Behind Maya: Building a Multilingual Vision Language Model

Nahid Alam[1*,2], Karthik Reddy Kanjula[2], Surya Guthikonda[3,2], Timothy Chung[4,2], Bala Krishna S Vegesna[5], Abhipsha Das[2], Anthony Susevski[2], Ryan Sze-Yin Chan[6], S M Iftekhar Uddin[2], Shayekh Bin Islam[7,2], Roshan Santhosh[8], Snegha A[9], Drishti Sharma[2], Chen Liu[10], Isha Chaturvedi[2], Genta Indra Winata[11*], Ashvanth.S[2], Snehanshu Mukherjee[12], Alham Fikri Aji[13]

Abstract

In recent times, we have seen a rapid development of large Vision-Language Models (VLMs). They have shown impressive results on academic benchmarks, primarily in widely spoken languages but lack performance on low-resource languages and varied cultural contexts. To address these limitations, we introduce Maya, an open-source Multilingual VLM. Our contributions are: 1) a multilingual image-text pretraining dataset in eight languages, based on the LLaVA pretraining dataset; and 2) a multilingual image-text model supporting these languages, enhancing cultural and linguistic comprehension in vision-language tasks. Code available at https://github.com/nahidalam/maya.

1. Introduction

Recent progress in Large Language Models (LLMs) and vision encoders like CLIP [28] and SigLIP [40] has greatly advanced Vision-Language Models (VLMs). Models such as Flamingo [2], LLaVA [19, 20], KOSMOS [25, 27], Florence-2 [37], and Molmo [9] excel at image captioning, VQA, and reasoning. Qwen2-VL's M-RoPE [32, 36] and PaLI's joint modality and cross-lingual scaling [6, 7] further improve multimodal understanding. We acknowledge that although Aya Vision [8] showcases multilingual capabilities in 23 languages, it was released subsequent to the development of Maya.

In general though, VLMs still underperform in low-resource languages, struggling with cultural context and localized visual concepts [15]. A key reason is the lack of high-quality multilingual multimodal data. Most pretraining datasets—COCO [18], Flickr30K [38], LAION [30], Visual Genome [16], and LLaVA [20]—are English-centric, others [10, 35] remain limited in scale and cultural coverage. To address these challenges, we introduce Maya, an open-source multilingual VLM that extends multimodal capabilities to eight languages. Our key contributions:

1. a novel multilingual image-text pretraining dataset consisting of *550K* samples for future development of multilingual VLMs,
2. a new multilingual VLM that demonstrates improved performance in understanding cultural and linguistic nuances compared to PALO-7B [22] on LLaVA-Bench-In-The-Wild [20], offering a multilingual alternative to LLaVA [20].

2. Dataset Creation and Filtering

2.1. Dataset Creation Methodology

Recently, PALO [22] and Pangea [39] have created a multilingual image-text dataset to build multilingual multimodal models. However, these multilingual datasets often suffer from data quality issues and distribution biases across languages. For example, in the PALO dataset, the distribution varies significantly between English and other languages [10, 13, 22]. To address these limitations, we present a novel pre-training dataset tailored for LLaVA's architecture that is both multilingual and optimized for diverse language representation. Our dataset introduces rigorous processes for toxicity analysis, distribution balance, and quality control, ensuring consistent and reliable performance in multiple languages and modalities, as shown in Figure 1. We expand the original English LLaVA dataset, which contains 550K samples, to include seven additional languages—Chinese, French, Spanish, Russian, Hindi, Japanese, and Arabic—yielding a total of 4.4 million samples, equally distributed across all eight languages.

Our approach encompasses three components: 1) parallel dataset creation using a hybrid translation method, 2) prompt engineering optimization, and 3) scalable generation of pre-training datasets. This pipeline integrates multiple language models, such as [3, 24, 33], alongside specialized multilingual models like Aya 35B [4], to ensure quality, cross-lingual data suitable for multilingual applications.

