Meta CLIP 2: A Worldwide Scaling Recipe

Yung-Sung Chuang[1,2][†] **Yang Li**[1], **Dong Wang**[1], **Ching-Feng Yeh**[1], **Kehan Lyu**[1],

Ramya Raghavendra[1], **James Glass**[2], **Lifei Huang**[1], **Jason Weston**[1], **Luke Zettlemoyer**[1],

Xinlei Chen[1][‡] **Zhuang Liu**[3], **Saining Xie**[4], **Wen-tau Yih**[1], **Shang-Wen Li**[1][§] **Hu Xu**[1][§]

[1]FAIR, Meta [2]MIT [3]Princeton University [4]New York University

Abstract

Contrastive Language-Image Pretraining (CLIP) is a popular foundation model, supporting from zero-shot classification, retrieval to encoders for multimodal large language models (MLLMs). Although CLIP is successfully trained on billion-scale image-text pairs from the English world, scaling CLIP's training further to learning from the worldwide web data is still challenging: (1) no curation method is available to handle data points from non-English world; (2) the English performance from existing multilingual CLIP is worse than its English-only counterpart, i.e., "curse of multilinguality" that is common in LLMs. Here, we present Meta CLIP 2, the first recipe training CLIP from scratch on worldwide web-scale image-text pairs. To generalize our findings, we conduct rigorous ablations with minimal changes that are necessary to address the above challenges and present a recipe enabling mutual benefits from English and non-English world data. In zero-shot ImageNet classification, Meta CLIP 2 ViT-H/14 surpasses its English-only counterpart by 0.8% and mSigLIP by 0.7%, and surprisingly sets new state-of-the-art without system-level confounding factors (e.g., translation, bespoke architecture changes) on multilingual benchmarks, such as CVQA with 57.4%, Babel-ImageNet with 50.2% and XM3600 with 64.3% on image-to-text retrieval. Code and model are available at https://github.com/facebookresearch/MetaCLIP.

1 Introduction

Contrastive Language-Image Pre-training (CLIP) [1] has become an essential building block of modern vision and multimodal models, from zero-shot image classification and retrieval to serving as vision encoders in multimodal large language models (MLLMs) [2, 3, 4, 5]. CLIP and its majority variants [6, 7] adopt an English-only setting, and Meta CLIP [7] introduces a scalable data curation algorithm to meticulously extract a billion-scale English dataset that exhausts long-tailed concepts in Common Crawl. The algorithm transforms the distribution of the raw Internet into *controllable* and *balanced* training distribution defined by metadata (e.g., visual concepts composed by human experts) and training distribution is known as one key contributor to performance. In contrast, popular CLIP reproductions outsource such key contributor to external resources, e.g., OpenCLIP [6] trained on LAION [8, 9] and DFN [10] rely on pretrained CLIP models for black-box *filtering* to keep only high-confidence data. Such approaches resemble distillation of an existing CLIP teacher model and produce *intractable* distributions owned by an outsourcing party.

[†] Work done during an internship at Meta.

[‡] Work done when working at Meta.

[§] Project leads.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

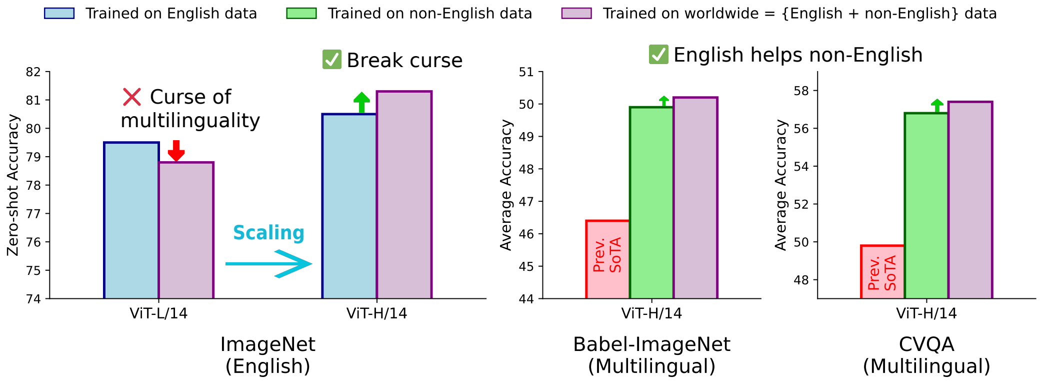

Figure 1: (Left) CLIP training suffers from the *curse of multilinguality* that the English performance of a CLIP model trained on worldwide (i.e., English + non-English), billion-scale data is worse than its English-only counterpart, even when applying our recipe on ViT-L/14; scaling to ViT-H/14 enables non-English data helps English-only CLIP. (Right) English data also helps non-English CLIP.

Although being the most widely used "foundation" models, most CLIP variants, including the scalable Meta CLIP, rely on English-only curation and thus discard the other, e.g., 50.9% [11] of non-English, worldwide web data. To extend CLIP training and data to the worldwide web for the next level of scaling, we inevitably have to handle these non-English image-text pairs—a barrier we refer to as the *worldwide scaling challenges*, which are issues not yet being solved after years of attempts to train CLIP on multilingual data:

Challenge #1: Lack of a fundamental data curation method to handle non-English data at scale. Existing attempts either conduct no curation on the raw, non-English image-text pair data at all (e.g., distilling from English CLIP [12] or machine translation [13, 14]), or rely on proprietary and private data sources (e.g., WebLI [15] that drives mSigLIP and SigLIP 2 [16, 17] is built from Google Image Search [18]).

Challenge #2: Worse English performance than English-only CLIP. This is also known as *curse of multilinguality* in text-only large language models (LLMs). For instance, mSigLIP is 1.5% worse than its English-only counterpart, SigLIP, on ImageNet [16], while SigLIP 2 [17] prioritizes English performance at the cost of even worse multilingual results than mSigLIP. Hence, disparate models have to be used to optimize English and non-English performance at the same time.

This work. We present **Meta CLIP 2**, the first ever recipe developing CLIP with training *from scratch* on *native* **worldwide** image-text pairs, without relying on outsourced resources, such as any private data, machine translation, or distillation. We empirically show that the curse of multilinguality in CLIP is the consequence of *insufficient scaling* due to the lack of a proper recipe for worldwide data curation and model training. When metadata, data curation, model capacity, and training are carefully designed and scaled *jointly*, we show that not only the performance trade-offs between English and non-English data disappear, but the two become *mutually beneficial*. Achieving such worldwide scaling is highly desirable, especially when English Internet data is exhausted soon [19].

Our Meta CLIP 2 recipe is built on top of English Meta CLIP, where overlapping with OpenAI CLIP's vanilla architecture is deliberately maximized. The overlap makes our findings generalizable to CLIP and its variants, compared to system works (cf. [16, 17, 20]) aiming at state-of-the-art (SoTA) performance with combination of all available techniques. Such combination involves confounding factors or comparison on outsourced resources instead of CLIP itself. The Meta CLIP 2 recipe introduces three principled innovations for scaling to worldwide. 1) *Metadata.* We scale the English Meta CLIP metadata to 300+ languages on Wikipedia and multilingual WordNet. 2) *Curation algorithm.* We build per-language substring matching and balancing to curate concept distribution for non-English data similar to the English counterpart. 3) *Training framework.* We design the first worldwide CLIP training framework, including an increase of seen image-text pairs during training proportional to the increased data size from the added non-English data examples, and a study on minimal viable model capacity to learn from worldwide scale data. As shown in Fig. 1, although a ViT-L/14 (the largest model size used by OpenAI) still suffers the curse of multilinguality, ViT-H/14 breaks the curse. English accuracy *rises* from 80.5% to 81.3% on ImageNet and surprisingly new SoTA is set with minimal CLIP architecture changes for multilingual image-to-text retrieval (XM3600 64.3%, Babel-ImageNet 50.2%, and CVQA 57.4%).

