure even when textual alignment signals are weak or absent.

Building on these ideas, Jain et al. propose MURAL, a multitask and multilingual retrieval framework that learns aligned representations across languages and modalities [4]. MURAL leverages contrastive learning with multilingual image–text pairs and demonstrates strong cross-lingual retrieval performance, especially in low-resource settings. However, the model still benefits from explicit multilingual balancing and task-specific supervision.

2.3 Multilingual Extensions of CLIP

CLIP-style contrastive learning has become a dominant paradigm for vision–language representation learning. Several works extend CLIP to multilingual scenarios. Chen et al. introduce mCLIP, which adapts CLIP to multiple languages via cross-lingual transfer [2]. mCLIP initializes from an English CLIP model and transfers knowledge to other languages using multilingual text encoders and translation-based supervision. While effective, this approach depends on translation quality and parallel data during training.

Ahmat et al. propose M^2-VLP, which enhances multilingual vision–language pre-training through multi-grained alignment, combining sentence-level and token-level objectives [1]. Their approach shows that finer-grained alignment signals can improve multilingual robustness, but at the cost of increased model complexity and training overhead.

Overall, multilingual CLIP variants demonstrate that strong multilingual performance is achievable, but many existing approaches rely on machine translation, parallel corpora, or explicit cross-lingual supervision.

2.4 Limitations of Prior Work and Positioning

Despite significant progress, existing multilingual multimodal approaches exhibit several limitations. Many methods rely on parallel text or translation pipelines, which may introduce noise and bias, especially for low-resource languages. Others require complex architectures or large-scale pre-training that limits reproducibility and accessibility.

In contrast, image-pivoted approaches suggest that visual grounding alone may suffice to induce cross-lingual alignment, provided that training is carefully designed. However, prior work has not fully characterized how such alignment emerges, which layers contribute most to cross-lingual convergence, and how representation geometry changes during training.

Our work builds on the image-pivot paradigm by systematically studying multilingual alignment in CLIP-style models trained on non-parallel multilingual image–caption pairs. We place particular emphasis on representation analysis—such as intrinsic dimensionality, neighborhood structure, and cross-lingual similarity—to better understand when and why image-pivot supervision succeeds or fails.

3 Method

3.1 Problem Formulation

We consider a multilingual vision–language learning setting with a set of images

$$\mathcal{I} = \{I_i\}_{i=1}^{N}, \tag{1}$$

and, for each image I_i, a set of captions

$$\mathcal{T}_i = \{T_i^{(l)} \mid l \in \mathcal{L}\}, \tag{2}$$

where $\mathcal{L}$ denotes a set of languages. During training, captions in different languages are never paired directly with each other; instead, each caption is paired only with its corresponding image.

Our goal is to learn a shared embedding space in which image representations and multilingual text representations are aligned, such that captions describing the same visual content—regardless of language—are mapped close to the associated image and, indirectly, close to each other.

3.2 Dual-Encoder Architecture

We adopt a dual-encoder architecture similar to CLIP, consisting of:

- an image encoder f_{img}, which maps an image I to a visual embedding
$$\mathbf{v} = f_{\text{img}}(I) \in \mathbb{R}^d, \tag{3}$$
- a multilingual text encoder f_{text}, which maps a caption T to a textual embedding
$$\mathbf{t} = f_{\text{text}}(T) \in \mathbb{R}^d. \tag{4}$$

The image encoder is initialized from a pretrained vision–language model and kept frozen in our main experiments. The text encoder is also pretrained and frozen; multilingual alignment is achieved through a lightweight trainable projection head applied to text embeddings, described next.

3.3 Image-Pivot Text Projection

To align multilingual text embeddings with the visual embedding space, we apply a trainable projection head

$$g_{\text{text}} : \mathbb{R}^d \to \mathbb{R}^d \tag{5}$$

on top of frozen text embeddings:

$$\tilde{\mathbf{t}} = g_{\text{text}}(\mathbf{t}). \tag{6}$$

We consider two lightweight parameterizations of g_{text}: a linear residual mapping and a small residual MLP. All projected text embeddings and image embeddings are ℓ_2-normalized prior to computing similarities.

3.3.1 Linear Residual Head

In the linear setting, we parameterize the projection as a residual affine mapping

$$W \in \mathbb{R}^{d\times d}, \qquad W = I + \Delta W, \tag{7}$$

where I is the identity matrix and ΔW is the only trainable parameter. Projected text embeddings are computed as

$$\tilde{\mathbf{t}} = \mathbf{t}W. \tag{8}$$

This formulation encourages the learned mapping to remain close to the original pretrained embedding space unless the data provides strong evidence for deviation, enabling controlled and interpretable alignment.