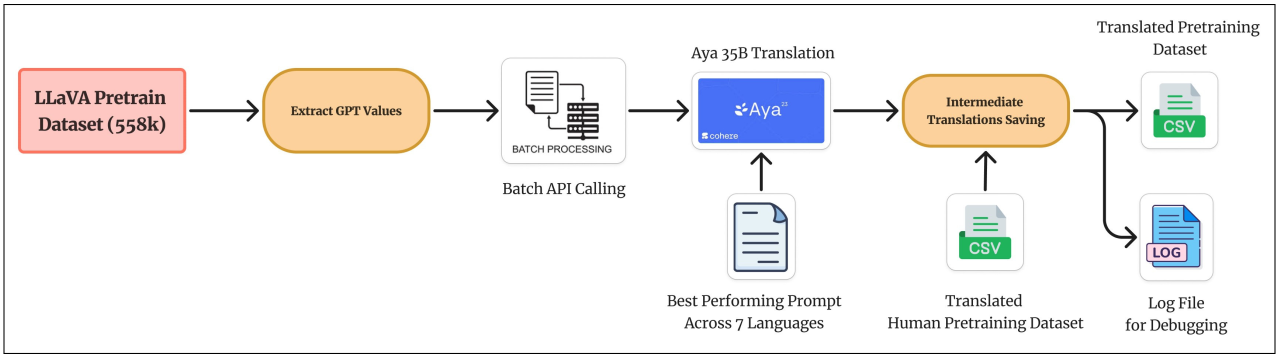

Figure 1. Pretrain Dataset Preparation Process

2.1.1. Multilingual LLaVA Pretrain Dataset

Our Multilingual Pretrain dataset builds on the LLaVA dataset [20], using image-text pairs and the corresponding GPT responses. Our approach started with sampling to select 30 diverse samples per language, guided by Length Analysis (LA), Flesch Reading Ease (FRE), and Flesch-Kincaid Grade Level (FKGL) metrics [34] to obtain a representative GPT response in English. A cascaded translation and verification process ensures quality: initial translation using Google Translate, followed by back-translation and final human review with the help of [1, 3, 33] to generate the prompt engineering dataset in 8 languages.

2.1.2. Prompt Engineering and Evaluation

During prompt engineering, we evaluate prompts per language using a BLEU score-based process. We construct a prompt evaluation dataset following Figure 1, translating six sample prompts into seven languages using Aya 35B. These translations are compared with reference translations of the prompt engineering dataset using BLEU [26] and N-gram scores [5, 31]. In Arabic, Chinese, French, Hindi, Japanese, Russian and Spanish, Preamble 6 consistently produces the highest BLEU scores per Ngram (typically 0.4–0.5), showing a clear improvement from Type 5. Figure 2 shows Preamble 6 with the largest area in 1- to 4-grams, indicating better fidelity to the phrase level and consistent performance across languages. We adopt Preamble 6 as our final prompt template, shown in Listing 1, and integrate it into our translation framework.

2.1.3. Translation Framework Design

Our translation framework uses the best preamble identified in the prompt engineering evaluation step. The prompt includes 1) standardized input-output formatting to maintain uniformity across languages, 2) example-driven instruction sets tailored for complex translations, and 3) integrated validation triggers to ensure quality control. Using Aya 35B translation capabilities, our framework achieves more than 0.47 average BLEU scores in seven languages.

Listing 1. Translation Instructions

```
## Instructions
You are an expert in translations.
Your job is to translate the input to Japanese in
the given chat.

Ensure that:
{Specific Things to Consider while Translating}

Note: {Extra Constraints on Output Generation}

## Examples
### Example 1
Input:
{Input Sentence}
Expected Output:
{Translated Sentence}
```