Together, Meta CLIP 2 enables the following desirable results by nature. 1) **Mutual benefits from the English and non-English worlds.** Non-English data now can better support an English-only model and vice versa, which is critical in the era when English data is depleting. 2) **Full multilingual support.** Meta CLIP 2 never drops image-text pairs simply by languages and yields models outperforming all the previous multilingual systems, such as mSigLIP [16] and SigLIP 2 [17]. 3) **Native-language supervision.** Models learn directly from alt-texts written by native speakers rather than synthetic machine translations [21, 14]. 4) **Cultural diversity.** Meta CLIP 2 retains the entire global distribution of images and thus inherits the comprehensive cultural and socioeconomic coverage advocated by [21]. Such coverage improves geo-localization and region-specific recognition. 5) **No-filter philosophy.** With the curation algorithm designed towards worldwide data, Meta CLIP 2 removes the last filter (i.e., whether the alt-text is in English) in pipeline, achieving better diversity and minimizing biases introduced by filters [21]. 6) **Broader impacts on foundation data.** This work provides a foundational data algorithm designed for worldwide scale, and benefits not only CLIP, but also efforts using CLIP data such as MLLM [2, 22], SSL (Web-DINO [23]) and image generation (DALL-E [24] and diffusion models [25]).

2 Related Work

2.1 Evolution of CLIP and its Data Processing

CLIP [1] and its variants [26, 6, 16] learn versatile image and text representations that are generally useful for downstream tasks [2, 27, 4]. Such multimodal contrastive learning and transformer architectures become standard components in vision and multimodal research. Data is a key contributor to CLIP's performance [28, 7]. Two major processing approaches for CLIP data emerge: curation[5] from scratch, and distillation from external resources. One key difference is that the former yields more *controllable distribution* and the latter has intractable distribution owned by an outsourcing party.

Curation from scratch. OpenAI CLIP [1] curates a training dataset of 400M image-text pairs from scratch and publicizes high-level curation guidance. Meta CLIP [7] makes OpenAI's guidance as a formal curation algorithm and scales the curation to 2.5B pairs. The algorithm is model-free, no blackbox filtering, and fully transparent to enable training entirely from scratch on public data source, where the data distribution is curated to align with metadata composed by human experts (e.g., WordNet and Wikipedia).

Distillation from external resources. Distillation-based methods usually have good performance and save compute by learning from teacher model's knowledge [30]. However, in the context of CLIP training the teacher is usually an external blackbox system, which introduces intractable bias. For example, LAION-400M/5B [8, 31] (used by OpenCLIP [6]) relies on OpenAI CLIP-filter and DFN [10] using a filter model trained on high-quality private data [32]. Recently, SigLIP [16] and SigLIP 2 [17] learn from data source WebLI [15], which is derived from Google Image Search [18].

2.2 Vision Encoding

CLIP-style models are widely used as vision encoders in MLLM, where language supervision in CLIP training helps to learn compact and semantic-rich visual representations. In contrast, traditional visual representation learning is based on self-supervised learning (SSL) methods like SimCLR [33], DINOv2 [34], and purely relies on the full visual signal without language bias. There are variants that take advantage of both. SLIP [35] combines language and SSL supervision; LiT [36] trains a vision encoder first and conducts language alignment later; Perception Encoder [20] shows early layers of CLIP representation yields vision-driven features with less semantic alignment. Recently, Web-DINO [23] shows SSL has better scalability on Meta CLIP curated large-scale data. In summary, CLIP focuses on human-aligned representations optimized for compact models and efficient downstream uses; SSL models aim to preserve all visual information as a general pretraining approach. We envision more synergy from the two research lines due to complementarity.

[5] Here, "curation" refers to select and align training data distribution with human from raw data source, excluding data filtering that is also referred to as curation in many works like DataComp [28, 29] and DFN [10].

2.3 Multilingual CLIP Models

Due to the lack of open source curation for public worldwide data, initial attempts to multilingual CLIP models are mainly distillation approaches. M-CLIP [13] and mCLIP [12] simply leverage existing English-only CLIP as the vision encoder and trains a multilingual text encoder with low-quality multilingual pairs. To incorporate non-English data, subsequent works [37, 14, 21] leverage machine translation techniques, either translating non-English captions into English or vice versa. These distillation-based models carry existing English CLIP bias or translation bias on nonhuman-captioned data. mSigLIP [16] substantially advanced multilingual performance by leveraging multilingual data from WebLI [15], which is an undisclosed dataset built with private data processing pipeline instead of publicly available worldwide data curation algorithm.

However, mSigLIP and other multilingual CLIP models suffer from the curse of multilinguality, e.g., mSigLIP is 1.5% worse in ImageNet accuracy than its English-only counterpart SigLIP. Recently, SigLIP 2 adopts a notably English-centric design of having 90% of its data in English, which is much higher than mSigLIP. Mixed results are also observed [38] on English benchmarks when scaling SigLIP from WebLI's 10B to 100B raw data, suggesting the challenges of scaling WebLI beyond.

3 The Meta CLIP 2 Recipe

Our recipe of scaling CLIP to native worldwide data and training comprises three steps shown in Fig. 2: (1) constructing worldwide metadata, (2) implementing worldwide curation algorithm, and (3) building training framework for worldwide model. For generalizable recipe and findings, Meta CLIP 2 is designed to maximize overlapping with OpenAI CLIP and Meta CLIP, and only adopts necessary changes to learn from worldwide data.

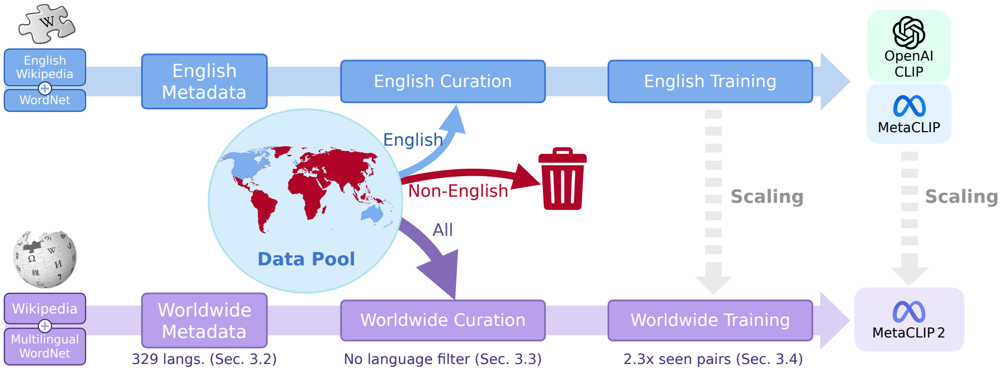

Figure 2: Overview of Meta CLIP 2 recipe: scaling CLIP data and training to worldwide scope.

3.1 Revisit of Meta CLIP Algorithm

We revisit the original Meta CLIP algorithm to illustrate how English-based CLIP data is curated with metadata constructed from human knowledge. The algorithm first constructs **metadata** $\mathcal{M}$, a list of high-quality visual concepts, from corpora written by human experts. $\mathcal{M}$ contains 500k entries, a combination and deduplication of entities from four high-quality sources: 1) all English WordNet Synsets, 2) Wikipedia English unigrams, and 3) bigrams, and 4) Wikipedia page titles. Then, the algorithm performs **substring matching** on each alt-text (from a given image-text pair in the data pool $\mathcal{D}$) using metadata $\mathcal{M}$ to obtain a list `matched_entry_ids`. **Global counting** is conducted to calculate the number of matches over $\mathcal{D}$ for each entry in $\mathcal{M}$ as `entry_count`. Finally, the algorithm applies **balancing** to transform the raw image-text pair distribution into a distribution that is balanced for head and tail concepts and ready for training, by associating each pair with a sampling probability. Specifically, the count per entry is first converted into a probability of sampling each entry, `entry_prob`, where tail entries (defined as `entry_count` $< t$) have a probability set to 1, and all the other head entries have $t/$`entry_count` as sampling probabilities. Each pair is then sampled based on probabilities of matched entries in its alt-text. Here, t is a threshold to decide head vs. tail entries and set to 20k in OpenAI CLIP; Meta CLIP raised t to 170k for scaling to billion English pairs.

3.2 Worldwide Metadata

We address the first challenge for worldwide scaling by constructing the missing metadata to cover the non-English world. We maintain independent metadata per language since such design is intuitive (e.g., the same word "mit" has different meaning in English and Germany), has better performance (see ablation in Sec. 4.2.2), and is flexible for adding and curating a new set of languages in future.

Our metadata is from the same four sources as OpenAI CLIP and Meta CLIP, but beyond English. Key changes are highlighted as follows. 1) Multilingual WordNet: we include all synsets from 31 languages. 2) Wikipedia Unigrams and 3) Bigrams: we process unigram and bigram from Wikipedia dumps dated on May 2024, which include corpora in 329 languages. We clean the corpora into plain text with WikiExtractor [39]. For most languages, we use space and punctuation to tokenize text into words, and then count unigrams and bigrams (with PMI scoring described in Appendix A.2). For languages without space separation (e.g., some Asian languages), we use open-source tokenizers (see Table 7 in Appendix A.1) developed by local communities to properly split text into words and meanwhile maintain the semantic integrity. 4) Wikipedia Titles: we use page titles from 40 random dates of Wikipedia snapshots and rank these titles by click-through traffic for each language.

