
Scaling Language-Centric Omnimodal Representation Learning

Chenghao Xiao¹ Hou Pong Chan^{1,†} Hao Zhang^{1,†}
Weiwen Xu¹ Mahani Aljunied¹ Yu Rong^{1,2,‡}

¹DAMO Academy, Alibaba Group ²Hupan Lab

Abstract

Recent multimodal embedding approaches leveraging multimodal large language models (MLLMs) fine-tuned with contrastive learning (CL) have shown promising results, yet the underlying reasons behind their superiority remain underexplored. This work argues that a crucial advantage of MLLM-based approaches stems from implicit cross-modal alignment achieved during generative pretraining, where the language decoder learns to exploit multimodal signals within a shared representation space for generating unimodal outputs. Through analysis of anisotropy and kernel similarity structure, we empirically confirm that latent alignment emerges within MLLM representations, allowing CL to serve as a lightweight refinement stage. Leveraging this insight, we propose a **Language-Centric Omnimodal Embedding** framework, termed **LCO-EMB**. Extensive experiments across diverse backbones and benchmarks demonstrate its effectiveness, achieving state-of-the-art performance across modalities. Furthermore, we identify a **Generation-Representation Scaling Law (GRSL)**, showing that the representational capabilities gained through contrastive refinement scale positively with the MLLM’s generative capabilities. This suggests that improving generative abilities evolves as an effective paradigm for enhancing representation quality. We provide a theoretical explanation of GRSL, which formally links the MLLM’s generative quality to the upper bound on its representation performance, and validate it on a challenging, low-resource visual-document retrieval task, showing that continual generative pretraining before CL can further enhance the potential of a model’s embedding capabilities.¹

1 Introduction

Cross-modal representation alignment, such as vision-language alignment, has traditionally relied on massive-scale contrastive learning (CL) over paired cross-modal data, as seen in CLIP-style models [37, 50, 83]. Prior work primarily focuses on scaling model size, dataset volume, and batch size during training [9, 25, 50, 58, 83]. While these strategies demonstrate benefits in tasks like linear probing [9, 25, 50] and zero-shot classification [50, 83], performance tends to plateau on complex tasks requiring deeper cross-modal comprehension, *e.g.*, multilingual image retrieval [57, 60], visual text representations [11, 19, 74], and tasks involving interleaved multimodal encodings [69].

Recent approaches utilize autoregressive multimodal large language models (MLLMs) as the backbone models, followed by CL fine-tuning, to enhance representational capabilities, leading to improved performance on these complicated tasks [8, 39, 84]. However, the underlying reasons for the performance advantages of MLLM-based embedding approaches over traditional CLIP-based ones

[†]Corresponding authors, kenchanhp@gmail.com, h Zhang26@outlook.com.

[‡]Project head.

¹Codes, models, and resources are available at <https://github.com/LCO-Embedding/LCO-Embedding>.

remain underexplored. This represents a critical research gap in understanding the limitations of CLIP-style models and the specific strengths MLLMs bring to these challenging scenarios.

To address this research gap, we conduct a systematic study of MLLM-based embedding models across modalities. First, we empirically investigate the embedding space patterns of MLLM representations before and after lightweight CL fine-tuning using only textual data, via anisotropy and kernel-level similarity. Our results show that *text-only* fine-tuning not only improves the discriminability of text embeddings but also generalizes to enhance the discriminability of embeddings in non-textual modalities. This finding reveals that *MLLMs achieve implicit cross-modal alignment during generative pretraining*, such that representation activation for one modality generalizes to others. We posit that the generative objective of MLLMs enables them to leverage multimodal information in the same semantic space by learning to generate textual outputs during pretraining. Thus, we argue that the knowledge foundation and intrinsic multimodal alignment established during generative pretraining grant MLLM-based embedding models the fundamental advantages.

Building on the observations, we propose a **Language-Centric Omnimodal Embedding** framework, termed LCO-EMB, that employs language-centric paired data for efficient CL refinement. We highlight that *CL can function as a lightweight, post-hoc refinement step for mapping pre-aligned generative embeddings into a similarity-matching space* in MLLMs, which differs sharply from the computationally intensive CL required by CLIPs for alignment. Accordingly, this emerging paradigm shifts emphasis towards preserving the cross-modal alignment structure established during MLLM pretraining. In line with recent work [24, 86], LCO-EMB adopts LoRA [27] for representation activation of MLLM, aiming to enhance its representation capability with minimal disruption to the pretrained generative capabilities and latent multimodal alignment.

Extensive experiments across diverse backbones and benchmarks show that LCO-EMB outperforms state-of-the-art multimodal embedding models trained with much larger multimodal training sets, with text-only training sets. Combining minimal additional multimodal paired data in diverse formats further calibrates the embedding space of LCO-EMB for downstream tasks, setting a new state-of-the-art on MIEB [76], while also providing competitive performance on audio and videos. Further analysis reveals that LoRA with language-centric contrastive learning yields superior results compared to alternative fine-tuning strategies, suggesting the importance of preserving the latent alignment structure during CL through minimal modification to the MLLM’s pretrained knowledge. CL acts less as a means of introducing new knowledge and more as a lightweight activation mechanism, serving primarily to project the embedding space into a similarity-matching subspace.

As LCO-EMB relies on the inherent multimodal alignment capability of MLLMs, we further investigate the relationship between potentials of representation quality and the underlying generative ability of MLLMs. Through experiments with backbones of various sizes and generation strengths, we identify a **Generation-Representation Scaling Law (GRSL)**, indicating that multimodal representational capabilities gained through contrastive refinement scales positively with the MLLM’s generative capabilities before CL. GRSL suggests that improving the MLLM’s initial generative capability—via continued generative pretraining or supervised fine-tuning—is an effective strategy for enhancing its potential in multimodal representations. We offer a theoretical explanation for GRSL through a PAC-Bayesian generalization bound, showing that an MLLM’s generative capability determines an upper bound for its representational potential. To empirically validate this, we introduce **SeaDoc**, the most difficult visual document retrieval task to date in low-resource Southeast Asian languages. Through continual OCR-intense pretraining in low-resource languages, we show that retrieval performance enhances after the same amount of text-only contrastive learning.

Our contributions are threefold: (1) We propose a language-centric omnimodal representation learning framework, achieving promising performance across various MLLM backbones and embedding benchmarks. (2) We identify a Generation-Representation Scaling Law (GRSL), that representational capabilities after CL scales positively with the MLLM’s generative capabilities. (3) We provide a theoretical justification for GRSL, followed by comprehensive empirical studies, demonstrating that generative capability sets a fundamental upper bound on representational quality in MLLMs.

2 Latent Cross-Modal Alignment in MLLMs

In this section, we conduct an in-depth empirical analysis of multimodal large language models (MLLMs) to investigate *whether their internal representations exhibit latent cross-modal alignment*

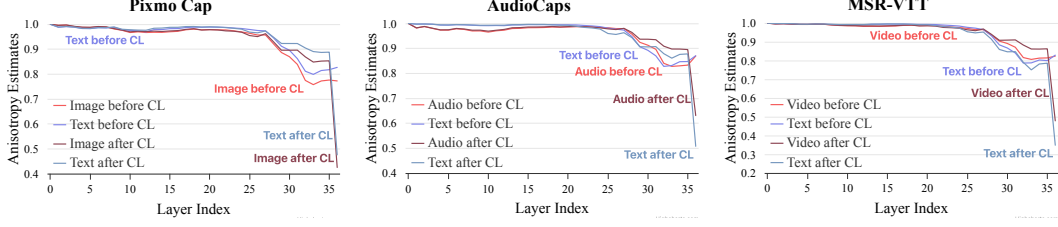


Figure 1: The anisotropy estimates of Qwen2.5-Omni-3B embeddings across text, image, audio, and video modalities. The vanilla model exhibits typical representation degeneration (anisotropy) for all modalities. After applying **text-only** contrastive learning, embeddings across modalities become more isotropic, indicating latent **language-centric** cross-modal alignment within the model.

through two geometric properties, *i.e.*, degree of anisotropy [23] and kernel-level similarity [29]. Specifically, starting with an MLLM, we directly take out its text decoder [30], *i.e.*, the LLM, and fine-tune it using text-only contrastive learning with LoRA on anchor-entailment text pairs from NLI datasets. Then, we merge the trained LoRA weights into the LLM and re-plug it into the original MLLM architecture. The detailed experimental settings are summarized in Section 4.1.

2.1 Analysis of Anisotropy Degrees

Language models trained on self-supervised objectives are known to suffer from anisotropy [17, 21], an embedding degeneration issue characterized by hidden representations collapsing into a confined region of representation space, resulting in high expected cosine similarity between random inputs. Contrastive learning is known to have the uniformity promise [68, 73] through enhancing discriminability across random negative pairs. Here, we employ contrastive learning to fine-tune multimodal large language models (MLLMs) exclusively with paired text data. We then compare the behaviors of models before and after fine-tuning to assess whether text-only training can effectively mitigate anisotropy for non-textual modalities, even in the absence of explicit multimodal training. The successful transfer of improvements across modalities would provide empirical evidence that MLLMs inherently preserve geometrically aligned latent spaces among different modalities.