3.3.2 Nonlinear MLP Head

To introduce additional capacity while remaining lightweight, we also consider a nonlinear projection head implemented as a small residual MLP. Given a frozen text embedding $\mathbf{t} \in \mathbb{R}^d$, the MLP head computes

$$\tilde{\mathbf{t}} = \mathrm{MLP}(\mathbf{t}), \tag{9}$$

where

$$\mathrm{MLP}(\mathbf{t}) = \mathbf{t} + \phi(\mathbf{t}W_1)\, W_2, \tag{10}$$

with $W_1 \in \mathbb{R}^{d\times h}$, $W_2 \in \mathbb{R}^{h\times d}$, and element-wise nonlinearity ϕ (e.g., ReLU or GELU).

The residual formulation preserves the inductive bias of staying close to the pretrained embedding space while allowing local nonlinear corrections. As in the linear case, embeddings are ℓ_2-normalized before similarity computation, and the same image-pivoted contrastive objective is used for training.

3.4 Image-Pivoted Contrastive Learning

Crucially, the objective contains only image–text terms: captions in different languages are neither paired as positives nor contrasted directly; cross-lingual coupling arises only through sharing the same image embedding within the batch.

Training is performed using an image-pivoted contrastive objective. For a mini-batch of images $\{I_i\}_{i=1}^{B}$, we sample one caption per language for each image, yielding a batch of $B \times |\mathcal{L}|$ text embeddings. The corresponding image embeddings are replicated across languages, ensuring that captions in different languages describing the same image are aligned through a shared visual anchor.

We use a symmetric InfoNCE loss. Let

$$V \in \mathbb{R}^{M \times d} \tag{11}$$

denote the batch of image embeddings and

$$\tilde{T} \in \mathbb{R}^{M \times d} \tag{12}$$

the batch of projected text embeddings, where $M = B \times |\mathcal{L}|$. The similarity logits are given by

$$S = \frac{\tilde{T} V^\top}{\tau}, \tag{13}$$

where τ is a temperature hyperparameter.

The training objective is

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\Big(\mathcal{L}_{\text{CE}}(S) + \mathcal{L}_{\text{CE}}(S^\top)\Big), \tag{14}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss with matching image–text pairs as positives and all others as in-batch negatives.

This formulation implicitly encourages captions from different languages that describe the same image to align, without requiring any explicit cross-lingual supervision.

3.5 Regularization and Stability

Beyond optimization stability, these terms act as inductive biases that preserve the geometry of the pretrained space, which is crucial for interpretability of representation-level changes.

To further stabilize training and preserve representational structure, we optionally apply two regularization terms.

Proximity-to-Identity Regularization.

$$\mathcal{L}_{\text{prox}} = \|\Delta W\|_F^2, \tag{15}$$

which discourages large deviations from the identity mapping.

Soft Orthogonality Regularization.

$$\mathcal{L}_{\text{ortho}} = \|W^{\top}W - I\|_F^2, \tag{16}$$

which encourages the projection to preserve angles and distances in the embedding space.

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{prox}}\mathcal{L}_{\text{prox}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}, \tag{17}$$

where λ_{prox} and λ_{ortho} are scalar hyperparameters.

3.6 Optimization

We optimize the model using Adam with weight decay and a cosine learning rate schedule with warm-up. Training is performed for a fixed number of epochs with language-balanced mini-batches to prevent dominance of high-resource languages. Early stopping is applied using retrieval performance on a held-out validation fold (macro-averaged Recall@1), and we evaluate using 5-fold cross-validation.

3.7 Discussion

By restricting learning to a lightweight projection head and using images as the sole cross-lingual signal, our method isolates the effect of visual grounding on multilingual alignment. This design enables a controlled study of representation geometry while remaining computationally efficient and easily reproducible.

4 Experimental Setup

4.1 Data

4.1.1 Dataset construction and final size

We build a custom multilingual image–caption dataset starting from Wikipedia-based image–text metadata (WIT v1 split files), and package it into WebDataset shards for efficient streaming. Each sample is keyed by a stable image identifier (derived from the image URL / metadata and used consistently across languages) to ensure that captions in different languages can be associated through a shared image pivot.

Filtering and validation. We apply the following filtering and QA steps: (i) we keep only samples with non-empty captions for each selected language; (ii) we enforce a strict language intersection constraint, retaining only images that have captions available in *all* the languages in $\mathcal{L}$ (ar, de, en, es, fr, it, ja, pt, zh); (iii) we validate images by decoding and integrity checks (PIL `verify`), while rejecting non-image payloads (e.g., HTML/JSON) and handling problematic formats (e.g., SVG rasterization, TIFF detection) before re-encoding to a robust RGB format.

Final dataset size. After filtering, the final aligned dataset contains $N = 2770$ unique images, each with $|\mathcal{L}| = 9$ captions (one per language), for a total of $N \times |\mathcal{L}| = 24930$ image–caption pairs. In our 5-fold cross-validation protocol, each holdout fold contains 554 images (and thus 554 queries per language for text-to-image retrieval), while the remaining 2216 images are used for training in that fold.