3.3 Curation Algorithm

Next, we scale curation to worldwide data language-by-language. The curation algorithm is detailed below and summarized in pseudo-code as Algorithm 1. First, we conduct **language identification** (LID) [40] to classify the language of the alt-text from an image-text pair, and choose language-specific metadata to match concepts. The sets of languages covered by LID and metadata sources (e.g., Wikipedia) are usually different, so we first establish a mapping between one language in LID to a unique set of languages in metadata entries. The languages in the metadata mapped to the same language in LID are merged into one group. This ends with a dictionary representation of metadata, `M`, where the keys are each language in LID and the values are the combined metadata of each group of languages. We also include a key "other" for metadata of languages that cannot be associated with any language in LID. Each alt-text (`text`) in $\mathcal{D}$ is applied with LID for predicting its language (`text.lang`). After that, as in the Meta CLIP algorithm summarized in Sec. 3.1, we run **substring matching** with metadata corresponding to predicted languages: `matched_entry_ids = substr_match(text, M[text.lang])`, and aggregate **global count**, the number of matches of each entry, in `entry_counts`.

With counts calculated, we **balance** occurrence of concepts across pairs. In data curation for English CLIP described above, a threshold t is designed to limit the *matches per metadata entry*, where entries with matches fewer than t are defined as tail entries (or concepts) and otherwise head. Image-text pairs from head concepts are downsampled by a sampling probability derived from t to balance training data distribution. Thus, t depends on the size of raw data pool (e.g., a larger pool has higher counts for the same entry). OpenAI CLIP sets t to 20k for 400M pairs; Meta CLIP [7] tunes t to 170k for scaling the training dataset to 2.5B and keeping the same ratio, 6% of matches from *tail concepts*, that OpenAI CLIP leverages to obtain the 400M pairs. For worldwide data, the data size and the counts of matches differ greatly across languages, so t should be language-*dependent*. Applying a single threshold t to all languages yields suboptimal performance, e.g., a larger t for a language with fewer pairs may yield too many pairs of head concepts and dilutes tail concepts in the curated data (see Sec. 4.2.2).

To derive t for each language, we leverage the *invariance* assumption adopted in Meta CLIP algorithm design, the percentage of *tail matches* (i.e., 6%), and apply it across languages. With this assumption, we determine t in two steps. (1) **From** t_{en} **to** p: we calculate the global tail proportion p for all languages, based on matches of English tail entries decided by t_{en}. (2) **From** p **to** t_{lang}: for each non-English language, we reversely find the language-specific threshold t_{lang} based on the calculated p to ensure the same tail proportion across all languages. Detailed implementation of these two steps is shown as the `t_to_p()` and `p_to_t()` functions in Algorithm 1. With t_{lang}, `entry_counts` is converted to `entry_probs` similarly as in Meta CLIP but for each language.

Putting everything together, Algorithm 1 takes raw image-text pairs $\mathcal{D}$, metadata `M`, and an arbitrary threshold for English t_{en} as input, and outputs a curated dataset of balanced and diverse training pairs, $\mathcal{D}^*$, with three stages. **Stage 1** performs language-specific substring matching for each alt-text, `text`, based on LID results and corresponding metadata, and obtains match counts, `entry_counts`, for each language and entry. **Stage 2** computes thresholds t_{lang} from t_{en}. **Stage 3** samples image-text pairs based on matched entries in `text` with probabilities `entry_probs`. Pairs matched with tail entries are always selected (i.e., probability = 1.0); pairs with head entries have sampling probabilities t_{lang} / `entry_counts[lang]`. Sampled pairs compose $\mathcal{D}^*$(see efficient implementation details in Appendix A.3).

Algorithm 1: Pseudo-code of Meta CLIP 2 Curation Algorithm in Python/NumPy.

```
"""
Input:
D(list) raw (image, text) pairs: each text is assigned with a language "text.lang" by LID;
M(dict) worldwide metadata: key->language code; value(list)->metadata for that language;
t_en(int) English threshold on counts of head/tail entry cutoff: OpenAI CLIP->20k, Meta CLIP
    ->170k;

Output:
D_star(list): curated image-text pairs;
"""

# helper functions to compute t for each language.
def t_to_p(t, entry_count):
    return entry_count[entry_count < t].sum() / entry_count.sum()

def p_to_t(p, entry_count):
    sorted_count = np.sort(entry_count)
    cumsum_count = np.cumsum(sorted_count)
    cumsum_prob = cumsum_count / sorted_count.sum()
    return sorted_count[(np.abs(cumsum_prob - p)).argmin()]

# Stage 1: sub-string matching.
entry_counts = {lang: np.zero(len(M[lang])) for lang in M}
for image, text in D:
    # call substr_match which returns matched entry ids.
    text.matched_entry_ids = substr_match(text, M[text.lang])
    entry_counts[text.lang][text.matched_entry_ids] += 1

# Stage 2: compute t for each langauge.
p = t_to_p(t_en, entry_counts["en"]); t = {}
for lang in entry_counts:
    t[lang] = p_to_t(p, entry_counts[lang])

# Stage 3: balancing via indepenent sampling per language.
entry_probs = {}
for lang in entry_counts:
    entry_counts[lang][entry_counts[lang] < t[lang]] = t[lang]
    entry_probs[lang] = t[lang] / entry_counts[lang]

D_star = []
for image, text in D:
    for entry_id in text.matched_entry_ids:
        if random.random() < entry_probs[text.lang][entry_id]:
            D_star.append((image, text))
            break
```

3.4 Training Framework

Adopting data prepared with worldwide curation in current CLIP training framework addresses the first challenge, but curse of multilinguality still exists as shown in Fig. 1. Thus, we further design the worldwide CLIP training framework. To make our framework and findings generalizable to CLIP and its variants, our framework follows OpenAI/Meta CLIP's training setting and model architecture with three additions: (1) a multilingual text tokenizer, (2) scaling seen training pairs, and (3) study of minimal viable model capacity. The first is required to support worldwide languages and discussed in Sec. 4.2.2 for various choices; details of the latter two are described below.

Scaling seen pairs. Expanding from an English-only dataset and distribution to worldwide naturally increases the number of available image-text pairs. Training CLIP for worldwide distribution with the same number of seen pairs as English CLIP downsamples English training pairs and harms English performance. Hence, we scale seen pairs proportionally to the growth of data size from non-English pairs, to ensure the amount of English seen pairs unchanged during the worldwide CLIP training. This is achieved by increasing the global training batch size, which encourages cross-lingual learning, and meanwhile keeping the other training hyperparameters unchanged. We choose a 2.3× scaling of global batch to reflect that English pairs constitute 44% of our training data. We ablate other choices of global batch size in Sec. 4.2.1.

Minimal viable model capacity. Lastly, we study the minimal model expressivity to enable learning on extra seen pairs and break the curse of multilinguality. As in Fig. 1, we find that even a ViT-L/14 (largest model provided by OpenAI) suffers from the curse due to deficient capacity, and ViT-H/14 is the inflection point to break the curse (strong performance improvement in both English and non-English tasks).

4 Experiment

4.1 Dataset and Training Setup

Following Meta CLIP pipeline, we collect 5B image-text pairs sourced from the Internet that are publicly available. After LID, there are about 44% of alt-texts are in English, which are on par with the scale of English-only data from Meta CLIP [7]. For generalizable recipe and findings, we base our training setup on OpenAI CLIP's ViT-L/14 and Meta CLIP ViT-H/14, except changes necessary for enabling worldwide capability, as described in Sec. 3.4 and ablated in later subsections. The full details can be found in Table 8 and Appendix B.

4.2 Evaluation

We first present the main ablations of Meta CLIP 2 on a wide range of English and multilingual zero-shot transfer benchmarks, along with other multilingual CLIP baselines for comparison (Sec. 4.2.1); then we conduct a comprehensive ablation study on the variants of metadata, curation and tokenizer (Sec. 4.2.2). Lastly, we evaluate the embedding quality of Meta CLIP 2 on downstream tasks for culture diversity (Sec. 4.3) and building MLLM (Sec. 4.4). Additionally, we conduct analysis on embedding alignment and uniformity [41] (Sec. 4.5) and distillation (Sec. 4.6).