We follow Ethayarajh [17] and Xiao et al. [73] to approximate the degree of anisotropy using the expected mean of cosine similarity between random data points. Let $\mathbf{h}_i, \mathbf{h}_j \sim \mathcal{D}$ be the embedding vectors sampled independently and identically distributed (*i.i.d.*) from the empirical distribution $\mathcal{D}$ of the representation space. Then, the degree of anisotropy is calculated as:

$$\text{Anisotropy} := \mathbb{E}_{\mathbf{h}_i, \mathbf{h}_j \sim \mathcal{D}} [\cos(\theta_{ij})] = \mathbb{E}_{\mathbf{h}_i, \mathbf{h}_j \sim \mathcal{D}} \left[\frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|} \right]. \quad (1)$$

In practice, we approximate it empirically using a finite sample of N embeddings $\{\mathbf{h}_1, \dots, \mathbf{h}_N\}$ as:

$$\hat{\mathbb{E}}[\cos(\theta)] = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}. \quad (2)$$

Specifically, we use Qwen2.5-Omni-3B [77] as the backbone model and fine-tune it with text-only contrastive learning. To ensure objective and fair semantic comparison between text and other modalities, we utilize paired datasets, *i.e.*, Pixmo Cap [13] for image-text, AudioCaps [32] for audio-text, and MSR-VTT [79] for video-text, for anisotropy comparison. The changes in the embedding spaces of different modalities after the text-only contrastive learning are depicted in Figure 1. As anticipated, the embedding space produced by Qwen2.5-Omni-3B initially exhibits a collapsed structure and poorly separated distribution across modalities. After text-only contrastive learning, embedding spaces of non-text modalities surprisingly generalize to become *more isotropic, dispersing more uniformly across the respective subspaces*. **The generalized reduction in anisotropy for image, audio, and video embeddings reflects an underlying latent semantic alignment with textual representations within the base model.**

2.2 Analysis of Kernel-level Similarity

Building on the latent cross-modal alignment in MLLMs, we further employ kernel-level similarity to analyze the improvement in similarity structure alignment across modalities after fine-tuning.

Given a function $f : \mathcal{X} \rightarrow \mathbb{R}^n$ that maps inputs to high-dimensional representations, the associated kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ characterizes the induced similarity structure via inner product $K(x_i, x_j) = \langle f(x_i), f(x_j) \rangle$, where $x_i, x_j \in \mathcal{X}$ and $K \in \mathcal{K}$. Then, a kernel alignment metric $m : \mathcal{K} \times \mathcal{K} \rightarrow \mathbb{R}$ is adopted to quantify the similarity between two kernels, *i.e.*, the “similarity of similarity structures”, by assessing *how closely the distance metric induced by one representation space aligns with that of another*. Prior work [29] examines these structures across independently trained models and finds convergence in their representations. For instance, despite being trained separately, LLaMA [64] and DINOv2 [49] exhibit comparable similarity perception of captions and images from paired datasets.

Similar to Huh et al. [29], we adopt mutual k NN to quantify the overlap in the top- k nearest neighbors of each data point shared across the similarity structures induced by two models, f and g . The data samples (x_i, y_i) are drawn in mini-batches of size b from a distribution $\mathcal{X}$. Taking the image-text alignment as an example, each (x_i, y_i) pair, *i.e.*, an image and its corresponding caption, is assumed to share the same semantic content, denoted as $x_i \triangleq y_i$. These paired samples serve as semantic anchors for evaluating the representations across modalities. Given the models f and g ,² the corresponding embeddings of each paired sample are attained as $\phi_i = f(x_i)$ and $\psi_i = g(y_i)$. For a mini-batch of b data samples, we can derive the feature sets $\Phi = \{\phi_1, \dots, \phi_b\}$ and $\Psi = \{\psi_1, \dots, \psi_b\}$. For each feature $\phi_i \in \Phi$ (and similarly $\psi_i \in \Psi$), the k NN set $\mathcal{S}(\phi_i)$ (or $\mathcal{S}(\psi_i)$) comprises the indices of its k nearest neighbors within its feature collection, excluding itself, which is determined by d_{knn} as:

$$\mathcal{S}(\phi_i) = d_{\text{knn}}(\phi_i, \Phi \setminus \{\phi_i\}), \quad \mathcal{S}(\psi_i) = d_{\text{knn}}(\psi_i, \Psi \setminus \{\psi_i\}). \quad (3)$$

The kernel-level similarity score m_{NN} for a specific feature pair (ϕ_i, ψ_i) is the normalized cardinality of the intersection of their k NN sets:

$$m_{\text{NN}}(\phi_i, \psi_i) = \frac{1}{k} |\mathcal{S}(\phi_i) \cap \mathcal{S}(\psi_i)|, \quad (4)$$

which indicates the proportion of shared nearest neighbors, where k denotes the number of nearest neighbors. The overall m_{NN} metric is computed as the average of the individual $m_{\text{NN}}(\phi_i, \psi_i)$ scores across the mini-batch.

Different from Huh et al. [29], we utilize kernel alignment metrics to inspect cross-modal alignment within the same model. Specifically, we attain hidden representations at all layers from the 3B and 7B variants of Qwen2.5-VL-Instruct, and assess the self-similarity between the vision and language kernels using Equation 4. The text-only contrastive learning is applied to the language decoder, and alignment scores are compared before and after fine-tuning. As illustrated in Figure 2, two notable findings emerge: (1) cross-modal kernel alignment improves after text-only contrastive learning, indicating the presence of inherent latent alignment across modalities; and (2) the 7B variant exhibits consistently stronger cross-modal kernel alignment than the 3B variant, both before and after fine-tuning. This advantage may be attributed to the expanded parameter space of the larger model, yielding better expressivity and a superior ability to capture latent cross-modal relationships during pre-training. Collectively, these findings suggest that **inherent cross-modal binding enables the optimization of representation in one modality to induce corresponding improvements in other modalities**.

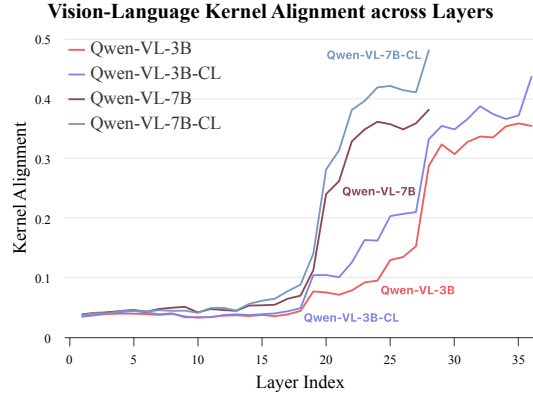


Figure 2: Layer-wise vision-language kernel alignment before and after text-only contrastive learning, evaluated on Qwen-VL models with 7B (28 layers) and 3B (36 layers) parameters. Note the 3B model has more layers than the 7B model.

3 Language-centric Omnimodal Representation Learning

The preliminary experiments reveal that MLLMs implicitly acquire cross-modal alignment during pretraining. Although initial embeddings are suboptimal for similarity matching, latent align-

²In this example, we use the same model to encode the image x_i and its caption y_i , *i.e.*, $f = g$.

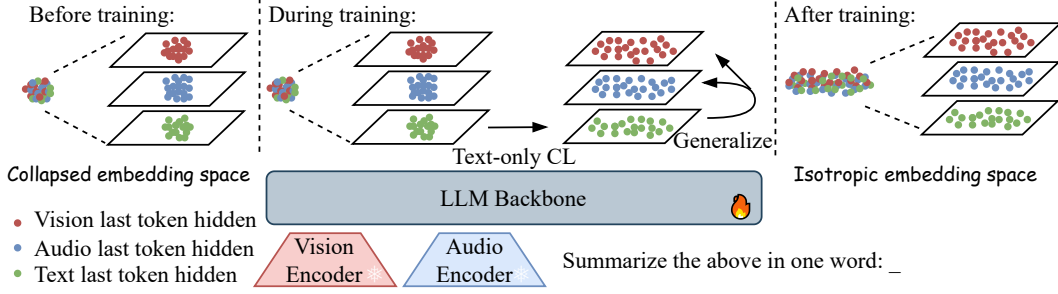


Figure 3: The power of language-centric omnimodal representation learning: Before text-only contrastive learning (CL), representations across modalities in multimodal large language models (MLLMs) exhibit anisotropy, collapsing into a confined subspace. Text-only CL disperses textual representations by increasing their separation, effectively reducing anisotropy. Notably, this process generalizes to alleviate anisotropy in non-textual modalities, despite the absence of direct supervision.

ment emerges in intermediate layers. This inherent alignment can be efficiently unlocked through lightweight text-only contrastive fine-tuning, enhancing representation quality across both textual and non-textual modalities. Building on this insight, we introduce **Language-Centric Omnimodal** representation learning (LCO-EMB), a framework that leverages language-centric data and lightweight contrastive learning to boost MLLM representation capabilities across modalities.

The contemporary architectures of MLLM are composed of modality-specific encoders, a projector, and a language decoder (*i.e.*, an LLM), with the projector aligning modality-specific representations to the decoder’s embedding space [2, 40, 59, 77]. For text-only variants of LCO-EMB, we *isolate and fine-tune only the language decoder via text-only contrastive learning*, while freezing the parameters of modality encoders and the projector. After training, the updated decoder is re-plugged into the original model. We further incorporate minimal multimodal paired data to calibrate the embedding space for downstream tasks, resulting in multimodal variants of LCO-EMB.

Central to our method is the preservation of the latent cross-modal alignment established during generative pretraining. This alignment, wherein multimodal embeddings are integrated into a shared latent subspace by the language decoder, is fundamental to the model’s multimodal representation capability. We employ LoRA [27], which introduces low-rank trainable parameters into select layers while freezing the original model. While LoRA is widely recognized for enabling parameter-efficient fine-tuning, its primary advantage in our context is its ability to minimally perturb the original model. This approach yields two critical benefits: (1) *it preserves the model’s generative capabilities through minimal weight modifications* [4]; and (2) *it maintains the latent cross-modal alignment, especially in the language decoder’s embedding layer, which is unaffected by the adaptation*.