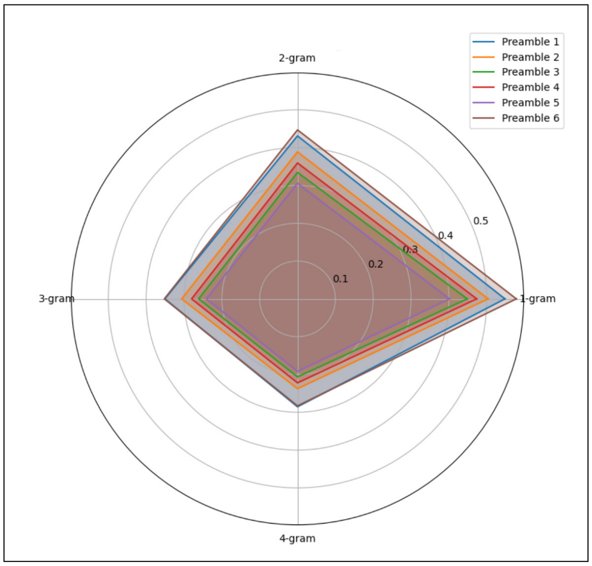

Figure 2. Radar chart of N-gram averages by preamble

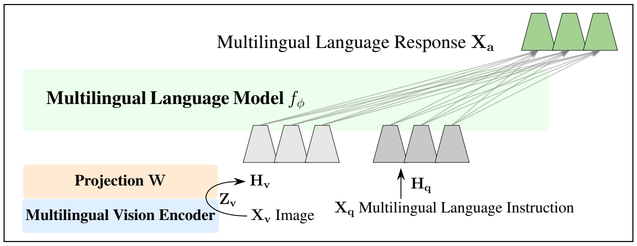

Figure 3. Maya Architecture adapted from LLaVA [21]

2.1.4. Scalable Dataset Generation

For large-scale dataset generation, we implement a batch processing pipeline integrated with Aya 35B API as shown in Figure 1. The extracted GPT values from LLaVA Pretrain Dataset are passed through the Aya 35B batch-optimized API calling with intermediate translation checkpointing. The pipeline does necessary error handling and comprehensive logging for quality tracking. We implement version control for intermediate translations and maintain detailed debug logs, ensuring reproducibility and enabling systematic error analysis. The translation pipeline enables efficient processing of the 550K samples while maintaining translation quality.

3. Multilingual Multimodal Instruction Tuning

3.1. Model Architecture

We draw inspiration from LLaVA 1.5 [19] for model architecture. We employ the Aya-23 8B [4] model as our LLM f_ϕ because of its multilingual capability. Aya-23 has 8 billion parameters, an 8K context window, and is trained across 23 languages. Our dataset, however, includes 8 of these 23 languages, aligning with our objective of optimizing Maya for a diverse yet focused linguistic range.

For the vision encoder, we opted for SigLIP[1] [40] rather than CLIP [28], which is traditionally used in LLaVA. This choice is motivated by SigLIP's strong performance, multilingual adaptability, and capacity for variable-length patch sizes. Unlike CLIP, SigLIP supports scalable positional embeddings, enabling it to accept inputs of varying dimensions through positional embedding interpolation. This flexibility makes SigLIP particularly suitable for our purpose. For each input image X_v, we get the visual features from SigLIP, $Z_v = g(X_v)$. A trainable projection matrix W then converts the image features Z_v into language features H_v.

3.2. Pretraining

For image-text alignment, we used a projection matrix W that brings image features X_v closer to language features. This projection matrix is a simple 2-layer MLP with GELU activation [12], as in LLaVA 1.5 [19]. Although we experimented with 4- and 8-layer MLPs, the 2-layer configuration consistently achieved the lowest training loss. Advanced alignment techniques, such as gated soft-attention in Flamingo [2], Q-Former in BLIP-2 [17], or pooling from MM1 [23] as alternatives to the projection layer, are set aside for future work. For each image X_v, we used the single-turn conversation data

$$(X_q^1, X_a^1, \ldots, X_q^T, X_a^T)$$

where T is the total number of turns from LLaVA [20]. We pretrained Maya on the translated dataset in eight languages. Image inputs were cropped to 256x256 for compatibility with the SigLIP encoder. For training, we used 8xH100 GPUs with 80GB DRAM, per-device batch size of 32 and a global batch size of 256. A learning rate of 1e-3 and a cosine scheduler were applied during training. The pretraining process only trains the projection matrix and takes about 20 hours.