We train and evaluate on a multilingual image–caption dataset where each image is associated with captions in multiple languages. Importantly, we do not use any text–text supervision: captions in different languages are never paired as positives, no machine translation is used, and no parallel sentence-level alignment is provided to the model. Cross-lingual coupling arises only indirectly through the shared image identifier (image pivot).

We consider a diverse set of languages spanning different language families and scripts, including both high-resource and lower-resource languages. To prevent dominance of high-resource languages during training, we employ language-balanced sampling, ensuring that each mini-batch contains an equal number of image–caption pairs per language.

For evaluation, we use held-out folds of the same dataset under cross-validation. Parallel captions are used only for cross-lingual caption retrieval via image pivoting and for RSA comparisons across languages; they are never used as training positives.

4.2 Model Architecture

Frozen encoders (exact backbones). Image embeddings are extracted with OpenCLIP `ViT-B-32` initialized with the `openai` pretrained tag, and ℓ_2-normalized for cosine similarity. Text embeddings are extracted with the Sentence-Transformers model `sentence-transformers/clip-ViT-B-32-multilingual-v1`, which maps multilingual text directly into a CLIP-compatible embedding space with matched dimensionality. This choice makes the identity baseline meaningful (CLIP-compatible, dimension-matched, cosine-ready embeddings) and isolates the contribution of image-pivot alignment learned by the lightweight projection head.

Our framework follows a CLIP-style dual-encoder architecture, consisting of a vision encoder and a text encoder that map images and captions into a shared embedding space.

The vision encoder is initialized from a pretrained visual backbone and kept frozen in the primary experimental setting, allowing us to focus on multilingual alignment on the text side. The text encoder is a multilingual CLIP-compatible text model (Sentence-Transformers) pretrained for cross-lingual alignment in the CLIP space, followed by a lightweight trainable projection head (either a linear residual mapping or a small MLP) that maps textual representations into the visual embedding space.

In our main setting, both encoders are frozen and we optimize only a lightweight text-side projection head (either linear residual or MLP-based) on top of the pretrained embeddings, which isolates the effect of image-pivot alignment while keeping computation minimal and analysis controlled.

4.3 Training Objective

We train the model using a symmetric contrastive InfoNCE loss over image–text pairs. Given a batch of image embeddings and their corresponding caption embeddings, the objective encourages matching image–text pairs to have high cosine similarity while treating all other pairs in the batch as negatives.

To encourage cross-lingual consistency, we exploit the image-pivot structure of the data: for a given image, captions from different languages are sampled together within the same batch. This design ensures that captions describing the same visual content—but expressed in different languages—are implicitly encouraged to align through their shared association with the image.

Note that the loss is defined only over image–text pairs; no caption–caption contrastive term is used. We adopt a temperature-scaled contrastive loss and train using the Adam optimizer with a cosine learning rate schedule and warm-up. Additional regularization terms are optionally applied to stabilize training and preserve representational structure, including a proximity-to-identity constraint on the projection matrix and a soft orthogonality penalty.

4.4 Baselines

Why frozen backbones. Freezing both towers is a deliberate design choice: it reduces compute, prevents overfitting in small datasets, and preserves the geometry of the pretrained space, making representation-level comparisons before/after alignment interpretable. In many practical deployments (small datasets, specialized domains), training only a lightweight head is also the most feasible adaptation strategy.

We intentionally avoid end-to-end fine-tuning baselines, as our goal is to isolate the effect of image-pivot supervision under minimal capacity, rather than maximizing absolute retrieval performance.

We consider the pretrained embedding space (identity mapping) as our primary baseline, and compare it against image-pivot alignment learned through a residual projection head (linear residual or MLP-based). This controlled setup isolates the contribution of the image-pivot objective without confounding factors such as translation or additional supervision.

4.5 Evaluation Tasks and Metrics

Retrieval pool size. To interpret Recall@K values, it is important to note that retrieval is performed within each held-out fold. In 5-fold CV, each evaluation pool contains 554 candidate images. Therefore, text-to-image retrieval uses 554 queries per language and 554 candidate images per fold (one correct match per query). This relatively small pool size makes Recall@1 values higher than large-scale benchmarks, and we report mean ± std across folds to mitigate variance.

We evaluate models primarily on multilingual image–text retrieval, considering both text-to-image and image-to-text directions. Performance is reported using standard retrieval metrics, including Recall@1, Recall@5, Recall@10, and Mean Reciprocal Rank (MRR), computed separately for each language and macro-averaged across languages.

To assess cross-lingual alignment beyond retrieval accuracy, we also perform cross-lingual caption retrieval via image pivoting: given a caption in one language, we retrieve the corresponding image and then measure how well captions in another language associated with that image are recovered.