4 Experiments

4.1 Experimental Settings

Backbones and hyperparameters. We use LLaVA-Next [41], Qwen2.5-VL [2], and Qwen2.5-Omni [77] as backbones, all conforming to the standard architecture of modality-specific encoders, a projector, and a language decoder. LLaVA-Next and Qwen2.5-VL focus on image/video-text, while Qwen2.5-Omni supports omnimodal inputs, covering text, image, video, and audio. We utilize the 8B variant of LLaVA-Next, the 3B and 7B variants for both Qwen2.5-VL and Qwen2.5-Omni. We adopt AdamW optimizer with a cosine learning rate schedule, a peak learning rate of 4×10^{-4} , and a batch size of 768,³ to train the model for 2 epochs. The LoRA rank (r) and α are set as 64 and 16 for text-only variants and 64 and 128 for multimodal variants, respectively. For multimodal variants of Qwen2.5-Omni-7B, we use a reduced learning rate of 3×10^{-4} due to the loss spike.

Datasets. (1) *Text-only Setting.* We consider two data settings: **all-NLI** and **Scale-1M**. The **all-NLI** combines MNLI [70] and SNLI [5], both frequently used for sentence representation learning. Each

³For multimodal variants with additional image-text and interleaved data, we scale batch size by the ratio of total to text-only data size. Thus, for our 370k data (276k text-only), the batch size is 1,052 (1.37×768).

Table 1: **MIEB-Lite (51 tasks) results broken down by task categories.** We provide averages of both English and multilingual tasks. Models are ranked by the Mean (m) column. Shortcuts are x=“Crosslingual”, m=“Multilingual”, en=“English”, and task categories from MIEB [76]. We refer to the latest MIEB leaderboard to obtain scores for the compared baselines.

MIEB-Lite (51 Tasks)													
Model Name (↓)	Dataset Size	Rtrv. (11)	Clus. (2)	ZS. (7)	LP. (8)	Cmp. (6)	VC. (5)	Doc. (6)	vSTS (en) (2)	Rtrv. (m) (2 (44))	vSTS (x&m) (2 (19))	Mean (en) (47)	Mean (m) (51)
<i>Encoder baselines</i>													
CLIP-ViT-bigG [34]	2B	34.2	80.8	72.4	77.8	35.0	43.0	35.5	73.4	26.2	34.5	56.5	51.3
SigLIP-so400m [83]	9B	32.4	75.9	73.8	78.8	32.8	48.0	46.9	69.6	35.4	41.4	57.3	53.5
<i>MLLM baselines</i>													
VLM2Vec-LoRA [31]	662k	21.0	66.4	32.1	64.8	29.4	65.3	42.7	70.9	24.8	42.2	49.1	46.0
E5-V [30]	276k	26.9	51.7	36.2	70.6	39.4	52.6	56.0	81.2	58.3	46.3	51.8	51.9
Voyage Multimodal 3 [65]	-	33.2	76.6	48.6	69.3	35.8	50.0	63.5	84.2	49.0	70.4	57.7	58.1
mmE5 (11B) [8]	2.1M	34.2	77.0	59.8	71.1	27.8	59.2	53.9	78.8	66.6	54.6	57.7	61.8
GME (7B) [84]	8.0M	37.9	69.6	55.5	68.7	52.2	55.4	86.1	81.8	62.4	75.4	63.4	64.5
<i>Our Text-only Variants</i>													
LCO-EMB-VL (3B)	276k	32.6	61.8	45.0	67.4	38.5	57.7	62.2	86.1	52.4	76.5	56.4	58.0
LCO-EMB-VL (7B)	276k	31.8	52.7	49.1	68.5	40.4	63.1	66.0	88.4	59.8	84.3	57.5	60.4
<i>Our Multimodal Variants</i>													
LCO-EMB-VL (3B)	370k	34.0	71.6	58.1	68.3	46.1	57.8	73.0	83.8	54.6	76.1	61.6	62.3
LCO-EMB-VL (7B)	370k	36.4	76.0	66.8	72.5	51.0	65.2	75.6	86.6	63.1	83.3	66.2	67.6
LCO-EMB-Omni (3B)	370k	34.5	78.3	66.1	72.8	48.2	59.2	73.4	85.7	54.5	80.4	64.8	65.3
LCO-EMB-Omni (7B)	370k	36.4	80.0	68.5	74.1	50.1	70.5	75.4	86.2	64.3	82.4	67.6	68.8

instance includes a premise with three hypotheses (entailment, neutral, contradiction). We use $\sim 276k$ triplets from all-NLI with entailments as positives and contradictions as hard negatives. We further construct **Scale-1M**, a curated collection of 1M sentence pairs sampled from 20M multilingual parallel corpora, including Global Voice [48], MUSE [53], News Commentary [62], Tatoeba [1], Talks [52], WikiMatrix [55], and other Sentence Transformers sources [51]. This design leverages diverse descriptive text to simulate image captions, activating image representations without direct image supervision, and integrates multilingual pairs to enhance cross-lingual alignment, which may in turn enhance multimodal alignment across languages. (2) *Multimodal Setting*. Building on all-NLI, we further add $\sim 94k$ synthetic multimodal pairs (Appendix B) to enhance alignment in the downstream task format space, yielding a final dataset of $\sim 370k$ triplets.

Benchmarks. For *image-text embedding tasks*, we adopt **MIEB-Lite** (51 tasks), the official lightweight version of MIEB (130 tasks; [76]), which covers eight categories (Appendix C), including Linear Probing, Retrieval (English and Multilingual), Zero-shot Classification, Compositionality Evaluation, Vision-centric QA, Document Understanding, Clustering, and Visual STS (English and Cross-lingual). For rapid iteration and ablation, we employ a compact subset of 18 overlapping tasks, termed **MIEB-Sub18** (Appendix D). For *audio-text embedding tasks*, we evaluate on AudioCaps [32] and Clotho [16] datasets. For *video-text embedding tasks*, we utilize MSR-VTT [79] and ActivityNet [26] datasets. The performance on these tasks provides complementary evidence supporting the universality and effectiveness of LCO-EMB, extending beyond the vision and language modalities. For both audio-text and video-text embedding tasks, we utilize the *Recall@1* as the evaluation metric.

4.2 Performance Comparison on the MIEB Benchmark

To better understand the results, we brief the goal and metric of each MIEB-Lite category: (1) **Visual STS** reformulates semantic textual similarity as a vision task by rendering text as images to test models’ semantic understanding, evaluated by *Spearman correlation*; (2) **Document Understanding/Visual Document Retrieval** measures a model’s ability to capture layout-aware textual semantics in visual documents and image-text alignment, evaluated by *nDCG@5*; (3) **Image Linear Probing** assesses the discriminative and transferable quality of frozen visual representations using *accuracy*; (4) **Compositionality Evaluation** tests fine-grained image-text alignment with *accuracy*; (5) **Vision-centric QA** evaluates visual reasoning and understanding through *accuracy*; (6) **Retrieval** measures modality-specific and joint encoding performance with *nDCG@10*; (7) **Zero-shot Classification** evaluates similarity-based classification using *accuracy*; and (8) **Clustering** examines the structural coherence of embeddings using the *NMI* metric. Detailed task descriptions are in Appendix C.

We evaluate LCO-EMB on the 51 tasks of the MIEB-Lite benchmark. As reported in Table 1, LCO-EMB consistently outperforms strong baselines, including E5-V [30], VLM2Vec [31], Voyage-

Table 2: Performance and efficiency comparisons of different training strategies using 3B and 7B variants of Qwen2.5-VL backbones. GPU hours are benchmarked by hours $\times$ number of H20 GPUs.

Training Strategy	Training Time (GPU Hours)	Multiling. Img. Rtr	V-STs (Eng.)	V-STs (cross)	V-STs (multi)	Doc. Und.	Linear Probe	Average
Qwen2.5-VL-3B	n/a	31.73	73.82	59.03	68.57	28.82	46.96	51.49
w/ CLIP-style CL (multimodal)	$\sim$ 453.0 Hours	25.15	72.51	67.45	65.22	48.91	41.05	53.38
w/ Linear Proj. (text-only)	$\sim$ 4.5 Hours	31.31	75.25	62.95	69.32	28.12	49.19	52.69
w/ Full-Finetune (text-only)	$\sim$ 8.5 Hours	<u>44.61</u>	<u>81.65</u>	<u>68.67</u>	<u>77.75</u>	<u>49.71</u>	<u>50.21</u>	<u>62.10</u>
w/ LoRA (text-only)	$\sim$ 4.7 Hours	51.61	81.88	74.97	78.30	57.90	53.05	66.28
Qwen2.5-VL-7B	n/a	40.31	73.82	59.03	68.56	28.82	46.96	52.92
w/ CLIP-style CL (multimodal)	$\sim$ 550.0 Hours	18.24	73.92	68.70	65.41	44.89	38.93	50.02
w/ Linear Proj. (text-only)	$\sim$ 8.8 Hours	40.29	72.05	65.46	70.88	35.69	52.96	56.22
w/ Full-Finetune (text-only)	$\sim$ 17.3 Hours	<u>44.05</u>	<u>83.15</u>	<u>79.09</u>	<u>81.28</u>	<u>58.02</u>	<u>53.34</u>	<u>66.49</u>
w/ LoRA (text-only)	$\sim$ 9.3 Hours	56.64	85.05	85.30	83.48	67.49	53.91	71.98

Multimodal-3 [65], mmE5 [8], and GME [84]. Remarkably, despite using only $\sim$ 0.37M training pairs—about $21\times$ less data than GME ($\sim$ 8M)—our multimodal variants set a new state-of-the-art on MIEB. Consistent with findings from Xiao et al. [76], MLLM-based embedding models excel at tasks leveraging MLLM backbones’ reasoning and cross-modal understanding abilities, such as multilingual alignment, compositionality, and document understanding. Beyond these strengths, LCO-EMB also attains competitive results on clustering, linear probing, and zero-shot classification—areas where MLLM-based representations typically lag behind CLIP-style models. Notably, even our text-only variants, trained with minimal text-only contrastive data, surpass advanced proprietary model Voyage-Multimodal-3. Incorporating only $\sim$ 94k additional multimodal samples (image–text and interleaved data; Appendix B) further calibrates the representation space for downstream task formats, resulting in a compact yet highly effective dataset of $\sim$ 370k triplets.