				English Benchmarks			Multilingual Benchmarks					
Model	**ViT Size (Res.)**	**Data**	**Seen Pairs**	**IN val**	**SLIP 26 avg.**	**DC 37 avg.**	**Babel -IN**	**XM3600** T→I I→T	**CVQA** EN LOC	**Flicker30k -200** T→I I→T	**XTD-10** T→I I→T	**XTD-200** T→I I→T
XLM-CLIP[6]	H/14(224)	LAION-5B	32B (2.5×)	77.0	69.4	65.5	34.0	50.4 / 60.5	56.1 / 48.2	43.2 / 46.2	87.1 / 88.4	42.5 / 45.2
mSigLIP[16]	B/16(256)	WebLI(12B)	40B (3.0×)	75.1	63.8	60.8	40.2	44.5 / 56.6	51.8 / 45.7	34.0 / 36.0	80.8 / 84.0	37.8 / 40.6
mSigLIP[16]	SO400M(256)	WebLI(12B)	40B (3.0×)	80.6	69.1	65.5	46.4	50.0 / 62.8	56.8 / 49.8	39.9 / 42.0	85.6 / 88.8	42.5 / 45.2
SigLIP 2[17]	SO400M(256)	WebLI(12B)	40B (3.0×)	83.2	73.7	69.4	40.8	48.2 / 59.7	58.5 / 49.0	36.6 / 40.3	86.1 / 87.6	40.3 / 44.5
Meta CLIP[7]	L/14(224)	English(2.5B)	13B (1.0×)	79.2	69.8	65.6	-	- -	- -	- -	- -	- -
	H/14(224)	English(2.5B)	13B (1.0×)	80.5	72.4	66.5	-	- -	- -	- -	- -	- -
Meta CLIP 2	L/14(224)	English	13B (1.0×)	79.5	69.5	66.0	-	- -	- -	- -	- -	- -
		Worldwide	29B (2.3×)	78.8	67.2	63.5	44.2	45.3 / 58.2	59.2 / 55.1	41.9 / 45.8	82.8 / 85.0	41.9 / 44.8
Meta CLIP 2	H/14(224)	English	13B (1.0×)	80.4	72.6	68.7	-	- -	- -	- -	- -	- -
		Non-Eng.	17B (1.3×)	71.4	63.1	61.7	49.9	46.9 / 59.9	59.8 / 56.8	47.5 / 50.5	83.2 / 85.7	46.6 / 49.2
		Worldwide	13B (1.0×)	79.5	71.1	67.2	47.1	49.6 / 62.6	59.9 / 56.0	49.1 / 52.1	85.2 / 87.1	47.0 / 49.7
		Worldwide	29B (2.3×)	81.3	74.5	69.6	50.2	51.5 / 64.3	61.5 / 57.4	50.9 / 53.2	86.1 / 87.5	48.9 / 51.0

Table 1: Main ablation: Meta CLIP 2 breaks the curse of multilinguality when adopting ViT-H/14, with seen pairs scaled (2.3×) proportional to the added non-English data. Meta CLIP 2 outperforms mSigLIP with fewer seen pairs (72%), lower resolution (224px vs. 256px), and comparable architectures (H/14 vs. SO400M). We grey out baselines those are SoTA-aiming systems with confounding factors. Here, numbers of seen pairs are rounded to the nearest integer (e.g., 12.8B->13B).

4.2.1 Main Ablation

We first ablate the effects of scaling seen training pairs and minimal viable model capacity that break the curse of multilinguality, with the following two groups of 6 training runs. Two trainings are in ViT-L/14 on worldwide curated data and its English portion, where global batch size and seen pairs are set to 2.3× and 1.0× compared to OpenAI CLIP and Meta CLIP setting (i.e., 1.0× has 12.8B seen pairs, or 400M for 32 epoches as in OpenAI CLIP). Four runs are on ViT-H/14 with different subsets of curated data to demonstrate the effects of English data helping multilingual performance and vice versa. We denote each run based on subsets trained with and corresponding seen pairs: 1) Worldwide

(2.3×) with the full-fledged worldwide curated data; 2) Worldwide (1.0×) with 1) downsampled; 3) English (1.0×) with English portion of 1); 4) Non-English (1.3×) with the non-English portion.

We adopt the following two groups of zero-shot transfer benchmarks and discuss the limitations in Appendix G: 1) *English-only* benchmarks on **ImageNet (IN val)** [42], **SLIP 26 tasks (SLIP 26 avg.)** [35], and **DataComp 37 tasks (DC 37 avg.)** [28]; 2) *multilingual* benchmarks on **Babel-ImageNet (Babel-IN)** [43] (averaged zero-shot classification on IN with classes and prompts translated into 280 languages), **XM3600** [44] (multilingual text-to-image, T→I, and image-to-text, I→T, retrieval with an averaged recall@1 on 36 languages), **CVQA** [45] (multilingual multi-choice visual question answering with English and local averaged answer accuracy), **Flickr30k-200** [46] (Flickr30k test set translated into 200 languages), **XTD-10** [47] (multilingual image-text retrieval on MSCOCO [48] averaged Recall@1 over 7 languages), and **XTD-200** [46] (XTD10 translated into 200 languages). In Table 1, we observe that Meta CLIP 2 on ViT-H/14 with worldwide data and scaled seen pairs consistently outperforms its counterparts English (1.0×) and Non-English (1.3×), on both English and multilingual tasks, effectively breaking the "curse of multilinguality". The curse still exists in non-scaled seen pairs, Worldwide (1.0×) or smaller ViT-L/14 model even with Worldwide (2.3×)). We further provide gradient conflict analysis to help understand the root of the curse in Appendix C.

Although SoTA is non-goal for Meta CLIP 2, its full recipe demonstrates strong performance with fewer seen pairs (72% of SigLIP series) and lower resolution (224px vs mSigLIP's 256). Meta CLIP 2 surpasses mSigLIP on IN, SLIP 26, and DC 37, and the recent SigLIP 2 on the latter two. More significantly, Meta CLIP 2 sets many SoTA multilingual benchmarks, e.g., Babel-IN (+3.8%), XM3600 (+1.1%/+1.5%), CVQA (+3%/+7.6%), Flicker-30k-200 (+7.7%/+7%), and XTD-200 (+6.4%/+5.8%). SigLIP 2 prioritizes English (90% of its training data in English), while it is worse than mSigLIP on multilingual tasks and Meta CLIP 2 on most English benchmarks except IN. We also provide per-language analysis in Appendix D, cross-lingual translation in Appendix F.

Ablation Steps	**Metadata**	**Alt-texts**	**IN**	**Babel-IN**	**XM3600**		**CVQA**	
					T→I	I→T	EN	LOCAL
1: English CLIP	English	English	**67.5**	-	-	-	-	-
2: remove English filter	English	all, in 1 set	66.9	-	-	-	-	-
3: no language isolation	all, in 1 set	all, in 1 set	62.1	31.2	37.8	49.7	49.8	45.8
4: language isolation with $t_{\text{lang}} = t_{\text{en}}$	all, by lang.	all, by lang.	61.1	**31.5**	37.9	49.4	49.0	46.5
5: language specific t_{lang}	all, by lang.	all, by lang.	64.7	**31.5**	**38.1**	**50.0**	**50.3**	**46.6**

Table 2: Ablation study of metadata and alt-texts combination on ViT-B/32 using English 1.0× and Worldwide 1.0× with mT5 tokenizer. $t_{\text{lang}}/t_{\text{en}}$ are the count thresholds for each language/English.

4.2.2 Ablation on Metadata, Curation, and Tokenizer

We further ablate the transition from metadata and curation focuses solely on English to their worldwide equivalents using the ViT-B/32 encoder for efficiency. We evaluate zero-shot transfer on IN for English and Babel-IN, XM3600 and CVQA for multilingual. As in Table 2, starting from English-only CLIP, we first remove the English filter on alt-texts so that all alt-texts are curated by English metadata, resulting in 0.6% drop on IN, indicating English isolation separating text or metadata by LID before matching is important. Then, we replace English metadata using all metadata merged without separation, yielding even worse English performance but start building up multilingual capability. Next, we isolate substring matching and curate alt-text language-by-language, with the same t_{en} over all languages. This further lowers English performance since t_{en} is too high for non-English and let head data dominate curation. Lastly, we compute t_{lang}, to keep the same ratio of head-to-tail concepts for each language. This improves English and non-English performance, while curse of multilinguality remains unresolved in ViT-B/32 until the main ablation described above.