4.6 Representation Analysis

A central focus of our evaluation is the analysis of learned representations before and after image-pivot alignment. We employ several complementary diagnostics:

- **Representational similarity analysis (RSA)**, using correlations between Gram matrices of text embeddings across languages to quantify structural alignment.
- **Intrinsic dimensionality**, measured via effective rank and the number of principal components required to explain a fixed fraction of variance.
- **Isotropy**, assessed through mean pairwise cosine similarity.
- **Feature activation statistics**, including the percentage of near-zero activations and per-dimension entropy, to detect representation collapse or degeneration.
- **Hubness**, measured via k-NN in-degree statistics (skewness and top-1% neighbor mass).
- **Language-identification probing**, via a linear probe trained to predict language from embeddings.
- **Neighborhood overlap**, measured as cross-lingual k-NN neighborhood consistency (reported for $k = 10$).

These analyses provide insight into how multilingual alignment reshapes the geometry of the embedding space and help distinguish meaningful semantic convergence from superficial alignment effects.

4.7 Implementation Details

All experiments are implemented in PyTorch. We optimize only the text-side projection head g_{text} on top of frozen image and text embeddings, using either the linear residual head ($W = I + \Delta W$) or the residual MLP head described in Section 3.3.2. Unless otherwise stated, model selection is performed via 5-fold cross-validation on the training data: for each fold, the head parameters are trained on 4 folds and early stopping is applied based on macro-averaged Recall@1 computed on the held-out fold. We report mean ± standard deviation across folds.

5 Results

5.1 Multilingual Image–Text Retrieval Performance

We first evaluate our approach on the task of multilingual image–text retrieval, which directly measures how well textual representations in different

	Identity (Before)	After (Linear)	After (MLP)
R@1	0.4735 ± 0.0077	0.4951 ± 0.0051	0.5049 ± 0.0055
R@5	0.7284 ± 0.0052	0.7481 ± 0.0054	0.7506 ± 0.0072
R@10	0.8091 ± 0.0067	0.8251 ± 0.0039	0.8260 ± 0.0053
MRR	0.5898 ± 0.0043	0.6096 ± 0.0026	0.6159 ± 0.0028

Table 1: Multilingual text-to-image retrieval on holdout folds. Mean ± std over folds, macro-averaged across languages. (**5-fold** CV)

	R@1					MRR				
Lang	Before	Linear	MLP	ΔLin	ΔMLP	Before	Linear	MLP	ΔLin	ΔMLP
ar	0.3335	0.3591	0.3685	+0.0256	+0.0350	0.4474	0.4740	0.4792	+0.0265	+0.0317
de	0.5043	0.5235	0.5325	+0.0192	+0.0282	0.6174	0.6356	0.6429	+0.0182	+0.0255
en	0.6019	0.6167	0.6275	+0.0148	+0.0257	0.7101	0.7223	0.7301	+0.0121	+0.0200
es	0.5159	0.5365	0.5520	+0.0206	+0.0361	0.6307	0.6479	0.6583	+0.0172	+0.0276
fr	0.5105	0.5249	0.5462	+0.0145	+0.0358	0.6260	0.6396	0.6503	+0.0136	+0.0243
it	0.4993	0.5202	0.5289	+0.0210	+0.0296	0.6170	0.6349	0.6425	+0.0178	+0.0254
ja	0.3562	0.3855	0.3847	+0.0293	+0.0285	0.4786	0.5093	0.5085	+0.0306	+0.0299
pt	0.5228	0.5484	0.5560	+0.0257	+0.0332	0.6375	0.6563	0.6621	+0.0189	+0.0247
zh	0.4169	0.4415	0.4473	+0.0246	+0.0304	0.5439	0.5665	0.5691	+0.0227	+0.0252

Table 2: Per-language text-to-image retrieval performance on holdout folds. Deltas are computed w.r.t. the identity baseline.

languages align with a shared visual embedding space. We report results for the text-to-image direction, using Recall@1, Recall@5, Recall@10, and Mean Reciprocal Rank (MRR). All metrics are computed per language and macro-averaged across languages.

5.1.1 Text-to-Image Retrieval

Unless otherwise specified, all retrieval results are reported as mean ± standard deviation over 5-fold cross-validation, macro-averaged across the 9 languages (ar, de, en, es, fr, it, ja, pt, zh).

We observe consistent improvements after alignment across all metrics. In particular, Linear improves macro R@1 by 0.0217 and macro MRR by 0.0198, while MLP yields larger gains (0.0314 R@1, 0.0260 MRR), and the effect is stable across folds, as indicated by the low standard deviation in Table 1.

Since we adapt only a lightweight projection head on top of frozen encoders, absolute improvements are bounded; nonetheless, gains are consistent across folds and languages, indicating a robust alignment effect.