4.3 Analysis of Representational Capability of LCO-EMB

To better analyze the representational capability of LCO-EMB, we use the text-only variants, *i.e.*, without utilizing the synthetic multimodal pairs, as evaluation targets and conduct extensive validation and ablation studies on **MIEB-Sub18** benchmark.

Main results. We assess text-only variants of LCO-EMB against the advanced embedding methods on the MIEB-Sub18. As illustrated in Figure 6 (Appendix D), LCO-EMB, trained on the 3B and 7B variants of Qwen2.5-VL-Instruct (VL) and Qwen2.5-Omni (Omni), consistently outperforms the leading embedding models across a variety of multimodal downstream tasks. On average across all evaluation categories, the text-only variants of LCO-EMB have outperformed E5-V [30] and Voyage-Multimodal-3 [65] by 21.69 and 13.00 points, where E5-V and Voyage-M3 are advanced open-source and proprietary MLLM embedding models, respectively. Notably, LCO-EMB delivers significant improvements on the Linear Probing, Cross-lingual Visual STS, and Multilingual Image Retrieval tasks, outperforming prior advanced methods by margins of 21.02, 10.26, and 15.35 points, respectively. The results highlight the effectiveness and generalizability of LCO-EMB. It is also noteworthy that while Voyage-M3 is a commercial model explicitly optimized on PDF–text pairs for document understanding tasks, LCO-EMB, trained solely on textual data, still achieve comparable results.

Comparison of different training strategies. We apply LoRA to enhance representational capacity while preserving latent cross-modal alignment. To assess this design, we experiment with Qwen2.5-VL 3B and 7B backbones, comparing LoRA against three baselines: (1) *standard CLIP-style contrastive fine-tuning* on 800K PixmoCaps image-caption pairs, (2) *full fine-tuning*, and (3) a *shallow projection* that adds a linear layer after the output. Reported in Table 2, the CLIP-style baseline underperforms text-only LoRA, requires $50\times$ more training time, and the shallow projection increases parameters but does not effectively leverage pretrained cross-modal structure, yielding only marginal gains over native embeddings. Full fine-tuning achieves reasonable results but remains notably inferior to LoRA. We attribute this gap to an objective mismatch: contrastive loss deviates from the model’s pretraining objective, and full fine-tuning consequently induces larger perturbations to the pretrained parameters, which are more likely to disrupt the established cross-modal alignment. Detailed analysis of LoRA hyperparameters⁴ is presented in Appendix E. We further analyze the

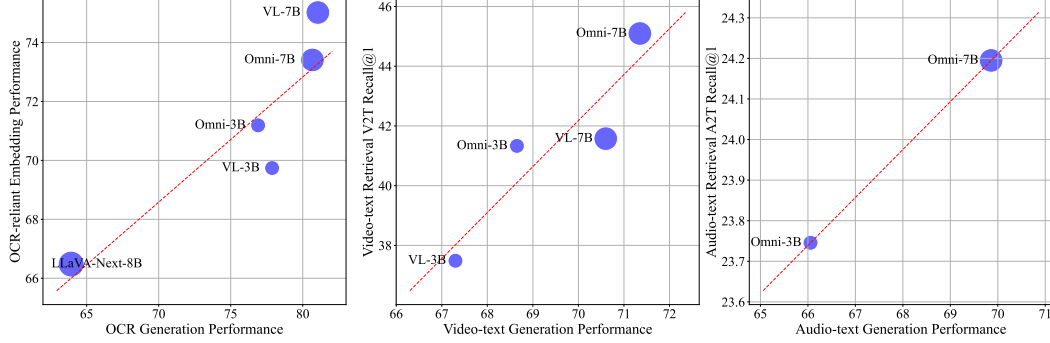


Figure 4: Scaling relationship between generation benchmark performance (X-axis) and representation benchmark performance after language-centric contrastive learning (Y-axis).

impact of different training datasets and model merging technique in Appendix F.

5 Generation-Representation Scaling Law

The superior performance of LCO-EMB is primarily attributed to the intrinsic multimodal alignment capabilities of the backbone MLLMs, which we activate through lightweight contrastive fine-tuning. This observation prompts a fundamental question: *What is the relationship between the inherent multimodal generative ability of MLLMs and their representation potential?* Through empirical analysis, we reveal a positive scaling correlation between these two aspects. Furthermore, we substantiate our empirical findings with a theoretical analysis with a PAC-Bayesian generalization bound, linking models’ generative capabilities with the upper bound of their representation performance.

5.1 The Relationship between Generative and Representational Capabilities

We conduct an empirical analysis to investigate the relationship between improving multimodal representation capabilities via text-only contrastive learning and the intrinsic generative capacity of MLLMs. Our analysis spans three different types of modality pairs, *i.e.*, OCR-based image-text tasks, video-text tasks, and audio-text tasks. The experimental setup is detailed below:

- **OCR-Based Image-Text Tasks:** We evaluate OCR-dependent capabilities through paired representation and generative tasks. For representation tasks, we average scores from Visual Semantic Textual Similarity (V-STs-English) and Document Understanding. Generative performance is measured by averaging results from TextVQA [56], DocVQA [45], OCRBench [42], and ChartQA [44].
- **Video-Text Tasks:** Representation capabilities are assessed using video-text Recall@1 scores averaged across MSR-VTT [78] and ActivityNet [6]. Generative performance combines results from Video-MME_{w/ sub} [20] and MVBench [38].
- **Audio-Text Tasks:** We compute Recall@1 scores using Clotho [16] and AudioCaps [32]. For generative evaluation, we average performance on MMAU [54] and VoiceBench [10], comprehensive benchmarks encompassing multiple sub-tasks.

Results. As depicted in Figure 4, we observe a consistently positive correlation between baseline generative performance before CL and the post-CL representational performance across different MLLM backbones on all task categories. This observation leads to the discovery of **Generation-Representation Scaling Law (GRSL)**, where *the representational abilities of MLLMs, enhanced through contrastive refinement, scale positively with the model’s original generative capability*. This insight suggests an alternative pathway for advancing multimodal models by harnessing the scaling effects of generative capacity. Next, we provide a theoretical analysis of GRSL, which formally links the MLLM’s generative quality to the upper bound on its final embedding performance.

⁴There is a slight difference in the evaluation resolution between Table 4 and Table 2 due to different codebase versions used for the experiments.

5.2 Theoretical Analysis of Generation-Representation Scaling Law

We aim to prove that a stronger generative prior of MLLMs leads to better representations after contrastive fine-tuning. We formalize this intuition using the PAC-Bayesian framework. We begin by defining our central hypothesis, deferring formal definitions of population/empirical risk ($\mathcal{L}_c^{\text{pop}}$, $\hat{\mathcal{L}}_c^{\text{emp}}$) and the generative quality of the prior ($I_P(X; Y)$) to Appendix G.

5.2.1 Central Hypothesis

The core of our argument is that a good generative prior provides a “warm start” for contrastive fine-tuning. We formalize this as follows.

Hypothesis 1 (Generative Warm Start). *Let P be a generative prior and Q the posterior after contrastive fine-tuning. The expected empirical loss under Q is bounded by:*

$$\mathbb{E}_{\theta \sim Q} [\hat{\mathcal{L}}_c^{\text{emp}}(\theta)] \leq \log N - I_P(X; Y) + \epsilon_P, \quad (1)$$

where $\epsilon_P \geq 0$ captures the gap between the information-theoretic optimum and the loss achieved after finite-step contrastive fine-tuning. A better prior (higher I_P) leads to a smaller ϵ_P .

A detailed justification for this hypothesis, linking $I_P(X; Y)$ to the pre-existing alignment of representations, is provided in Appendix G.2.

5.2.2 Main Theoretical Result

Theorem 1 (Generative-Contrastive PAC-Bayes Bound). *Let P be a generative prior and Q the posterior after contrastive fine-tuning on n samples. Under Hypothesis 1, with probability at least $1 - \delta$, the expected population contrastive risk is bounded by:*

$$\mathbb{E}_{\theta \sim Q} [\mathcal{L}_c^{\text{pop}}(\theta)] \leq \underbrace{\log N - I_P(X; Y)}_{\text{Generative Bottleneck}} + \underbrace{\epsilon_P}_{\text{Inefficiency Gap}} + \underbrace{\sqrt{\frac{\text{KL}(Q \| P) + \log(1/\delta)}{2n}}}_{\text{PAC-Bayes Complexity Penalty}}. \quad (2)$$

The proof is provided in Appendix H.

Corollary 1 (Generative Performance Governs Representation Bound). *Using the approximation $I_P(X; Y) \approx H(Y) - \mathcal{L}_g(P)$ (see Appendix G), the bound is directly governed by the prior’s generative loss:*

$$\mathbb{E}_{\theta \sim Q} [\mathcal{L}_c^{\text{pop}}(\theta)] \lesssim \mathcal{L}_g(P) + (\log N - H(Y)) + \epsilon_P + \sqrt{\frac{\text{KL}(Q \| P) + \log(1/\delta)}{2n}}. \quad (3)$$

This result formalizes our central claim: a **lower generative loss** $\mathcal{L}_g(P)$ directly tightens the upper bound on the final contrastive performance.

Interpretation of the Bound. The theorem and its corollary reveal three distinct factors that govern the final representation quality:

1. **The Generative Bottleneck** ($\log N - I_P(X; Y)$): The final quality is limited by the mutual information, $I_P(X; Y)$, captured by the generative prior. A stronger generative model (higher I_P) lowers this performance floor.
2. **The Optimization Inefficiency** (ϵ_P): A better prior creates a more favorable optimization landscape, resulting in a smaller “inefficiency gap” ϵ_P .
3. **The Fine-tuning Cost** ($\sqrt{\dots}$): The PAC-Bayes complexity penalty justifies using parameter-efficient methods (like LoRA) to keep $\text{KL}(Q \| P)$ small, thus retaining the benefits of the strong generative prior.