Tokenizer	**Vocab. Size**	**IN val**	**Babel-IN avg.**	**XM3600**		**CVQA**	
				T→I	I→T	EN	LOCAL
mT5 (mSigLIP) [49]	250k	**64.7**	31.5	38.1	50.0	50.3	46.6
Gemma (SigLIP 2) [50]	256k	63.7	26.1	36.1	47.8	48.3	44.0
XLM-Roberta [51]	250k	64.0	31.1	38.0	49.8	49.8	46.1
XLM-V [52]	900k	**64.7**	**32.7**	**40.0**	**51.4**	**50.4**	**47.4**

Table 3: Ablation study of various multilingual tokenizers with ViT-B/32 and Worldwide 1.0×.

To minimize changes in model architecture, we only swap the English tokenizer for a multilingual one. Four popular tokenizers are studied on our zero-shot benchmarks. As shown in Table 3, the XLM-V vocabulary yields the strongest performance in both the English and non-English world.

Model	**Data**	**Seen Pairs**	**Dollar Street**		**GLDv2**	**GeoDE**
			Top-1	Top-5		
mSigLIP [16]	WebLI(12B) [15]	40B (3.0×)	36.0	62.5	45.3	94.5
SigLIP 2 [17]	WebLI(12B) [15]	40B (3.0×)	36.7	61.9	48.5	95.2
Meta CLIP 2	English	13B (1.0×)	37.2	63.3	52.8	93.4
	Non-English	17B (1.3×)	35.7	61.3	68.6	91.7
	Worldwide	13B (1.0×)	37.2	63.7	65.8	94.3
	Worldwide	29B (2.3×)	37.9	64.0	69.0	93.4

Table 4: Zero-shot classification accuracy on cultural diversity benchmarks. Meta CLIP 2 models are in ViT-H/14 and mSigLIP/SigLIP 2 are in ViT-SO400M. mSigLIP/SigLIP 2 are SoTA-aiming systems with many factors changed and thus greyed out.

4.3 Cultural Diversity

Following protocols in [21] and [38], we perform zero-shot classification and few-shot geo-localization on a range of geographically diverse benchmarks. Specifically, we include zero-shot classification with Dollar Street [53], GeoDE [54], and GLDv2 [55] in Table 4, and few-shot geo-localization [21] on Dollar Street, GeoDE and XM3600 in Fig. 3. We find that only changing the training data distribution, from 13B *English* to 13B *worldwide* pairs, yields significantly better performance, and scaling to 29B *worldwide* pairs improves further, except for the on-par, probably saturated performance in GeoDE. Fig. 3 shows similar trend for evaluating on few-shot geo-localization.

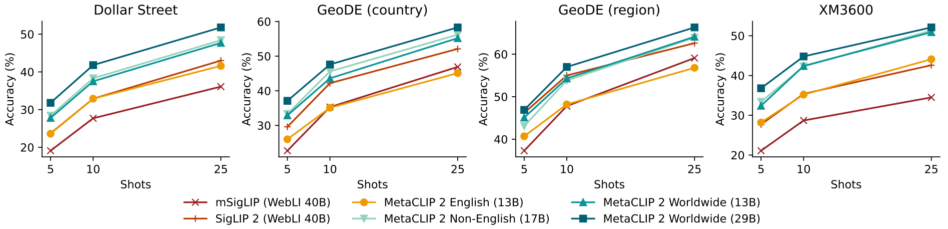

Figure 3: Few-shot geo-localization accuracy on cultural diversity benchmarks.

4.4 Building Multi-modal LLM with Meta CLIP 2

We evaluate the efficacy of Meta CLIP 2 being used as a vision encoder in downstream multilingual MLLMs with a frozen-encoder setup [56], with details in Appendix E. Table 5 shows that switching the frozen vision encoder from mSigLIP to Meta CLIP 2 consistently improves MLLM performance over the wide range of evaluation, including both English and multilingual tasks. Scaling Meta CLIP 2 from 13B to 29B seen pairs shows better results. These results show that curating worldwide data not only enhances retrieval or classification but also transfers to MLLMs.

Model (ViT Size)	**Data**	**Seen Pairs**	*Culture Understanding*				*Captioning*		*Short VQA*				*Multi-subject Reasoning*			
			CVQA		**MaRVL**		**XM100**		**xGQA**		**MaXM**		**xMMMU**		**M3Exam**	
			en	mul	en	mul	en	mul	en	mul	en	mul	en	mul	en	mul
mSigLIP [16] (SO400M)	WebLI [15]	40B (3.0×)	63.2	55.8	86.8	82.9	**30.5**	16.4	63.5	59.5	51.4	52.1	45.4	44.7	57.6	49.1
Meta CLIP 2 (H/14)	English	13B (1.0×)	46.0	55.9	88.1	83.7	30.1	16.6	64.2	60.2	**54.5**	**53.5**	43.4	43.4	59.3	48.5
	Non-Eng.	17B (1.3×)	52.3	57.7	86.5	82.8	30.0	16.4	64.3	**60.5**	53.3	50.4	45.7	45.0	57.6	49.1
	Worldwide	13B (1.0×)	67.1	59.4	87.7	83.5	30.3	16.3	64.1	60.2	52.9	52.9	**47.2**	45.4	**59.6**	47.5
	Worldwide	29B (2.3×)	**67.5**	**59.9**	**88.1**	**83.8**	30.3	**16.8**	**64.3**	60.3	53.3	50.3	46.4	**45.9**	58.9	**50.4**

Table 5: Multilingual MLLM tasks from PangeaBench [56].

We also observed the English-only Meta CLIP 2 performs much worse on CVQA (translated) English benchmark, indicating the importance of training on non-English data. Interestingly, while some tasks like CVQA and M3Exam show clear improvement trends after adding non-English data, some other tasks, e.g. XM100, xGQA and MaXM, exhibit similar performances after switching from English-only to multilingual models. This indicates these benchmarks can be insensitive to the improvement on culturally diverse visual features, but rely more on language ability of MLLMs.

4.5 Alignment and Uniformity

Following [41], we further measure the embeddings quality across different CLIP models. To avoid various unknown biases from different benchmarks, we use 5k holdout image-text pairs not used in our training and report alignment and uniformity scores, where alignment measures the relevance of an image and a text and uniformity measures how images distributed in vision encoder's embedding space. Note that we have no control on whether these 5k pairs are leaked in other baselines. From Fig. 4, we can see that Meta CLIP 2 exhibits good scores in both alignment and uniformity (lower is better), whereas mSigLIP or SigLIP 2 may have non-trivial bias on our collected holdout data.

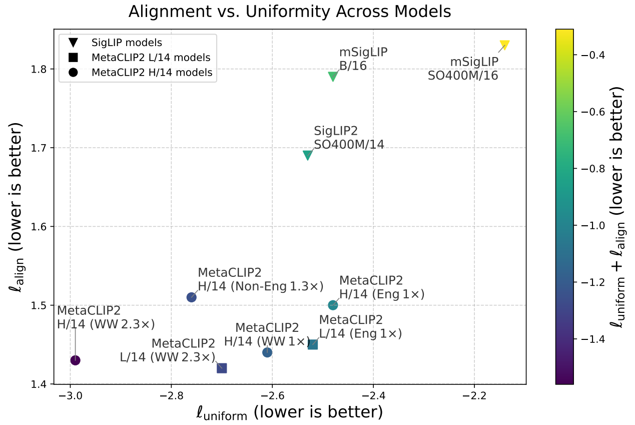

Figure 4: Alignment and uniformity scores [41] calculated on our collected 5k holdout data, WW indicates worldwide data.

4.6 Distilling ViT-H/14 into Smaller Models

To reduce the inference cost while maximizing performance, we distill the ViT-H/14 model into the smaller models such as ViT-L/14. Our results in Table 6 demonstrate that although the teacher model, ViT-H/14, is considerably large, its knowledge can be effectively compressed into a smaller ViT-L/14 student through distillation. The distilled model trained on worldwide data performs better than the from scratch worldwide model on all benchmarks and the English-only model in most cases, while still suffering the curse of multilinguality on ImageNet. More models, including ViT-S/16, ViT-M/16, ViT-B/32, ViT-B/16, ViT-bigG/14, and their corresponding text encoders using the smaller mT5 tokenizer, are available on the official website.