	Identity (Before)	After (Linear)	After (MLP)
R@1	0.4808 ± 0.0056	0.4442 ± 0.0091	0.4551 ± 0.0045
R@5	0.7255 ± 0.0068	0.6871 ± 0.0125	0.7000 ± 0.0070
R@10	0.8064 ± 0.0100	0.7745 ± 0.0116	0.7840 ± 0.0091
MRR	0.5937 ± 0.0047	0.5567 ± 0.0098	0.5676 ± 0.0049

Table 3: Multilingual image-to-text retrieval on holdout folds. Mean ± std over folds, macro-averaged across languages. (**5-fold** CV)

Interestingly, the MLP head yields the strongest gains in retrieval performance, whereas the linear residual head produces the most consistent improvements in cross-lingual representational similarity (RSA), indicating a trade-off between instance-level retrieval accuracy and global structural alignment.

5.1.2 Per-Language Analysis

Table 2 reports Recall@1 and MRR (closely related to AP in a single-positive retrieval setup) for each individual language. Performance improvements are observed across all languages, including those with different scripts and morphological properties.

Languages with weaker initial alignment, such as Arabic, Japanese, and Chinese, benefit particularly from image-pivot training, showing larger relative gains in both Recall@K and MRR. High-resource languages such as English, German, and French also improve, albeit with smaller absolute gains, reflecting their stronger initial alignment inherited from pretrained representations.

This pattern suggests that image-pivot supervision is especially effective at reducing cross-lingual disparities, narrowing the performance gap between high-resource and lower-resource languages.

5.1.3 Image-to-Text Retrieval

Asymmetry between text-to-image and image-to-text retrieval. We observe that, in contrast to the consistent gains in text-to-image retrieval, image-to-text retrieval performance decreases after image-pivot alignment, with a larger drop under the linear head and a partial recovery under the MLP head (Table 3).

This behavior is expected given our training setup.

	R@1					MRR				
Lang	Before	Linear	MLP	ΔLin	ΔMLP	Before	Linear	MLP	ΔLin	ΔMLP
ar	0.3436	0.3103	0.3194	-0.0332	-0.0242	0.4562	0.4210	0.4292	-0.0352	-0.0270
de	0.5108	0.4671	0.4801	-0.0437	-0.0307	0.6221	0.5818	0.5950	-0.0403	-0.0272
en	0.6055	0.5726	0.5827	-0.0329	-0.0228	0.7104	0.6799	0.6902	-0.0305	-0.0202
es	0.5285	0.4931	0.5036	-0.0354	-0.0249	0.6393	0.6031	0.6151	-0.0362	-0.0242
fr	0.5231	0.4787	0.4964	-0.0444	-0.0267	0.6301	0.5922	0.6065	-0.0379	-0.0236
it	0.5090	0.4751	0.4859	-0.0340	-0.0231	0.6231	0.5858	0.5975	-0.0373	-0.0256
ja	0.3678	0.3298	0.3396	-0.0379	-0.0282	0.4893	0.4473	0.4576	-0.0420	-0.0317
pt	0.5217	0.4870	0.5000	-0.0347	-0.0217	0.6357	0.5993	0.6119	-0.0364	-0.0238
zh	0.4173	0.3837	0.3880	-0.0336	-0.0293	0.5366	0.4993	0.5054	-0.0372	-0.0311

Table 4: Per-language image-to-text retrieval performance on holdout folds. Deltas are computed w.r.t. the identity baseline.

	Before	After (Linear)	After (MLP)
R@1	0.3265 ± 0.0078	0.3171 ± 0.0019	0.3313 ± 0.0030
R@5	0.5518 ± 0.0072	0.5435 ± 0.0076	0.5638 ± 0.0055
R@10	0.6410 ± 0.0092	0.6374 ± 0.0067	0.6567 ± 0.0050
MRR	0.4336 ± 0.0063	0.4253 ± 0.0032	0.4410 ± 0.0032

Table 5: Cross-lingual caption retrieval via image pivot (macro-averaged over ordered language pairs).

Although we optimize a symmetric contrastive objective, adaptation is restricted exclusively to the text-side projection head, while the image encoder remains frozen.

As a result, text embeddings are explicitly optimized to better retrieve their paired images, but image embeddings are not optimized to retrieve text.

This induces a geometric asymmetry in the shared space, favoring text-to-image retrieval at the expense of image-to-text performance.

Introducing a trainable projection on the image side would restore symmetry between the two retrieval directions, but would also undermine the role of images as fixed language-agnostic pivots.

We therefore intentionally accept this trade-off in order to isolate and analyze the effects of image-pivot supervision on multilingual text representations.

5.1.4 Cross-Lingual Caption Retrieval via Image Pivot

Cross-lingual caption retrieval via image pivot highlights a complementary trade-off between alignment capacity and structural convergence.

While the linear head improves global representational similarity, it yields marginal or slightly negative effects on pivot-based retrieval, which requires accurate instance-level alignment across two retrieval steps.

In contrast, the MLP head significantly improves pivot retrieval performance, suggesting that local nonlinear corrections are more effective for cross-lingual coupling mediated by visual grounding.