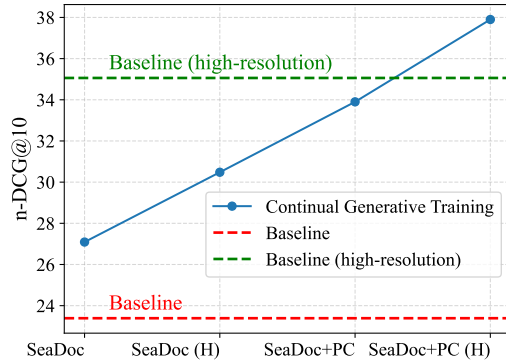


Figure 5: Retrieval performance of Qwen2.5-VL-3B fine-tuned on various continual generative fine-tuning strategies before CL on SeaDoc benchmark. PC denotes PixmoCaps. H denotes training with high resolution.

We further empirically investigate whether improving generative capabilities before contrastive learning raises the representational upper bound, which is the key aspect of GRSL. We construct a difficult Visual Document Retrieval (VDR) task: **SeaDoc**, a VDR task in low-resource SouthEast Asian languages. Figure 5 summarizes the results for Qwen2.5-VL-3B with four continual pre-training settings: SeaDoc-train, SeaDoc-train + PixmoCaps (PC) to preserve general image perception capabilities, and training both settings at high resolution (H). The trend confirms that representation performance after CL improves as generative capability strengthens. Full details of SeaDoc construction, experimental settings, and analyses are provided in Appendix J.

6 Related Work

Omnimodal Representation Learning. Existing approaches to omnimodal representation learning [22, 67] typically rely on large-scale cross-modal pairs to train modality-specific encoders. Recent progress [8, 39, 84] highlights the potential of MLLMs for image–text alignment. However, the effectiveness of exploiting the latent alignment inherent in MLLMs’ generative capabilities for omnimodal representation learning—and its underlying theoretical basis—remains unexplored.

Modality-centric Representation Learning. Prior work explores representation learning for a single modality. For instance, ImageBind [22] leverages the image modality as the anchor for contrastive learning to align with all other modalities. Web-SSL [18] explores language-free (thus “vision-centric”) visual representation learning, which scales data volume to be on par with CLIPs to train DINOv2. By scaling up data volume, the vision-centric self-supervised learning can achieve OCR performance on par with CLIP, which is typically thought to attain through textual supervision [63]. E5-V [30] leverages text-only learning to generalize to images and composed retrieval tasks. We extensively study the language-centric view to train omnimodal representation models.

Representation Capabilities. Through investigating 50 models across 130 tasks in 39 languages, Xiao et al. [76] report that CLIP’s performance gains from scaling data, batch size, and model size have largely plateaued on advanced representation benchmarks, including interleaved encodings [69, 85], compositionality [61], textual visual representations [19, 74], and image–multilingual text alignment [57]. They further highlight MLLM-based embedding models as a promising alternative, motivating our exploration of the relationship between representational and generative capabilities in MLLMs. Prior work has explored this connection: Cambrian-1 [63] combines a shared language decoder with various vision encoders for training MLLMs and demonstrates that the downstream performance of MLLMs scales with the representation capabilities of the vision encoders, while Yang et al. [80] formalizes the law between visual representation and MLLM generative capabilities. In contrast, we explore a fundamentally different concept: the “*Generation-Representation Scaling law*” between generation and representation capabilities of the MLLM itself. We see above as the “*Representation-Generation Scaling Law*” where the MLLM’s generation performance scales with the strength of modality-specific encoders. In this work, we explore a fundamentally different concept: the “*Generation-Representation Scaling law*” where the MLLM’s representation abilities scale with its own generation capabilities. Our findings align closely with Xiao et al. [75], who demonstrate that LLM-based embeddings excel at instruction following and reasoning-oriented retrieval.

7 Conclusion

In this work, we reveal that the superior performance of MLLM-based embedding approaches originates from implicit cross-modal alignment established during generative pretraining, wherein the language decoder learns to integrate multimodal information within a unified representation space. Leveraging this insight, we develop LCO-EMB, a language-centric omnimodal embedding framework that treats contrastive learning as a lightweight refinement stage, enhancing representation quality while preserving the model’s generative capabilities. Building on this formulation, we introduce the Generation-Representation Scaling Law (GRSL), which establishes a positive correlation between a model’s generative capacity and the representation upper bound. Our theoretical analysis, through a PAC-Bayesian generalization bound, together with extensive empirical validation on diverse and challenging benchmarks, confirms both the efficacy of LCO-EMB and the generality of GRSL. Collectively, these findings re-conceptualize the role of contrastive learning and position generative pretraining—not merely the expansion of cross-modal data—as the central driver of scalable, efficient, and robust multimodal representation learning. We discuss the limitations of this work in Appendix A.

References

- [1] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. URL <https://aclanthology.org/Q19-1038/>.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. URL <https://arxiv.org/abs/2502.13923>.
- [3] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026, 2014. URL [10.1109/CVPR.2014.259](https://doi.org/10.1109/CVPR.2014.259).
- [4] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=aIoEru2qCG>.
- [5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL <https://aclanthology.org/D15-1075/>.
- [6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nibbles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.pdf.
- [7] Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Deli Zhao, Anh Tuan Luu, and Yu Rong. Geopqa: Bridging the visual perception gap in mllms for geometric reasoning. *ArXiv*, abs/2509.17437, 2025. URL <https://arxiv.org/abs/2509.17437>.
- [8] Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mmE5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8254–8275, Vienna, Austria, July 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.findings-acl.433/>.
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/chen20j.html>.
- [10] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *ArXiv*, abs/2410.17196, 2024. URL <https://arxiv.org/abs/2410.17196>.
- [11] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*. Association for Computational Linguistics, 2025. URL <https://arxiv.org/abs/2411.06176>.
- [12] Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261, 2025. URL <https://arxiv.org/abs/2507.06261>.

- [13] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 91–104, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Deitke_Molmo_and_PixMo_Open_Weights_and_Open_Data_for_State-of-the-Art_CVPR_2025_paper.pdf.
- [14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. URL 10.1109/CVPR.2009.5206848.
- [15] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *ArXiv*, abs/2503.17352, 2025. URL <https://arxiv.org/abs/2503.17352>.
- [16] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. *ArXiv*, abs/1910.09387, 2019. URL <https://arxiv.org/abs/1910.09387>.
- [17] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-1006/>.
- [18] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. *ArXiv*, abs/2504.01017, 2025. URL <https://arxiv.org/abs/2504.01017>.
- [19] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Celine Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=ogjBpZ8uSi>.
- [20] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xianwu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.pdf.
- [21] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=SkEYojRqtm>.
- [22] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.pdf.

- [23] Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-long.3/>.
- [24] Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *ArXiv*, abs/2506.18902, 2025. URL <https://arxiv.org/abs/2506.18902>.
- [25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.pdf.
- [26] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nibbles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, June 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.pdf.
- [27] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [28] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *ArXiv*, abs/2503.06749, 2025. URL <https://arxiv.org/abs/2503.06749>.
- [29] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/huh24a.html>.
- [30] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *ArXiv*, abs/2407.12580, 2024. URL <https://arxiv.org/abs/2407.12580>.
- [31] Ziyang Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=TEOKOzWYAF>.
- [32] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL <https://aclanthology.org/N19-1011/>.
- [33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. URL 10.1109/ICCVW.2013.77.
- [34] laion. Clip-vit-bigG-14-laion2b-39b-b160k. <https://huggingface.co/laion/CLIP-ViT-bigG-14-laion2B-39B-b160k>, 2024.
- [35] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *ArXiv*, abs/2408.12637, 2024. URL <https://arxiv.org/abs/2408.12637>.

- [36] Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *ArXiv*, abs/2509.21268, 2025. URL <https://arxiv.org/abs/2509.21268>.
- [37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/li22n.html>.
- [38] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.pdf.
- [39] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=i45NQb2iK0>.
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.
- [41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [42] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. URL <https://arxiv.org/abs/2305.07895>.
- [43] Feipeng Ma, Hongwei Xue, Guangting Wang, Yizhou Zhou, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Multi-modal generative embedding model. *ArXiv*, abs/2405.19333, 2024. URL <https://arxiv.org/abs/2405.19333>.
- [44] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.findings-acl.177>.
- [45] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209, January 2021. URL https://openaccess.thecvf.com/content/WACV2021/papers/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.pdf.
- [46] Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=BC41IvfSzv>.
- [47] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB leaderboard. <https://huggingface.co/spaces/mteb/leaderboard>, 2025.

- [48] Khanh Nguyen and Hal Daumé III. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-5411/>.
- [49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=a68SUt6zFt>.
- [50] 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. 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, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- [51] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-1410/>.
- [52] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.365/>.
- [53] Vin Sachidananda, Ziyi Yang, and Chenguang Zhu. Filtered inner product projection for crosslingual embedding alignment. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=A2gNouoXE7>.
- [54] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=TeVAZXr3yv>.
- [55] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wiki-Matrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.eacl-main.115/>.
- [56] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.pdf.
- [57] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery. URL <https://doi.org/10.1145/3404835.3463257>.
- [58] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *ArXiv*, 2303.15389, 2023. URL <https://arxiv.org/abs/2303.15389>.

- [59] Kimi Team. Kimi-vl technical report. *ArXiv*, abs/2504.07491, 2025. URL <https://arxiv.org/abs/2504.07491>.
- [60] Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 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, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.45/>.
- [61] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.pdf.
- [62] Jörg Tiedemann. News commentary v16. <https://opus.nlpl.eu/News-Commentary/corpus/version/News-Commentary>, 2012. Accessed: May 16, 2024.
- [63] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf.
- [64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL <https://arxiv.org/abs/2302.13971>.
- [65] Voyage AI. voyage-multimodal-3: all-in-one embedding model for interleaved text, images, and screenshots. <https://blog.voyageai.com/2024/11/12/voyage-multimodal-3/>, Nov 2024.
- [66] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. VI-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025. URL <https://arxiv.org/abs/2504.08837>.
- [67] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *ArXiv*, abs/2305.11172, 2023. URL <https://arxiv.org/abs/2305.11172>.
- [68] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/wang20k.html>.
- [69] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/11927.pdf.
- [70] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL <https://aclanthology.org/N18-1101/>.