3.3. Finetuning

We instruction-finetuned our pretrained Maya model using the PALO 150K instruction-tuning dataset [22]. During finetuning, we observed that Low Rank Adaptation (LoRA) [14] produced suboptimal results, particularly when both adapter matrices A and B were updated with the same learning rate [11]. Based on these findings, we opted against using LoRA and instead implemented full finetuning on 8xH100 GPUs. Our finetuning configuration used a per-device batch size of 16 and a global batch size of 128. Finetuning process took about 48 hours.

We kept both the vision encoder and the language encoders frozen during the training process. We did pretraining and finetuning for both versions of our dataset: the pretraining dataset translated into 7 languages.

4. Results

We evaluate Maya on the PALO multilingual benchmark [22], as shown in Table 1. Although Maya was pretrained on only eight languages, it was finetuned on the PALO instruction dataset covering ten languages, and is thus evaluated on all ten. Maya outperforms all 7B models and is competitive with 13B models, surpassing LLaVA-13B on average and trailing PALO-13B by a small margin. Among the eight common languages, Maya outperforms PALO-7B in five. We attribute this to Maya's multilingual pretraining, in contrast to PALO's English-only pretraining. Notably, Maya outperforms both PALO and LLaVA in Arabic for both size classes, likely due to its handling of root-based morphology and effective prompt design. We also evaluate Maya on English only dataset across various benchmarks as shown in Table 2.

[1] `siglip-base-patch16-256-multilingual` from https://huggingface.co/google/siglip-base-patch16-256-multilingual

Model	English	Chinese	French	Spanish	Russian	Japanese	Arabic	Hindi	Bengali	Urdu	Avg.
Maya (8B)	61.5	61.7	61.0	60.4	62.2	63.7	**63.4**	64.0	50.8	55	60.4
LLaVA-7B	67.9	55.7	62.4	64.5	55.3	59.2	38.9	29.4	13.9	21.8	46.9
PALO-7B	64.2	55.7	58.3	61.0	57.4	57.5	57.8	57.6	51.7	55.3	57.7
	-6.4	+6.0	-1.4	-4.1	+4.8	+6.2	+5.6	+6.4	-0.9	-0.3	+2.7
LLaVA-13B	**69.5**	**62.9**	**67.5**	64.6	62.3	**65.3**	37.2	27.8	20.4	22.1	49.9
PALO-13B	65.5	62.1	66.4	**65.9**	**62.4**	60.6	56.9	**66.8**	**53.5**	**59.6**	**61.9**
	-8.0	-1.2	-6.5	-5.5	-0.2	-1.6	+6.5	-2.8	-2.7	-4.6	-1.5

Table 1. A comparison of LLaVA and PALO with Maya on eight languages adapted from LLaVA-Bench (In-the-Wild). Values underlined indicate best performance within size class and values in **bold** indicate best performance across all models tested. We provide performance differences between Maya and the best competing model within the size classes where red indicates where Maya is performing better and blue indicates where Maya is performing worse than the best in the size class. “Avg.” represents the average over all the languages.

GQA	VizWiz	ScienceQA	TextVQA	POPE-random	MMBench	MM-VeT	MME (P+C)
57.79%	34.92%	70.27%	47.01%	85.30%	71.10%	29.8	72.45%

Table 2. Accuracy of Maya models on English Language across multiple benchmarks. Abbreviations: P+C = Perception + Cognition.

Figure 4. Example image from LLaVA-Bench (In-the-Wild) [20].

Figure 5 shows VQA outputs in various languages for Figure 4. The Bengali response is the most detailed, identifying both meat and the wooden table. Spanish, French, and Hindi mention the meat but miss the table. Chinese and Japanese outputs are similar to English in detail.