To directly assess cross-lingual coupling induced by visual grounding, we evaluate cross-lingual caption retrieval via image pivoting: given a caption in a source language, we retrieve the corresponding image and then retrieve captions in a target language associated with that image.

Table 5 shows that the MLP head improves pivot-based cross-lingual retrieval, whereas the linear head yields marginal or slightly negative changes, consistent with the capacity-versus-structure trade-off observed in our representation analysis.

5.1.5 Effect of Balanced Multilingual Batching

We use language-balanced batching (equal number of samples per language within each batch) to prevent high-resource languages from dominating the contrastive objective. We found this choice important for stable multilingual alignment in practice.

5.1.6 Summary

Overall, these results demonstrate that image-pivoted contrastive learning is sufficient to induce strong multilingual alignment, yielding consistent retrieval improvements across a diverse set of languages without relying on parallel text or translation-based supervision.

6 Representation Analysis

While retrieval metrics quantify the effectiveness of multilingual alignment, they do not reveal how alignment reshapes the underlying representation space. We therefore conduct an in-depth analysis of text embeddings before and after image-pivot alignment, focusing on representational similarity, geometric properties, and feature statistics. All analyses are performed separately per language and summarized using macro averages.

	Before	After (Linear)	After (MLP)
Effective rank	112.24 ± 1.71	107.15 ± 1.96	112.18 ± 1.61
PCA 90% components	146.04 ± 1.27	142.09 ± 1.69	145.38 ± 1.22
Mean cosine similarity	0.660 ± 0.004	0.618 ± 0.014	0.618 ± 0.004
PoZ	0.0363 ± 0.0003	0.0354 ± 0.0006	0.0349 ± 0.0003
Entropy	0.745 ± 0.002	0.781 ± 0.009	0.777 ± 0.002
Gram corr. mean	0.345 ± 0.021	0.351 ± 0.021	0.346 ± 0.022
Neighborhood overlap ($k = 10$)	0.083 ± 0.003	0.088 ± 0.003	0.083 ± 0.002
Hubness skew ($k = 10$)	4.889 ± 0.173	4.985 ± 0.172	5.045 ± 0.174
Hub ratio (top 1%)	0.135 ± 0.005	0.135 ± 0.005	0.140 ± 0.006
Lang-ID probe acc.	0.210 ± 0.008	0.210 ± 0.007	0.213 ± 0.007

Table 6: Representation diagnostics on holdout folds (macro-averaged across languages, mean ± std over folds).

6.0.1 Cross-Lingual Representational Similarity

To measure structural alignment across languages, we perform Representational Similarity Analysis (RSA) by computing correlations between Gram matrices of text embeddings across languages. Specifically, for each language we compute the cosine similarity matrix over a shared set of samples, and then measure Pearson correlation between the flattened upper triangles of these matrices across language pairs.

Gram matrices capture relational structure up to orthogonal transformations, making them a natural tool to compare embedding geometries across languages.

After image-pivot alignment, cross-lingual representational similarity increases under the linear residual head, while the MLP head improves retrieval without necessarily increasing RSA (Table 6). On the holdout folds, the mean off-diagonal Gram correlation increases from 0.345 ± 0.021 to 0.351 ± 0.021 under the linear head, while remaining stable under the MLP head. These results suggest that alignment is not only reflected in instance-level retrieval, but also in a modest, consistent convergence of relational structure across languages.

This divergence between retrieval improvements and RSA gains highlights that stronger downstream performance does not necessarily imply deeper convergence of representational geometry. This increase appears broadly distributed across language pairs rather than driven by a small subset of dominant languages, suggesting a global effect rather than a collapse-driven artifact.

We complement RSA with a cross-lingual neighborhood overlap metric, which measures the consistency of local k-NN neighborhoods across languages. As shown in Table 6, neighborhood overlap slightly increases under linear alignment, indicating improved local structural consistency across languages, while remaining largely unchanged under the MLP head.

6.0.2 Intrinsic Dimensionality and Effective Rank

We next analyze how image-pivot alignment affects the intrinsic dimensionality of text embeddings. We measure effective rank, defined as the exponential of the entropy of the singular value spectrum, as well as the number of principal components required to explain 90% of variance.

Across languages, image-pivot alignment leads to a moderate reduction in effective rank on the holdout folds (from 112.24 ± 1.71 to 107.15 ± 1.96 under the linear head), while the MLP head largely preserves the original dimensionality. Similarly, the number of components required to explain 90% of variance decreases by approximately 4–5 components on average.

This behavior suggests that alignment induces a controlled compression of the representation space, removing redundant or language-specific variation while preserving semantic capacity. Crucially, the reduction in dimensionality is gradual and consistent across languages, rather than abrupt, which would be indicative of representational collapse.

Notably, this controlled reduction in effective rank is primarily observed under the linear head, while the MLP largely preserves the original dimensionality, consistent with its weaker impact on global structural alignment.