- [71] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/wortsman22a.html>.
- [72] Chenghao Xiao, Yizhi Li, G Hudson, Chenchua Lin, and Noura Al Moubayed. Length is a curse and a blessing for document-level semantics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1385–1396, Singapore, December 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.emnlp-main.86/>.
- [73] Chenghao Xiao, Yang Long, and Noura Al Moubayed. On isotropy, contextualization and learning dynamics of contrastive-based sentence representation learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12266–12283, Toronto, Canada, July 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.findings-acl.778/>.
- [74] Chenghao Xiao, Zhuoxu Huang, Danlu Chen, G Thomas Hudson, Yizhi Li, Haoran Duan, Chenchua Lin, Jie Fu, Jungong Han, and Noura Al Moubayed. Pixel sentence representation learning. *ArXiv*, abs/2402.08183, 2024. URL <https://arxiv.org/abs/2402.08183>.
- [75] Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. Rar-b: Reasoning as retrieval benchmark. *ArXiv*, abs/2404.06347, 2024. URL <https://arxiv.org/abs/2404.06347>.
- [76] Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. *ArXiv*, abs/2504.10471, 2025. URL <https://arxiv.org/abs/2504.10471>.
- [77] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *ArXiv*, abs/2503.20215, 2025. URL <https://arxiv.org/abs/2503.20215>.
- [78] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/Xu_MSR-VTT_A_Large_CVPR_2016_paper.pdf.
- [79] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/Xu_MSR-VTT_A_Large_CVPR_2016_paper.pdf.
- [80] Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu. Law of vision representation in mllms. *ArXiv*, abs/2408.16357, 2024. URL <https://arxiv.org/abs/2408.16357>.
- [81] Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. VI-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *ArXiv*, abs/2507.22607, 2025. URL <https://arxiv.org/abs/2507.22607>.
- [82] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=KRLUvxh8uaX>.

- [83] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.pdf.
- [84] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *ArXiv*, abs/2412.16855, 2024. URL <http://arxiv.org/abs/2412.16855>.
- [85] Xin Zhang, Ziqi Dai, Yongqi Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, Jun Yu, Wenjie Li, and Min Zhang. Towards text-image interleaved retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4254–4269, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.214. URL <https://aclanthology.org/2025.acl-long.214/>.
- [86] 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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *ArXiv*, abs/2506.05176, 2025. URL <https://arxiv.org/abs/2506.05176>.
- [87] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.acl-demos.38/>.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Yes, the abstract and introduction clearly state claims and conclusion made in our paper. Specifically, we propose a language-centric omnimodal representation learning framework, discover the Generation-Representation Scaling Law (GRSL), and provide a theoretical justification for the GRSL.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations of the work in the Limitation section of the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provide the full set of assumptions and a complete proof in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We fully disclose all the information needed to reproduce the main experimental results in the experiment settings sections.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide source code in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We disclose all the detail that is necessary to appreciate the results in the experiment settings sections.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Results of embedding evaluation are not affected by random seed. And due to the large scale of trained models and compute requirements for MLLM training and evaluation, we were not able to perform each of our 20+ training configurations multiple times and evaluate their checkpoints’ performance across random seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the experiments compute resources in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The paper conform the NeurIPS Code of Ethics in every respect. We preserve anonymity and the datasets that we used are open-sourced datasets that do not have privacy issues.

Guidelines: The paper conform the NeurIPS Code of Ethics in every respect. We preserve anonymity and the datasets that we used are open-sourced datasets that do not have privacy issues.

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the potential positive societal impacts of our work in the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This method proposed in this paper is an embedding model. It will not generate harmful contents.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly cited the original paper that produced the code package or dataset, as well as indicating the dataset version we used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We describe our dataset construction methods in Section 3.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[No\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[No\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [\[Yes\]](#)

Justification: Yes, the usage of LLMs in our method is described in Section 3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Limitations

In this work, we have studied the scaling law between generative capabilities of pretrained MLLMs, their latent multimodal alignment, and their representational capabilities after contrastive learning. We use MLLMs that have gone through generative pretraining and those that have attained different levels of generative capabilities, and let them go through lightweight contrastive learning. During contrastive learning, model weights are minimally adjusted, through low-rank adaptation, to project the original knowledge space into an embedding space suitable for similarity matching. However, we do note that one can also jointly train generative loss and contrastive loss [43, 46] to maintain a model’s knowledge (through continual generative training), and enhance its representational power (through continual contrastive learning). Due to the high computational cost of this approach, we leave it as a promising direction for future work in the context of omnimodal representation learning.

B Details of Additional Multimodal Data

On top of the main text-only setting, the all-NLI training corpus—which plays a crucial role in unlocking the model’s representational capacity—we further construct approximately 94k multimodal training samples to align the embedding space with the downstream multimodal task space. Specifically, we include: (1) **Visual Document**. Unlike most prior studies, we intentionally construct only about 23k triplets from Colpali [19] and Docmatix [35], rather than performing exhaustive data exposure. We found that large-scale visual document data, when not balanced with text and other task datasets, can degrade overall task generalization. (2) **Retrieval and Compositionality**. We include only 3k triplets from MS-COCO, aiming to introduce basic image–text alignment. To enhance robustness to varying input lengths, we apply augmentation techniques from LA(SER)³ [72]. Interestingly, this not only improves length robustness but also enhances the model’s spatial perception and image–text compositional reasoning. (3) **Multilingual/Diverse Text Data**. To enhance linguistic and contextual diversity, we sample several thousand examples from our Scale-1M dataset introduced in the main paper. (4) **General Synthetic Data**. We further construct around 60k synthetic samples in diverse formats to maintain and reinforce the model’s instruction-following and interleaved alignment capabilities—which is important for tasks like VQA under the Reasoning-as-Retrieval paradigm. The diverse synthetic data also benefits classification tasks, improving both probing and zero-shot performance.

C Details of MIEB-Lite Benchmark

The MIEB-Lite benchmark comprises 51 tasks in 8 categories, where the details of each category are summarized as follows:

- **Visual STS** [74]: It conceptualizes traditional semantic textual similarity (STS) as a vision task by rendering text as images and evaluating the semantic understanding of visual encoders. Similarity scores are computed from the embeddings of image-text pairs and compared against human annotations using Spearman correlation. This task comprises three subcategories: *English* (STS 13 and STS 15), *cross-lingual* (STS-17, with image pairs in different languages, *e.g.*, Arabic–English), and *multilingual* (STS-b, with pairs in the same language, *e.g.*, Italian–Italian). Visual STS naturally assesses a model’s interleaved encoding ability to capture semantic meaning from text in image form, with **Spearman correlation** as the primary evaluation metric.
- **Document Understanding/Visual Document Retrieval**: MIEB-lite selects 6 tasks from the Vidore benchmark [19], which is to retrieve visual documents that contain information to solve the problem in the query. This task assesses a model’s ability to understand the complex layouts and textual information in visual documents, and the interleaved image-text alignment. Here we use **nDCG@5** as the evaluation metric.
- **Image Linear Probing**: MIEB-lite selects 8 challenging linear-probing datasets, including Country211, DTD, EuroSAT, GTSRB, OxfordPets, PatchCamelyon, RESISC45, and SUN397, which MLLMs typically struggle compared to CLIP-style models, as indicated by the MIEB benchmark leaderboard. We follow Xiao et al. [76] to adopt 16-shot linear probing, which closely preserves ranking compared to full-dataset probing, and report **accuracy** as the metric.
- **Compositionality Evaluation**: It evaluates fine-grained alignment of image-text features, requiring retrieving the groundtruth texts corresponding to the correct composition of all elements, *e.g.*, an

accurate fine-grained caption of an image, and vice versa for images given texts. This category includes ARO-Benchmark [82] and Winoground [61]. Here we use **accuracy** as the evaluation metric.

- **Vision-centric QA:** Given an interleaved input composed of a question conditioned on an image, the task requires models to retrieve the correct answer under the reasoning-as-retrieval paradigm [75]. This task category is mostly made of tasks assessing vision-centric capabilities [63], such as spatial relation perception, depth estimation, and relative distance. Here we use **accuracy** as the evaluation metric.
- **Retrieval:** MIEB-Lite adopts 11 retrieval tasks, consisting of image-only retrieval, image-text retrieval, and interleaved retrieval, providing a comprehensive assessment of models’ modality-specific and composed encoding capabilities. In addition, it also selects WIT datasets [57] and XM3600 [60], totally covering image retrieval tasks across 38 different languages, *i.e.*, multilingual image retrieval, to assess a model’s alignment capability between image and multilingual text embeddings, using **nDCG@10** as the primary metric.
- **Zero-shot Classification:** Zero-shot Classification assesses classification in a similarity-matching fashion. We use text prompts like “an image of a {label}” following common practices Radford et al. [50] and Xiao et al. [76]. MIEB-lite selects 7 challenging fine-grained zero-shot classification tasks, including CIFAR100, Country211, FER2013, FGVCAircraft, Food101, OxfordPets, and StanfordCars. Here we use **accuracy** as the evaluation metric.
- **Clustering:** Clustering provides an extra lens to inspect the clustered structure of embeddings. MIEB-lite adopts two clustering tasks, including fine-grained tasks such as Imagenet-Dog15 [14], which MLLM-based embedding models typically fail compared to CLIP-style models [76]. The **Normalized Mutual Information (NMI)** is utilized as the main evaluation metric.