Training	ViT Size (Res.)	Data	Seen Pairs	English Benchmarks			Multilingual Benchmarks					
				IN val	SLIP 26 avg.	DC 37 avg.	Babel -IN	XM3600 T→I I→T	CVQA EN LOC	Flicker30k -200 T→I I→T	XTD-10 T→I I→T	XTD-200 T→I I→T
From Scratch	L/14(224)	English	13B (1.0×)	79.5	69.5	66.0	-	- -	- -	- -	- -	- -
		Worldwide	29B (2.3×)	78.8	67.2	63.5	44.2	45.3 / 58.2	59.2 / 55.1	41.9 / 45.8	82.8 / 85.0	41.9 / 44.8
Distilled	L/14(224)	Worldwide	29B (2.3×)	79.2	70.9	67.4	45.7	47.5 / 60.2	59.8 / 56.5	46.8 / 49.2	83.9 / 86.0	45.0 / 47.2

Table 6: Distillation into smaller models: we show that the distilled ViT-L/14 can be close to the performance of its English counterpart.

5 Conclusion

We present Meta CLIP 2, the first CLIP trained with worldwide image-text pairs from scratch. Existing CLIP training pipelines, designed primarily for English, cannot straightforwardly generalize to a worldwide setting without incurring an English performance degradation due to lack of curation for worldwide data or the "curse of multilinguality". Our careful study suggests that the curse can be broken by scaling metadata, curation, and training capacity, where English and non-English world benefit each other. Specifically, Meta CLIP 2 (ViT-H/14) surpasses its English-only counterpart on zero-shot IN (80.5% → 81.3%) and sets new SoTA on multilingual benchmarks such as XM3600, Babel-IN and CVQA with one single model. We envision our findings along with the fully open-sourced metadata, curation and training code encourage the community to move beyond English-centric CLIP and embrace the worldwide multimodal web.

Acknowledgments

We thank Thao Nguyen and Bernie Huang for insightful discussions, Wei Cheng and Guan Pang for evaluation on downstream use cases. At MIT, Yung-Sung was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[3] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[6] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[7] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024.

[8] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[9] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[10] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[11] Wikipedia. Languages used on the internet. `https://en.wikipedia.org/wiki/Languages_used_on_the_Internet`, 2025. It reports 50.9% of web content is non-English by 2025. Accessed: 2025-05-15.

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

[13] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6848–6854, 2022.

[14] Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459, 2024.

[15] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023.

[16] 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*, pages 11975–11986, 2023.

[17] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[18] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. Graph-rise: Graph-regularized image semantic embedding. *arXiv preprint arXiv:1902.10814*, 2019.

[19] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*, 2022.

[20] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[21] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496, 2024.

[22] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[23] 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 preprint arXiv:2504.01017*, 2025.

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[25] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[27] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[28] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[29] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[31] Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022.

[32] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024.

[35] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

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

[37] Gabriel Oliveira dos Santos, Diego AB Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, et al. Capivara: Cost-efficient approach for improving multilingual clip performance on low-resource languages. *arXiv preprint arXiv:2310.13683*, 2023.

[38] Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025.

[39] Giusepppe Attardi. Wikiextractor. `https://github.com/attardi/wikiextractor`, 2015.

[40] Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[41] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[43] Gregor Geigle, Radu Timofte, and Goran Glavaš. Babel-imagenet: Massively multilingual evaluation of vision-and-language representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5064–5084, 2024.

[44] 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, 2022.

[45] David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, Munkh-Erdene Otgonbold, Munkhjargal Gochoo, Guido Ivetta, Luciana Benotti, Laura Alonso Alemany, Hernán Maina, Jiahui Geng, Tiago Timponi Torrent, Frederico Belcavello, Marcelo Viridiano, Jan Christian Blaise Cruz, Dan John Velasco, Oana Ignat, Zara Burzo, Chenxi Whitehouse, Artem Abzaliev, Teresa Clifford, Gráinne Caulfield, Teresa Lynn, Christian Salamea-Palacios, Vladimir Araujo, Yova Kementchedjhieva, Mihail Minkov Mihaylov, Israel Abebe Azime, Henok Biadglign Ademtew, Bontu Fufa Balcha, Naome A Etori, David Ifeoluwa Adelani, Rada Mihalcea, Atnafu Lambebo Tonja, Maria Camila Buitrago Cabrera, Gisela Vallejo, Holy Lovenia, Ruochen Zhang, Marcos Estecha-Garitagoitia, Mario Rodríguez-Cantelar, Toqeer Ehsan, Rendi Chevi, Muhammad Farid Adilazuarda, Ryandito Diandaru, Samuel Cahyawijaya, Fajri Koto, Tatsuki Kuribayashi, Haiyue Song, Aditya Nanda Kishore Khandavally, Thanmay Jayakumar, Raj Dabre, Mohamed Fazli Mohamed Imam, Kumaranage Ravindu Yasas Nagasinghe, Alina Dragonetti, Luis Fernando D'Haro, Olivier NIYOMUGISHA, Jay Gala, Pranjal A Chitale, Fauzan Farooqui, Thamar Solorio, and Alham Fikri Aji. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[46] Alexander Visheratin. Nllb-clip–train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*, 2023.

[47] Pranav Aggarwal and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*, 2020.

[48] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[49] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[50] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[51] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[52] Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. *arXiv preprint arXiv:2301.10472*, 2023.

[53] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35:12979–12990, 2022.

[54] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137, 2023.

[55] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[56] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024.

[57] Wikipedia. Scriptio continua.

[58] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[59] Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, 2020.

[60] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, 2021.

[61] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland, 2022. Association for Computational Linguistics.

[62] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *ArXiv preprint*, abs/2209.05401, 2022.

[63] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505, 2023.

[64] 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. Ieee, 2009.

[65] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[66] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019.

[67] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

A Implementation Details for Metadata and Curation

A.1 Unigram and Bigram Tokenizer for Special Languages

Most modern languages around the world adopt writing systems that use "spaces" to separate words, except for some of the Asian languages, known as "scriptio continua"[57]. We find several open source tokenizers for many of these languages developed by local communities, as shown in Table 7, in order to properly split text into words while preserving semantic integrity. Note these tokenizers are only used to process Wikipedia dump labeled with the listed wiki codes (e.g., not on alt-texts' substring matching).

Wiki Code	Tokenizer Name	URL
bo,dz	Tibetan Tokenizer	`https://github.com/OpenPecha/Botok`
ja,ryu	Japanese Tokenizer	`https://github.com/SamuraiT/mecab-python3`
km	Khmer Tokenizer	`https://github.com/phylypo/segmentation-crf-khmer`
lo	Lao Tokenizer	`https://github.com/wannaphong/LaoNLP`
my	Myanmar Tokenizer	`https://github.com/ThuraAung1601/mmCRFseg`
th	Thai Tokenizer	`https://github.com/Querela/thai-segmenter`
zh,zh_classical,zh_yue	Chinese Tokenizer	`https://github.com/ckiplab/ckip-transformers`

Table 7: Tokenizers for special languages.

A.2 Fix on Ranking Bigram with Raw PMI

Although Meta CLIP follows OpenAI's description on ranking bigrams by point-wise mutual information (PMI), we observed that raw PMI for bigrams overemphasizes extremely rare pairs (e.g., a bigram appearing only once as a typo), yielding unintuitive high scores. For example "AAAAAB CCCCCB" appears high. To mitigate this, we (i) temper rarity by multiplying PMI with a sublinear count factor and (ii) subtract a baseline using a lower-percentile PMI threshold that roughly marks the onset of meaningfulness.

Let $c(w_1, w_2)$ be the bigram count, i.e., the times that w_1 and w_2 co-exist adjacently in the corpus, $c(w)$ the unigram count, and N the total token count in the corpus. We define

$$\mathrm{PMI}(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1), p(w_2)} = \log \frac{c(w_1, w_2)N}{c(w_1)c(w_2)}.$$

Let $\mathrm{PMI}_{30\%}$ denote the 30^{th} percentile of the empirical PMI distribution over all observed bigrams in a language (a baseline for "starts-to-be-meaningful" associations). Our final bigram score is

$$\mathrm{Score}(w_1, w_2) = \big[c(w_1, w_2){+}1\big]^{0.7} \times \big(\mathrm{PMI}(w_1, w_2) - \mathrm{PMI}_{30\%}\big).$$

This new formulation down-weights spurious high-PMI, low-count bigrams while preserving genuine high-frequency associations; the percentile shift suppresses background noise from weakly associated pairs. After replacing bigram ranking with the new scoring metric, we got the following top-5 bigrams: "United States", "of the", "New York", "such as", "has been".