6.0.3 Isotropy and Hubness

To assess the global geometry of the embedding space, we compute mean pairwise cosine similarity among text embeddings as a proxy for isotropy. High mean similarity values indicate anisotropic spaces dominated by a small number of directions, which are known to harm retrieval performance.

After alignment, mean pairwise cosine similarity decreases substantially on the holdout folds under both heads (from 0.660 ± 0.004 to 0.618 ± 0.014 under the linear head, and to 0.618 ± 0.004 under the MLP head), indicating improved isotropy. Direct hubness measurements based on k-NN in-degree statistics remain broadly stable (Table 6); the MLP head shows a slight increase in hub ratio (top 1%), but there is no qualitative change that could explain the retrieval gains. These trends suggest that image-pivot alignment improves global geometry without inducing additional hub dominance.

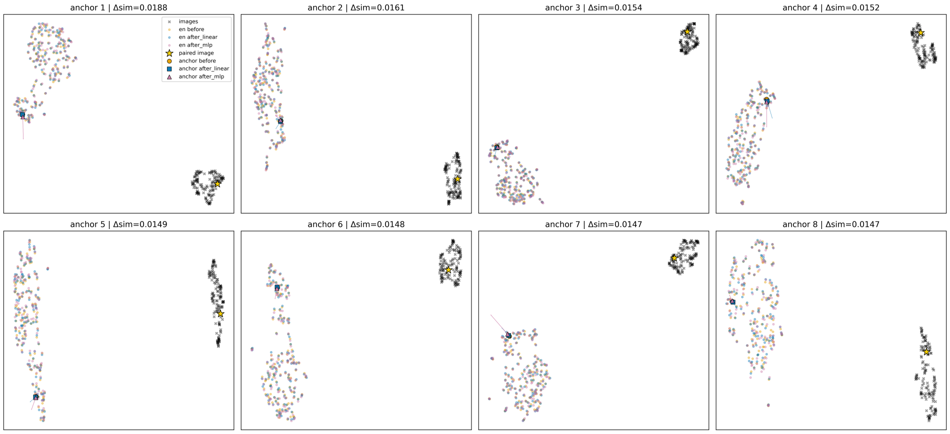

Figure 1: Local UMAP visualization of image–text alignment micro-corrections (fold 0). For each panel, we select an English anchor caption and visualize a local neighborhood of images (black ×) and text embeddings. Colored circles, squares, and triangles denote the anchor caption before alignment, after linear alignment, and after MLP alignment, respectively. Arrows indicate the displacement from the pretrained embedding (before) to the representations obtained after linear and MLP-based image-pivot alignment (arrows are scaled for visibility). The gold star marks the paired image associated with the anchor caption. The figure illustrates that alignment emerges through small, structured local corrections rather than global rearrangements of the embedding space.

This geometric improvement is consistent with the observed retrieval gains, supporting the hypothesis that improved isotropy—rather than increased hubness—contributes to better cross-lingual alignment.

6.0.4 Local Micro-Corrections (Qualitative UMAP)

Global 2D projections of the full embedding space are visually uninformative in our setting because the learned alignment heads apply small, local updates. To make these micro-corrections visible, we visualize local neighborhoods around captions that exhibit the largest increase in cosine similarity to their paired image after alignment.

For each anchor caption (English pivot), we fit UMAP on a local set containing neighboring images and the corresponding text embeddings (before, after-linear, after-MLP), and overlay the anchor displacement. This visualization is qualitative and is not used to derive quantitative claims.

6.0.5 Feature Activation Statistics

Finally, we examine low-level feature statistics to detect potential degeneration of representations. We measure the percentage of near-zero activations (PoZ) and the entropy of unit activations for each embedding dimension.

PoZ is computed as the fraction of activations with $|x| < 10^{-3}$; entropy is computed via per-dimension histograms with 30 bins over $[-1, 1)$.

Across all languages, PoZ slightly decreases on the holdout folds (from 0.0363 ± 0.0003 to 0.0354 ± 0.0006 under the linear head), while average entropy increases (from 0.745 ± 0.002 to 0.781 ± 0.009 under the linear head). These trends are consistent with slightly more evenly distributed activations and improved feature utilization after alignment.

6.0.6 Language-ID Probe

To assess how much language-identifying information is retained after alignment, we train a linear language-ID probe (multinomial logistic regression) to predict the caption language from frozen text embeddings. Probe accuracy remains stable after image-pivot alignment (Table 6), suggesting that visual grounding improves multilingual structure without aggressively erasing language-specific signals.

6.0.7 Summary

Overall, image-pivoted contrastive learning induces multilingual alignment through structured geometric convergence when using a linear residual head, while nonlinear MLP heads primarily improve retrieval performance through increased capacity. The combined diagnostics indicate that alignment improves isotropy and cross-lingual structure without inducing representational collapse, excessive hubness, or strong language invariance. These results underscore the importance of complementing downstream metrics with representation-level diagnostics when evaluating multilingual multimodal alignment.