D Details of MIEB-Sub18 Benchmark

We further select a smaller-scale subset than MIEB-lite, including 18 tasks from MIEB Xiao et al. [76] as **MIEB-Sub18**, which comprises 47 subtasks that are considered most challenging to the image-text embedding models, particularly in evaluating the capabilities of visual text representation, multilingual understanding, and interleaved encodings. Specifically, we focus on Visual STS [74], multilingual image retrieval [57], and document understanding from Vidore [19]. Additionally, we assess three image linear probing tasks where MLLM embeddings underperform relative to CLIP and self-supervised vision models, as reported on the MIEB leaderboard [47]. All evaluations are conducted using the official MIEB codebase [76].

- **Visual STS [74]:** It conceptualizes traditional semantic textual similarity (STS) as a vision task by rendering text as images and evaluating the semantic understanding of visual encoders. Similarity scores are computed from the embeddings of image-text pairs and compared against human annotations using Spearman correlation. This task comprises three subcategories: *English* (STS-12~16), *cross-lingual* (STS-17, with image pairs in different languages, *e.g.*, Arabic–English), and *multilingual* (STS-b, with pairs in the same language, *e.g.*, Italian–Italian). Visual STS naturally assesses a model’s interleaved encoding ability to capture semantic meaning from text in image form, with **Spearman correlation** as the primary evaluation metric.
- **Multilingual Image Retrieval:** We utilize the WIT datasets [57] and select its image retrieval subtasks across 11 different languages. This task accesses a model’s alignment capability between image and multilingual text embeddings with **nDCG@10** as the main metric.
- **Document Understanding:** We select 7 tasks from the Vidore benchmark [19], which is to retrieve visual documents that contain information to solve the problem in the query. This task assesses a model’s ability to handle the complex layouts in visual documents and the interleaved image-text alignment. Here we use **nDCG@5** as the evaluation metric.
- **Image Linear Probing:** We evaluate three linear probing tasks—Stanford Cars [33], BirdSnap [3], and Country211 [50]—which MLLMs struggle the most, as indicated by the MIEB benchmark leaderboard. We follow Xiao et al. [76] to adopt 16-shot linear probing, which closely preserves ranking compared to full-dataset probing, and report **accuracy** as the metric.

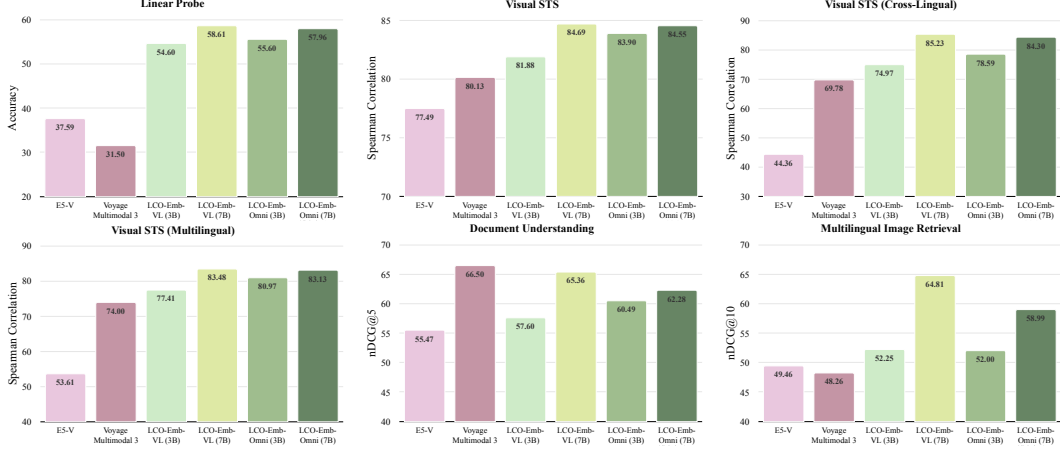


Figure 6: Comparison between the **text-only** variants of LCO-EMB with advanced open-source (E5-V [30]) and proprietary (Voyage Multimodal 3 [65]) embedding models on **MIEB-Sub18**.

Table 3: Comparison of different LoRA ranks and alpha values using Qwen2.5-VL-7B as the backbone. $*r = 256$, $\alpha = 512$ setting experiences unrecoverable loss spikes in training.

Rank (r)	Alpha (α)	Comp.	VC-QA	Multiling. Img. Rtr	V-STs (eng.)	V-STs (cross)	V-STs (multi)	Doc. Und.	Linear Probe	Average
8	16	48.35	58.08	56.64	85.05	85.30	83.48	67.49	53.91	67.29
64	16	55.64	60.62	55.62	84.60	85.16	83.40	65.76	52.44	67.90
64	128	43.40	51.86	58.93	84.98	84.44	83.39	67.66	57.24	66.49
256	16	52.29	57.30	57.49	84.88	85.68	83.61	66.95	53.36	67.70
256	128	43.07	57.56	56.32	85.89	84.82	83.98	67.33	55.51	66.81
256	512	85.52*	39.24*	0.70*	5.90*	12.90*	7.80*	0.90*	1.50*	19.31*

E Impact of LoRA Hyperparameters for LCO-EMB

Our approach, LCO-EMB, employs LoRA fine-tuning for lightweight contrastive learning, which aims to refine MLLM representations while minimally perturbing the model’s intrinsic knowledge and abilities, thereby effectively preserving its inherent cross-modal alignment capability. We encapsulate this benefit as the “learn less, forget less” characteristic of LoRA. Here, we further analyze the impact of two critical LoRA hyperparameters—rank (r) and α —on the performance of LCO-EMB.

In LoRA, rank (r) and alpha (α) jointly control the capacity for new knowledge integration and the extent to which it modulates existing knowledge. The rank defines the dimensionality of the trainable weight matrices used to approximate the original model’s weight updates; a higher rank thus increases the capacity for injecting new knowledge. Conversely, alpha scales the contribution of these matrices to the overall model weights, meaning a larger alpha amplifies the extent to which this new knowledge is infused into the model.

Table 3 presents the results of LCO-EMB under different values of rank and alpha. We observe distinct patterns for different task categories, and there doesn’t exist a global optimal setting of LoRA hyperparameters. For instance, LoRA hyperparameters bring minimal variations to tasks optimized by the training (*e.g.*, V-STs, whose textual counterpart STS is deemed directly optimized by All-NLI in text embedding literature, is invariant to LoRA hyperparameters). The optimal performance for multilingual retrieval, document understanding, and image linear probe generally occurs when alpha α is scaled up appropriately to rank r , such as $r = 8, \alpha = 16$ and $r = 64, \alpha = 128$. However, we notice that for tasks whose capabilities assessed largely differ from those which the training set optimizes, *e.g.*, compositionality and vision-centric QA, a larger alpha α generally brings significant performance degradation, showing the importance of the preservation of the base model’s knowledge for generalization to OOD tasks. We observe that with rank 256 and alpha 512, models experience unrecoverable loss spikes in training.

Table 4: Exploring the impact of training dataset utilization and model ensemble on LCO-EMB, where LCO-EMB-Ens denotes the ensemble model produced by applying the model soup [71] technique to LCO-EMB variants fine-tuned on all-NLI and Scale-1M.

Model	Data Source	Linear Prob.	v-STs (Eng.)	v-STs (cross)	v-STs (multi)	Doc. Und.	Multi. Img. Rtr.	Avg.
LCO-EMB	all-NLI	51.86	84.69	85.23	83.48	65.36	56.37	71.17
LCO-EMB	Scale-1M	58.61	81.27	81.42	78.42	62.12	64.81	71.11
LCO-EMB-Ens	-	<u>55.69</u>	<u>83.79</u>	<u>84.88</u>	<u>82.82</u>	<u>63.15</u>	<u>62.67</u>	72.17

We acknowledge that an optimal rank and alpha likely exist for models of each size, striking a balance between introducing new knowledge and the extent to which it modifies pretrained model weights. We leave a more comprehensive empirical analysis and theoretical study to quantify this relationship for future work.

F Impact of Different Training Datasets and Model Merging for LCO-EMB

Recognizing that language models fine-tuned on different datasets often demonstrate distinct strengths, we independently fine-tune LCO-EMB, using Qwen-2.5-VL-Instruct as the backbone model, on all-NLI and Scale-1M via contrastive learning, then assess the performance of each variant in isolation. Subsequently, we investigate the effect of model ensembling by applying the model soup [71] technique, which merges the parameters of multiple separately fine-tuned models by averaging their weights. The results, presented in Table 4, provide the following three key insights:

- **Performance of all-NLI fine-tuned variant.** LCO-EMB trained by all-NLI excels in Visual STS and Document Understanding, indicating that NLI supervision sharpens not only textual similarity perception but also generalizes to improve their ability to preserve vision-text semantic similarity.
- **Performance of Scale-1M fine-tuned variant.** LCO-EMB adapted by Scale-1M leads on Linear Probing and Multilingual Image Retrieval tasks. Since Scale-1M supplies semantically rich descriptions of real-world scenes, LCO-EMB fine-tuned on this corpus appears to emulate image-caption pre-training, thereby activating image representations without explicit visual data.
- **Performance of model ensemble.** The LCO-EMB-Ens, through merging the LCO-EMB variants trained by all-NLI and Scale-1M, achieves the best overall performance, demonstrating that the model ensemble strategy effectively integrates the complementary strengths of each checkpoint.