A.3 Scaling Curation

Worldwide scaling of data curation significantly increases time and space complexity due to storing metadata across hundreds of languages. To efficiently handle this complexity, we leverage several efficient implementations:

- **Efficient String Matching**: We adopt the Aho-Corasick algorithm [6],[7], which utilizes prefix trees (tries), for rapid substring matching. The matching speed is about 2k times faster than Meta CLIP's brute-force implementation, enabling matching with million-scale metadata.

[6] `https://en.wikipedia.org/wiki/Aho-Corasick_algorithm`

[7] `https://pypi.org/project/pyahocorasick`

- **Lazy Metadata Loading**: We pre-build and store the metadata into an Aho-Corasick automaton for each language separately, loading these automaton dynamically and only when encountering a new language for alt-text during processing, thereby minimizing the total number of languages encountered for each shard of data and saving re-compiling time for automation on a new shard.
- **Memory Management for Probabilities**: To address memory constraints during sampling for balancing, we utilize memory-mapped file loading (mmap) to efficiently access counts per entry across all languages, preventing out-of-memory errors caused by loading all the counts from different languages.

Mitigation and Benchmark Deduplication We run a state-of-the-art safety classifier to remove NSFW contents (e.g., adult, sexual, violence) from training data. We also apply face detector to remove human biometrics and personally identifiable information from data. To avoid benchmark leakage, we remove any overlap with ImageNet evaluation sets by performing deduplication using 64-bit hashes. These hashes are generated by applying random projection to feature embeddings from a similarity search model, reducing them to 64 dimensions followed by sign-based quantization.

B Training Setup

To remove confounding factors and generalize our findings, we follow OpenAI CLIP and Meta CLIP training setup with changes for worldwide scaling, detailed in Table 8.

Hyperparameter	OpenAI CLIP / Meta CLIP	Meta CLIP 2
Activation Function	QuickGELU	QuickGELU
Seen Pairs	12.8B	29B (2.3×)
Batch Size	32768	75366 (2.3×)
Learning Rate	4.0e-4 (L/14, H/14)	4.0e-4 (H/14)
Warm-up	2k	2k

Table 8: Hyperparameters of OpenAI CLIP / Meta CLIP vs Meta CLIP 2.

C Analysis on Cross-Lingual Gradient Conflicts

We hypothesize that the primary cause of the "curse of multilinguality" is insufficient *model capacity* to acquire new capabilities (e.g., concepts, domains, and languages) without harming existing ones. A practical indicator of this phenomenon is *language interference* inside the model. Inspired by PCGrad [58], originally proposed for multitask learning and later extended to multilingual XLM settings [59], we design a *gradient conflict* analysis to diagnose interference. Concretely, using XM3600 (36 languages), we compute gradients from model checkpoints and measure cross-lingual interference via cosine similarity between gradients from English examples and those from each non-English language, then average across all non-English languages. All checkpoints are pretrained on the Worldwide 29B schedule; we report the midway (epoch 16) and final (epoch 32) checkpoints.

Gradient Similarities	Worldwide (29B)	
	ViT-L/14	ViT-H/14
Midway (Epoch 16)	0.508	**0.688**
Final (Epoch 32)	0.546	**0.697**

Table 9: Average cosine similarity of gradients (English vs. each non-English language, then averaged) on XM3600. Higher is better (fewer gradient conflicts).

We observe that smaller models (L/14) exhibit lower similarities—i.e., stronger interference and gradient conflicts—than larger ones (H/14) throughout training. With more conflicts, L/14 spends valuable optimization steps mitigating cross-lingual disagreement rather than learning semantics, leading to degraded English performance when trained on multilingual data versus English-only. In contrast, H/14 shows consistently higher similarities even early in training, suggesting reduced conflict that allows the model to *jointly* learn from English and non-English data. This alleviates interference, enables positive transfer, and could be the potential reason why it *breaks* the curse of multilinguality.

D Correlation Between Training Data Volume and Performance Among Languages

We examine XM3600 zero-shot retrieval for the *top-10* languages by training volume versus the remaining languages. We have the following observations: (1) **Volume effect exists.** Top-10 languages average 62.6/75.6 (T→I/I→T) versus 47.2/59.9 for others, indicating a clear volume effect on average. (2) **English is not best.** Despite the largest volume, English lags behind German (the best among all listed; 69.2/83.6). (3) **Strong tail performers exist.** 18 non-top-10 languages (e.g., `it`, `hu`, `ro`, `uk`) exceed 50% on both directions, showing that factors beyond raw volume matter.

What is beyond the volume effect? We hypothesize two additional drivers: (i) *linguistic/cultural proximity* (benefiting from transfer with closely related or culturally overlapping languages), and (ii) *structural characteristics/expressiveness* of the language (e.g., morphology, tokenization efficiency, domain overlap with pretraining corpora). These factors can amplify or dampen the benefit of volume during multilingual pretraining.

Language	Text→Image	Image→Text
en	51.6	62.2
es	57.2	72.5
fr	67.1	78.5
zh	61.1	72.6
ru	67.8	79.9
ja	65.1	79.9
id	65.8	78.3
pt	60.4	72.6
de	**69.2**	**83.6**
vi	61.1	76.2
Avg (Top-10)	62.6	75.6

Language	Text→Image	Image→Text
ar	47.4	60.8
bn	39.4	47.1
cs	51.0	66.1
da	61.0	75.1
el	52.1	68.4
fa	56.9	70.3
fi	59.3	73.7
fil	24.8	36.7
hi	26.1	41.8
hr	57.3	72.9
hu	63.9	76.5
it	64.0	78.2
he	60.8	76.2
ko	54.8	70.1
mi	0.5	1.2
nl	53.2	66.9
no	57.7	73.2
pl	61.4	75.9
quz	2.5	6.5
ro	64.8	77.8
sv	57.6	73.8
sw	10.0	16.6
te	26.1	37.1
th	57.7	71.4
tr	55.7	68.4
uk	60.0	74.7
Avg (Non–Top-10)	47.2	59.9

Table 10: XM3600 Recall@1 (higher is better). Left: top-10 languages by training volume. Right: languages outside the top-10.

E Setup and Details of MLLM Evaluation

While zero-shot classification and retrieval benchmarks demonstrate the standalone capabilities of Meta CLIP 2, real-world applications often require grounding in generative models. Thus, we conduct the experiment of using Meta CLIP 2 model as vision encoder in MLLM in Sec. 4.4. Here, we provide more details about the settings and task details.

E.1 Training Setup

For evaluation, we leverage the open-sourced MLLM [56] implementation and apply exact the same model setup except that we vary the vision backbone with each vision encoder to be evaluated. The MLLM is trained as the following. First, a vision-language connector is trained to align the vision encoder features to the language backbone. Then, we fine-tune the MLLM with 6M samples spanning 39 languages. We followed the same training recipe, including learning rate 1e-3 and batch size 128 for vision-language connector training, and learning rate 2e-5 and batch size 512 for finetuning. Both stages are coupled with a cosine learning rate scheduler with warmup ratio of 0.03. For evaluating the quality of embeddings from vision encoder, we make one change in the training that during the finetuning stage, we freeze the vision backbone, while all weights are finetuned in the original setting.

E.2 Task Details

The trained MLLM models with varying vision encoders are then evaluated on PangeaBench [56], which includes following tasks:

Culture Understanding: *CVQA* evaluates model's capability in cultural reasoning using visual questions with diverse global contexts across 31 languages and 13 scripts [45]. Unlike our embedding-only experiment in Table 1, here we follow the generative setting to select answers based on the MLLM's output probabilities. *MaRVL* tests cross-lingual visual reasoning with culturally grounded entailment tasks in multiple non-English languages [60].

Captioning: *XM100* is a compact multilingual captioning benchmark with 100 diverse images selected from XM3600 across 36 languages for efficient and diverse evaluation [44].

Short VQA: *xGQA* extends the GQA dataset to multilingual settings to measure cross-lingual VQA performance [61]. *MaXM* offers multilingual VQA tasks covering different scripts and question types to test model understanding beyond English [62].

Multi-subject Reasoning: *xMMMU* is a translated subset of MMMU validation questions into six languages to evaluate academic reasoning in a multilingual setup. *M3Exam* poses real-world multimodal exam questions across subjects, requiring both visual and textual comprehension [63].