7 Limitations and Ethical Considerations

7.1 Limitations

While our results demonstrate that image-pivoted contrastive learning is effective for multilingual alignment, our approach has several limitations.

First, our method relies on the availability of images paired with captions in multiple languages. Although we do not use parallel text during training, the presence of multilingual captions for the same image is still more common for high-resource languages and popular visual content. As a result, alignment quality may degrade for languages or domains where such data is scarce.

Second, we restrict learning to lightweight text-side projection heads on top of frozen pretrained encoders (linear residual or small MLP), rather than end-to-end fine-tuning. While this design choice enables controlled analysis and computational efficiency, it may limit the expressiveness of the alignment compared to adapting the text encoder (or both encoders). More complex nonlinear mappings or deeper adaptation could further improve performance, at the cost of interpretability and training stability.

Third, our evaluation focuses primarily on image–text retrieval and representation geometry. While these tasks are well-suited to measure cross-lingual alignment, they do not capture all aspects of multilingual understanding, such as compositional reasoning, pragmatic meaning, or culturally grounded concepts that may not be visually grounded.

We quantify language-identifying signal via a linear language-ID probe on text embeddings; however, this is only a coarse diagnostic and does not fully separate beneficial alignment from potential loss of language-specific nuances.

Finally, although we analyze representation health using several diagnostics, we do not fully disentangle language invariance from language loss. Some degree of language-specific information may be beneficial for certain downstream tasks, and stronger alignment does not necessarily imply better performance in all multilingual settings.

7.2 Ethical Considerations

Our work uses large-scale web data, which may contain biases, stereotypes, or uneven representation across languages and cultures. Image–caption datasets are known to reflect societal biases present in online content, and multilingual captions may differ systematically in style, descriptiveness, or cultural framing. These factors can influence both model behavior and evaluation outcomes.

By relying on images as alignment pivots, our method may also amplify visual bias, privileging concepts that are easily depicted visually while underrepresenting abstract, relational, or culturally specific meanings. This limitation is particularly relevant for languages whose semantic distinctions are less frequently grounded in visual content.

We emphasize that our goal is analysis and understanding, rather than deployment in high-stakes applications. Any real-world use of multilingual vision–language models should be accompanied by careful dataset auditing, bias evaluation, and task-specific validation, especially when applied to underrepresented languages or communities.

8 Conclusion

In this work, we investigated multilingual vision–language representation learning through an image-pivoted contrastive framework, where images serve as language-agnostic anchors and no parallel text is used during training. By restricting learning to a lightweight text-side projection head (primarily a residual linear mapping) and leveraging visual grounding as the sole cross-lingual signal, we were able to isolate and study how multilingual alignment emerges in CLIP-style models.

Our experiments show that image-pivoted alignment consistently improves multilingual image–text retrieval across a diverse set of languages, including those with different scripts and resource levels. Importantly, improvements are measured on held-out folds under cross-validation, indicating that the alignment generalizes beyond the training subset used for optimization and is not driven by translation-based supervision or explicit cross-lingual constraints.

Beyond performance, we conducted an in-depth representation analysis to understand the structural effects of multilingual alignment. We showed that alignment manifests as a gradual convergence of representational geometry across languages, characterized by increased cross-lingual similarity, improved isotropy, and more evenly distributed activations and improved feature utilization. At the same time, intrinsic dimensionality is reduced in a controlled manner, suggesting that alignment removes redundant language-specific variation without inducing representational collapse.

Additional diagnostics confirm that these gains are not explained by degenerate effects. Direct hubness measurements remain stable after alignment, and a linear language-identification probe shows stable accuracy, indicating that improved alignment does not arise from increased hub dominance or aggressive suppression of language-identifying signal.

Taken together, our findings show that visual grounding alone constitutes a strong supervisory signal for multilingual alignment, and that the nature of the learned alignment depends critically on the capacity of the projection head: linear mappings favor structural convergence, while nonlinear heads yield larger retrieval gains with weaker changes in global cross-lingual geometry (e.g., RSA and neighborhood overlap), and without clear signs of degeneration in hubness or activation statistics. They also highlight the importance of going beyond downstream metrics and analyzing representation geometry when studying multilingual multimodal models.

We hope that this work encourages further research into lightweight, interpretable approaches to multilingual alignment, as well as more systematic use of representation-level diagnostics to better understand the behavior and limitations of large multimodal models.

References

[1] Ahtamjan Ahmat. M^2-vlp: Enhancing multilingual vision-language pre-training via multi-grained alignment. *OpenReview*, 2024.

[2] Guanhua Chen, Zhenzhong Han, and Guang Zeng. mclip: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[3] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[4] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask retrieval across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[5] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[6] Alireza Mohammadshahi, Rémi Lebret, and Karl Aberer. Aligning multilingual word embeddings for cross-modal retrieval task. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[7] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.