G Preliminaries of Theoretical Analysis

G.1 Definitions

Definition 1 (Population and Empirical Risk). Let $\mathcal{D}$ be the data distribution. The **population contrastive risk** for a model θ is its true expected InfoNCE loss:

$$\mathcal{L}_c^{\text{pop}}(\theta) := \mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}_{\text{InfoNCE}}(X, Y; \theta)]. \quad (4)$$

Given a training set $\mathcal{S} = \{(X_i, Y_i)\}_{i=1}^n$ of size n , the **empirical contrastive risk** is defined as:

$$\hat{\mathcal{L}}_c^{\text{emp}}(\theta) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\text{InfoNCE}}(X_i, Y_i; \theta). \quad (5)$$

Definition 2 (Generative Quality of the Prior). Let P be the prior distribution over the parameters of a pre-trained autoregressive generative model. In the common case where $P = \delta_{\theta_0}$ is a point mass of the model parameters at a pretrained checkpoint θ_0 , we define its **generative quality** via the mutual information captured by θ_0 :

$$I_P(X; Y) := I_{\theta_0}(X; Y). \quad (6)$$

Under the standard approximation that the generative cross-entropy loss $\mathcal{L}_g(P)$ estimates the conditional entropy (ref. Appendix I), we have the following approximation:

$$\mathcal{L}_g(P) \approx H(Y) - I_P(X; Y) \implies I_P(X; Y) \approx H(Y) - \mathcal{L}_g(P). \quad (7)$$

Here, $H(Y)$ is the Shannon entropy of the target data distribution, which quantifies the inherent diversity and complexity of the target modality (e.g., text). Thus, for a fixed dataset, a lower generative loss $\mathcal{L}_g(P)$ corresponds to a higher mutual information and therefore a higher generative quality.

G.2 Justification for Hypothesis 1

A high $I_P(X; Y)$ implies that the representation of the generative model before contrastive finetuning, $f_P(X)$, is already predictive of Y . This pre-existing alignment means positive pairs are closer in representation space, enabling contrastive optimization to reach a lower empirical loss. The term $\log N - I_P(X; Y)$ is the theoretical lower bound on InfoNCE loss under ideal conditions. This hypothesis aligns with our empirical findings in Section 5.1, which show that stronger pretrained generative models yield better representations for downstream tasks, and is consistent with recent literature on decoder-based embedding models (e.g., [86]).

H Proof of Theorem 1

Proof. We begin with the standard PAC-Bayesian generalization bound, which holds with probability at least $1 - \delta$ for any posterior Q :

$$\mathbb{E}_{\theta \sim Q} [\mathcal{L}_c^{\text{pop}}(\theta)] \leq \mathbb{E}_{\theta \sim Q} [\hat{\mathcal{L}}_c^{\text{emp}}(\theta)] + \sqrt{\frac{\text{KL}(Q \| P) + \log(1/\delta)}{2n}}. \quad (8)$$

According to Hypothesis 1, the empirical risk is bounded as:

$$\mathbb{E}_{\theta \sim Q} [\hat{\mathcal{L}}_c^{\text{emp}}(\theta)] \leq \log N - I_P(X; Y) + \epsilon_P. \quad (9)$$

Substituting (9) into (8) yields Theorem 1. $\square$

I Relationship Between Generative Loss and Conditional Entropy

Definition 3 (Generative Quality of the Prior). Let P be the prior distribution over the parameters of a pre-trained autoregressive generative model, centered at θ_0 . We define its **generative quality** via the mutual information $I_P(X; Y) := I_{\theta_0}(X; Y)$ that its representations capture between the input X and the target output Y .

The mutual information is defined as $I_P(X; Y) = H(Y) - H(Y|X)$. While the true conditional entropy $H(Y|X)$ is unknown, it can be estimated by the model’s generative cross-entropy loss, $\mathcal{L}_g(P)$. The formal relationship is:

$$\mathcal{L}_g(P) = H(Y|X) + D_{\text{KL}}(p_{\text{data}}(Y|X) \| p_{\theta_0}(Y|X)), \quad (10)$$

where p_{θ_0} is the model’s predictive distribution. For a well-trained MLLM, the goal of minimizing generative loss is to minimize this KL divergence. Thus, for a strong prior, we can use the approximation $H(Y|X) \approx \mathcal{L}_g(P)$. Substituting this into the definition of mutual information yields:

$$I_P(X; Y) \approx H(Y) - \mathcal{L}_g(P). \quad (11)$$

Here, $H(Y)$ is the entropy of the target data, which is constant for a given dataset. Therefore, a **lower generative loss** $\mathcal{L}_g(P)$ directly corresponds to **higher mutual information** and thus a higher generative quality of the prior.

J Improving Representation Bounds via Enhancing Generative Capability

To further investigate the hypothesis that enhancing an MLLM’s generative ability improves its representations, a key aspect of the Generation-Representation Scaling Law, we introduce a challenging cross-lingual multimodal document retrieval task, **SeaDoc**. This task enables a comprehensive evaluation of MLLM’s representational capacity. In this task, an English query is used to retrieve a corresponding multimodal document page in a low-resource target language.

J.1 Data Curation

SeaDoc is a cross-lingual visual document retrieval benchmark specifically designed for low-resource SouthEast Asian (SEA) languages. While building upon foundational concepts from existing visual document understanding benchmarks like ViDoRe [19], SeaDoc uniquely challenges MLLMs’ visual document understanding capabilities on non-English languages at an unprecedented scale.

Specifically, we utilize Gemini-2.5-Flash [12] to annotate each PDF page by sequentially applying OCR, translating the content into English, and generating an English query answerable exclusively from that specific page. This results in 5,055 annotated {OCR, English translation, English query} triplets. To construct a high-quality query pool for the retrieval dataset in SeaDoc, we implement a three-stage quality control process:

1. Qwen2.5-7B-Instruct is first used to filter out functional pages (*e.g.*, title pages, author pages, tables of contents), which reduces the dataset to 4,491 content page annotations.
2. The same model then scores these annotations for *Quality* and *Groundedness* on a 10-point scale. Only questions with a quality score of at least **9** and a groundedness score of **10** are retained. Note that *Quality* measures the informativeness of the content and relevance of the query, and *Groundedness* measures the exclusivity of the answer to the page.
3. Our in-house linguists conduct a final review of the remaining triplets to ensure their quality.

As a result, we derive 1,001 high-quality queries to be used for retrieval tasks within the 5,055 page corpus.

For conducting additional OCR-intensive generative training, we construct a training set leveraging images that do not correspond to retrieval test set queries, resulting in 4k seed images. We construct 5 SFT tasks per image: 1) OCR the image. 2) OCR the image, then generate a question from the image. 3) Provide the English translation given the OCR’d text. 4) Provide the English translation directly from the image. 5) Provide the answer to the generated query. Note that compared to the SeaDoc test set, the training set is separately generated and includes an additional “provide answer to the generated question” part in the seed prompt. This process leads us to an around 20k training set to enhance targeted generative capability on low-resource visual documents, which we also explore combining with the PixmoCap dataset (710k) for general capability preservation in the main experiments.

J.2 Experimental Settings

We use Qwen2.5-VL-3B as the backbone and establish a baseline with lightweight contrastive learning. To assess whether enhanced generative ability benefits embedding quality, we further train a variant with additional generative pre-training before lightweight contrastive learning.

We apply supervised fine-tuning to enhance the model’s image-to-text generative capability. This stage utilizes a mixture of image-to-text training data, comprising *OCR data in SEA languages* (derived from the training split of SeaDoc) and *general-domain image captioning data*, *i.e.*, PixmoCaps [13]. The OCR data strengthens its capability in generating SEA languages from visual documents, while the inclusion of general image captioning data helps preserve its semantic alignment between image and text modalities in the general domain.

Given that text in multimodal documents can be small, requiring higher image resolution for MLLMs to accurately read textual content, we further employ two settings—high- and low-resolution—to assess the impact of image resolution. For the low-resolution setting, we follow standard practice by using a maximum of 262,144 pixels [87].

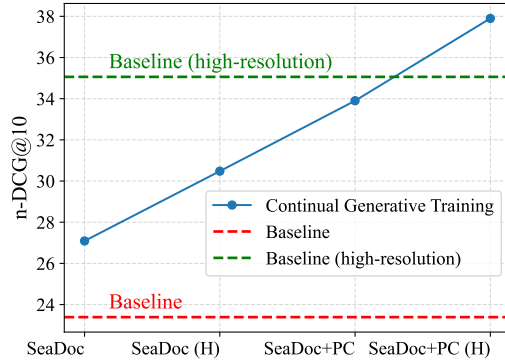


Figure 7: Retrieval performance of Qwen2.5-VL-3B fine-tuned on various continual generative fine-tuning strategies before CL on SeaDoc benchmark, where “PC” represents PixmoCaps and “H” denotes high-resolution. The results suggest that enhancing the generative ability of MLLMs before CL can enhance their embedding capability.

For the high-resolution setting, we use a $10\times$ larger maximum of 2,621,440 pixels. Here, we adopt **nDCG@10** as the primary metric.

J.3 Experimental Results

Figure 7 summarizes the retrieval performance of the same backbone model fine-tuned using different continual SFT strategies before the same CL tuning process on our SeaDoc benchmark.⁵ We draw the following key observations:

- (1) When SFT training is conducted exclusively on OCR-intensive data, *i.e.*, SeaDoc-train, at lower resolution, the model experiences a significant capability collapse compared to the baseline (Qwen2.5-VL-3B after lightweight CL). This SFT-induced degradation aligns with observations in recent multimodal reasoning research [7, 15, 28, 36, 66, 81]. Since foundation models have already undergone extensive SFT and RL, continual SFT can lead to overfitting and degrade models’ generalization capability.
- (2) Training on SeaDoc with higher resolution partly mitigates this collapse. This is because the text in visual documents is typically small; training with higher resolution allows for better grounding of the generated output in the visual text of the source image, as opposed to overfitting to example-level visual cues.
- (3) Incorporating PixmoCaps captions into the training set further boosts visual document retrieval performance post-CL. This is because general-domain image captioning data helps preserve the latent image-text alignment learned by MLLMs during pre-training. This preserved alignment is crucial for its effective exploitation by the subsequent text-only contrastive finetuning process.

⁵Unless otherwise specified, we evaluate model performance at the maximum resolution used during training.