F Cross-Lingual Translation Capability

We probe whether the model acquires cross-modal *translation* behavior without explicit supervision. Given an image that visually depicts the Chinese character "狗" ("dog"), we compute cosine similarities between the image embedding and candidate text prompts across languages, then rank the candidates. As expected, the exact Chinese character "狗" yields the highest score. Notably, the Japanese word "いぬ" (dog) ranks highest within Japanese candidates and achieves a substantially higher similarity than English "dog," suggesting stronger cross-lingual coupling between Chinese and Japanese scripts.

Word	Description	Cosine Sim.
狗	"dog" in Chinese (exactly visualized on image)	0.54325
犬	"dog" in Chinese, literary/ancient usage	0.04636
猫	"cat" in Chinese	0.00025
豺	"jackal" / 'wild dog" in Chinese	0.03427
狼	"wolf" in Chinese	0.01405
dog	English "dog"	0.08239
diagram	Unrelated word	0.00143
cat	English "cat"	0.00005
puppy	English "puppy"	0.02826
hound	English "hound"	0.05586
いぬ	"dog" in Japanese	0.19320
ねこ	"cat" in Japanese	0.00064

Table 11: Cosine similarities between the image of the character "狗" and multilingual text prompts. Higher is better.

We make the following observations: (1) **Cross-modal alignment reflects cross-lingual relations.** The strong scores for "狗" and "いぬ" indicate the model maps visual text to semantically equivalent words across languages. (2) **Script proximity matters.** Japanese scores exceed English for this example, plausibly due to closer linguistic/script overlap with Chinese. (3) **Robustness amid noise.** Despite inevitable Internet-scale noise and partial misalignment between OCR-like visual tokens and alt-text, the model still exhibits emergent cross-lingual translation behavior—supporting the premise that remaining faithful to natural data distributions can mitigate noise effects.

G Limitation on Benchmark

High-quality benchmarks are essential for researchers to understand the efficacy of proposed changes. After decades of meticulous efforts, the community has established reliable and diverse datasets to enable research advancement in vision and multimodal areas [64, 42, 1]. However, these datasets consist mainly of content scraped from North America and Western Europe (NA+EU) and focus on English [65, 66]. It is a long and resource-intensive endeavor to build similar benchmarks for unbiased and comprehensive evaluation of worldwide data and resulting representations, for the world outside NA+EU or English-speaking community, due to the complexity of covering diverse concepts across geo-locations, cultures, and languages. XM3600 [44] aims to build geographically diverse datasets by selecting images from Open Images Dataset [67] based on metadata of GPS coordinates, but later research [21] suggests Open Images Dataset is biased towards Western images or specific activities (e.g., tourism). GeoDE [54] recruits human workers on crowdsourcing platform to collect geographically diverse images for predefined object classes. Crowdsourcing is an economic way to collect human annotations, but the demographic background and proficiency of the workers are not guaranteed, nor is the quality of the collected data. Few efforts such as CVQA [45] attempt to scale annotation and control quality simultaneously by utilizing experts in machine learning community or existing materials as seeds. These efforts offer relatively unbiased evaluation with reasonable coverage in capabilities (e.g., cultural diversity, multimodal problem solving for exam questions across countries) of interests. We believe benchmarks of similar quality but built for evaluating more general and comprehensive capabilities will reveal the true potential of worldwide data and resulting representations developed in this work.

H Licenses for Existing Assets

Below we list the licenses for all existing assets used in this work, including code, models, and datasets.

- **Meta CLIP Code:**
  - License: CC-BY-NC
  - URL: `https://github.com/facebookresearch/MetaCLIP`
- **Wikipedia Dumps:**
  - License: CC BY-SA 4.0
  - URL: `https://dumps.wikimedia.org/`
- **WordNet:**
  - License: WordNet License 3.0
  - URL: `https://wordnet.princeton.edu/`
- **Multilingual WordNet:**
  - License: WordNet License 3.0
  - URL: `https://omwn.org/`
- **Models Used for Evaluation:**
  - **XLM-CLIP**
    - License: MIT
    - URL: `https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k`
  - **mSigLIP-SO400M**
    - License: Apache 2.0
    - URL: `https://huggingface.co/google/siglip-so400m-patch16-256-i18n`
  - **mSigLIP-B/16**
    - License: Apache 2.0

        - URL: `https://huggingface.co/timm/ViT-B-16-SigLIP-i18n-256`
    - **SigLIP 2-SO400M**
        - License: Apache 2.0
        - URL: `https://huggingface.co/timm/ViT-SO400M-14-SigLIP2`
- **Code and Data for Evaluation:**
    - CLIP benchmark, including Flickr30k-200, XTD-10, XTD-200
        - License: MIT
        - URL: `https://github.com/LAION-AI/CLIP_benchmark`
    - DataComp evaluation
        - License: MIT
        - URL: `https://github.com/mlfoundations/datacomp`
    - SLIP evaluation
        - License: MIT
        - URL: `https://github.com/facebookresearch/SLIP`
    - Pangea, including the MaRVL, XM100, xGQA, xMMMU, M3Exam datasets.
        - License: Apache 2.0
        - URL: `https://github.com/neulab/Pangea`
    - XM3600
        - License: CC BY 4.0
        - URL: `https://google.github.io/crossmodal-3600/`
    - CVQA
        - License: CC BY-SA 4.0
        - URL: `https://cvqa-benchmark.org/`
    - GLDv2
        - License: CC BY 4.0
        - URL: `https://github.com/cvdfoundation/google-landmark`
    - Babel-ImageNet
        - License: BabelNet Non-Commercial License and MIT license
        - URL: `https://github.com/gregor-ge/Babel-ImageNet`
    - ImageNet
        - License: non-commercial research and educational purposes, detailed in `https://www.image-net.org/download.php`
        - URL: `https://www.image-net.org`
    - Dollar Street
        - License: CC BY-SA 4.0
        - URL: `https://www.kaggle.com/datasets/mlcommons/the-dollar-street-dataset`
    - GeoDE
        - License: CC BY 4.0
        - URL: `https://geodiverse-data-collection.cs.princeton.edu/`

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: In Sec. 3, we introduce the metadata construction and scalable curation algorithm that enable our models to learn from 300+ languages without relying on translation or private data. In Sec. 4, we demonstrate that our approach breaks the curse of multilinguality by scaling model capacity and seen training pairs, showing mutual benefits between English and non-English performance. We also provide comprehensive ablations (Sec. 4.2.2) on metadata, curation, multilingual tokenizers, and additional downstream evaluations on cultural diversity tasks (Sec. 4.3). Taken together, the contributions and scope outlined in the abstract and introduction are fully supported by the methodology and analysis throughout the paper and appendix.

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: In Sec. 1, we discussed explicitly that the goal of this paper is to offer generalizable recipes and comparable results to mainstream CLIP architectures. Pushing SoTA performance is not the direct goal and we encourage the community to adopt our recipe for their own CLIP system. More details are discussed in Appendix G.

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: [NA]

Justification: This paper does not include theoretical results.

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 disclose all the experiment details in Sec. 3, 4.2 and Appendix A. We further provide the source code in supplemental material to make the results reproducible and verifiable.

   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 the source code on `https://github.com/facebookresearch/MetaCLIP` with descriptions and instructions. We use open-access public data from the Internet and the open benchmarks for evaluation. We disclose all the experiment details in Sec. 4 and Appendix A, B, for reproducing the experiment results.

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 experiment details in Sec. 4, Appendix A (data curation details), and Appendix B (training details) for reproducing the experiment results.

   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: [Yes]

   Justification: In our experiment, we run the training of ViT-B/32 model for three times report the standard deviation for ImageNet accuracy in Appendix B. For larger models that take weeks to train, i.e. ViT-L/14 and ViT-H/14, we only run the experiment once because it is difficult and expensive to repeat the large-scale pretraining.

   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: Discussed in Appendix B.

   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: Yes, this research conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - 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 include the discussion of broader impacts in the end of introduction.

    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: [Yes]

    Justification: As discussed in Appendix A.3, we use safety classifier to remove NSFW, CSAM and human face content when processing the training data.

    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 cite assets, such as code, data, and models, by their papers or repositories and follow licenses and terms properly. More details about the licenses of used assets are discussed in Appendix H.

    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: Yes, check `https://github.com/facebookresearch/MetaCLIP` for details.

    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: [NA]

    Justification: This 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: [NA]

    Justification: This 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: [NA]

    Justification: We only fine-tune LLMs and test them on the standard MLLM benchmarks to evaluate the feature quality of our visual encoders for solving standard VQA tasks, as described in Sec. 4.4 and Appendix E. Otherwise, we do not use LLM in developing any other components in this research.

    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.