English	A bowl of noodles with vegetables in it.
Spanish	Un plato de fideos asiáticos con verduras y carne.
French	Un bol de nouilles asiatiques avec des légumes et de la viande.
Chinese	一个装满面条和蔬菜的碗，可能是一碗汤面。
Japanese	麺が野菜と一緒に入ったボウルで、おそらくラーメンです。
Russian	Миска с лапшой, возможно, раменом, с различными овощами сверху.
Arabic	وعاء من النودلز مع مكونات مختلفة، ربما حساء.
Hindi	एक कटोरा जिसमें नूडल्स होते हैं, जो संभवतः एक एशियाई सूप होता है, जिसमें सब्जियाँ और मांस होते हैं।
Bengali	একটি বাটি নুডলস সুপ যা সবজি এবং মাংস দিয়ে পূর্ণ করা হয়েছে একটি কাঠের টেবিলে রাখা হয়েছে।
Urdu	یک باؤل نوڈلز جس میں سبزیوں کے ساتھ ہے اور ایک چمچ کے ساتھ ہے-

Figure 5. Maya output for prompt (with image from Figure 4): Please describe the food in {language} in 1 sentence.

5. Conclusion

Maya enables high-quality multilingual, multimodal AI content generation, addressing gaps in low-resource languages. We curate data to minimize harmful content, though some traces may remain. Future work includes refining cross-modal alignment with alternative projection layers, unfreezing decoder layers for fine-tuning, and expanding pretraining to Bengali and Urdu. We also plan to scale the instruction-tuning dataset to 665K examples, improve translation via language-specific preambles, and benchmark on datasets like PangeaBench [39] and CVQA [29] for broader, robust evaluation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 3

[3] Anthropic. Introducing the next generation of Claude, 2024. 1, 2

[4] Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. Aya 23: Open Weight Releases to Further Multilingual Progress. *arXiv preprint arXiv:2405.15032*, 2024. 1, 3

[5] Eric Brill, Radu Florian, John C. Henderson, and Lidia Mangu. Beyond n-grams: Can linguistic sophistication improve language modeling? In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. 2

[6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *arXiv preprint arXiv:2209.06794*, 2022. 1

[7] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*, 2023. 1

[8] Cohere. Aya Vision, 2025. 1

[9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. *arXiv preprint arXiv:2409.17146*, 2024. 1

[10] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, 2016. Association for Computational Linguistics. 1

[11] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient Low Rank Adaptation of Large Models. *arXiv preprint arXiv:2402.12354*, 2024. 3

[12] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 3

[13] Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shaoyen Tseng, and Vasudev Lal. Why do LLaVA Vision-Language Models Reply to Images in English? *arXiv preprint arXiv:2407.02333*, 2024. 1

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[15] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, 2020. Association for Computational Linguistics. 1

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2023. 1, 3

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023. 1, 2, 3, 4

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 3

[22] Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. PALO: A Polyglot Large Multimodal Model for 5B People. *arXiv preprint arXiv:2402.14818*, 2024. 1, 3

[23] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 3

[24] OpenAI. Hello GPT-4o, 2024. 1

[25] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-G: Generating Images in Context with Multimodal Large Language Models. *ArXiv*, abs/2310.02992, 2023. 1

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2

[27] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding

Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*, 2023. 1

[28] 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. 1, 3

[29] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. *arXiv preprint arXiv:2406.05967*, 2024. 4

[30] 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. 1

[31] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2

[32] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, 2021. 1

[33] 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. 1, 2

[34] Textstat Contributors. Textstat: An Easy to Use Library to Calculate Statistics from Text, 2024. Python package for calculating readability statistics. 2

[35] Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*, 2022. 1

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[37] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 1

[38] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1

[39] 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. 1, 4

[40] 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. 1, 3

6. Appendix

6.1. Author Affiliations

[1] Cisco Meraki
[2] Cohere Labs Community
[3] Indiana University Bloomington
[4] Imperial College London
[5] Georgia Institute of Technology
[6] The Alan Turing Institute
[7] Bangladesh University of Engineering and Technology
[8] University of Pennsylvania
[9] IIT Bombay
[10] TU Darmstadt
[11] Capital One
[12] IIT Dhanbad
[13] MBZUAI

6.2. Acknowledgment

We cannot express our gratitude enough to the Cohere Labs Community for bringing the open research community together which was instrumental to building Maya. The generous API credit for multilingual image-text translation from Cohere Labs Community has a direct impact on where the model exists today.