Anthropogenic Regional Adaptation in Multimodal Vision-Language Model

Samuel Cahyawijaya[1,2], **Peerat Limkonchotiwat**[3,2], **Tack Hwa Wong**[4,2], **Hitesh Laxmichand Patel**[5], **Amit Agarwal**[5], **Manuel Antonio Rufino**[6], **Carlos Rafael Catalan**[6], **Muhammad Reza Qorib**[7], **Vicky Feliren**[8], **Holy Lovenia**[2], **Aye Hninn Khine**[9,2], **Frederikus Hudi**[10,2], **David Anugraha**[11], **Alham Fikri Aji**[12,2], **Romrawin Chumpu**[13], **Viet-Thanh Pham**[14], **Minghan Wang**[14], **Mohamed Fazli Imam**[12], **Ruochen Zhang**[15,2], **Joseph Marvin Imperial**[16,31], **Khumaisa Nur'aini**[8], **Do Xuan Long**[13], **Musa Izzanardi Wijanarko**[8], **Joel Ruben Antony Moniz**[17], **Patrick Amadeus Irawan**[12], **Hanif Muhammad Zhafran**[18], **Isaiah Flores**[19], **Salsabila Zahirah Pranida**[12], **Jun Kevin**[20], **Jostin Jerico Rosal**[21], **Patricia Nicole Monderin**[6], **Kun Kerdthaisong**[22], **Ahmad Mustafid**[23], **My Chiffon Nguyen**[2], **Natchapon Jongwiriyanurak**[24], **Siva Worajitwannakul**[16], **Haochen Li**[25], **Adrian Xuan Wei Lim**[13], **Bin Wang**[26], **Muhammad Ravi Shulthan Habibi**[27,2], **Lynnette Hui Xian Ng**[7], **Mithil Bangera**[28], **Yeshil Bangera**[28], **Priyaranjan Pattnayak**[23], **Dun Li Chan**[29], **Sherissa Caren Djuniwar**[30], **Cho Chan Myei Oo**[32], **Hee Ming Shan**[12]

[1]Cohere [2]SEACrowd [3]AI Singapore [4]Universiti Teknologi PETRONAS
[5]Oracle [6]Samsung R&D Institute Philippines [7]Carnegie Mellon University
[8]Monash University, Indonesia [9]King Mongkut's University of Technology Thonburi
[10]Nara Institute of Science and Technology [11]Stanford University [12]MBZUAI
[13]National University of Singapore [14]Monash University, Australia [15]Brown University
[16]University of Bath [17]Mila - Quebec AI Institute [18]Institut Teknologi Bandung
[19]Ateneo de Manila University [20]Universitas Pelita Harapan
[21]Seoul National University of Science and Technology [22]Thammasat University
[23]Independent [24]University College London [25]Nanyang Technological University
[26]MiroMind AI [27]University of Indonesia [28]University of New Haven
[29]INTI International University and Colleges [30]Binus University
[31]National University Philippines [32]ThoughtFull

Abstract

While the field of vision-language (VL) has achieved remarkable success in integrating visual and textual information across multiple languages and domains, there is no dedicated framework for assessing human-centric alignment in vision-language systems. We offer two contributions to address this gap. First, we introduce **Anthropogenic Regional Adaptation**: a novel paradigm that aims to optimize model relevance to specific regional contexts while ensuring the retention of global generalization capabilities. Second, we present a simple, but effective adaptation method named Geographical-generalization-made-easy (GG-EZ), which utilizes regional data filtering and model merging. Through comprehensive ablations on 3 VL architectures: large vision-language models (LVLM), text-to-image diffusion models, and vision-language embedding models, and a case study in Southeast Asia (SEA) regional adaptation, we demonstrate the importance of Anthropogenic Regional Adaptation and GG-EZ, showing 5-15% gains in cultural relevance metrics across SEA while maintaining over 98% of global performance and even occasionally surpassing it. Our findings establish Anthropogenic Regional Alignment as a foundational paradigm towards applicability of multimodal VL architectures in diverse regions and demonstrate a simple-yet-effective baseline method that optimizes regional value alignment while preserving global generalization.[1]

1 Introduction

Representation alignment in underrepresented regional contexts of foundation models has been a longstanding problem in AI (Jain et al., 2024; Winata et al., 2024; Anugraha et al., 2025), limiting their effectiveness and applicability across

[1]All datasets and models can be found here.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 695–712
November 6-10, 2026 ©2026 Association for Computational Linguistics

diverse global populations. This challenge is particularly acute in under-developed and developing regions – such as African (Adelani et al., 2021; Wang et al., 2024a), Indian (Khan et al., 2024; Verma et al., 2025), Middle Eastern, Southeast Asian (SEA) (Cahyawijaya et al., 2023a; Lovenia et al., 2024; Cahyawijaya et al., 2025b; Winata et al., 2023), etc – where existing resources exhibit severe underrepresentation. Building on this critical gap, we observe that existing vision-language (VL) AI solutions, despite their capabilities to generalize across different languages and domains, often exhibit cultural insensitivity (Naous et al., 2024; Patel et al., 2025; Cahyawijaya et al., 2025a), stereotypical outputs (Liu et al., 2025; Cahyawijaya et al., 2025c), and reduced task performance when deployed in underrepresented regions (Kabra et al., 2023; Cecilia Liu et al., 2024; Agarwal et al., 2025). This misalignment stems from training data dominated by certain regions with limited exposure to diverse contextual nuances prevalent in other regions (Cahyawijaya et al., 2023b; Cahyawijaya, 2024; Adilazuarda et al., 2024), undermining their real-world applicability.

To bridge this divide, we propose Anthropogenic Regional Adaptation, a foundational paradigm designed to systematically evaluate human-centric alignment of VL models across regional contexts while preserving global generalization capabilities. Within this framework, we identify two primary model archetypes: (1) global models that achieve strong performance across broad geographic contexts but struggle with underrepresented regions, and (2) regionally specialized models that excel on locale-specific metrics, yet falter when confronted with broader global contexts. Neither archetype is ideal, creating a critical gap in the utilization of existing systems for targeted regional applications.

Overcoming this limitation, we specifically designed a simple yet effective method, Geographical-generalization-made-easy (GG-EZ), which adapts an existing global model to a regional-specific context with minimal degradation on the global context. Inspired by recent advancements in training strategies of large language models (LLMs) (Cohere et al., 2025; Kocmi et al., 2025; Zhang et al., 2025; Guo et al., 2025), GG-EZ operationalizes regional adaptation through a two-level approach: (1) regional data filtering to curate culturally relevant training subsets, and (2) model merging to integrate region-specific adaptations without catastrophic forgetting of global knowledge.

We validate Anthropogenic Regional Adaptation and GG-EZ through rigorous experiments across three VL architectures—large vision-language models (Gemma-3 27B), text-to-image diffusion models (SDXL), and vision-language embedding models (SigLIP-2) – focusing on SEA as a case study. Our results demonstrate that GG-EZ achieves 5-15% improvements in cultural relevance metrics, including cultural context accuracy and local visual context understanding, while retaining over 98% of global performance on standard benchmarks. These findings establish Anthropogenic Regional Adaptation as a critical step toward equitable vision-language solutions and position GG-EZ as a practical baseline for anthropogenic regional adaptation in vision-language systems.

2 Related Work

Regional-specific models have been reported to outperform one-size-fits-all models on regional context (Nguyen et al., 2024). Several datasets (Lovenia et al., 2024; Cahyawijaya et al., 2025b) and language models (Ng et al., 2025; Nguyen et al., 2026) have been developed for the Southeast Asia (SEA) region. However, in computer vision, work on regional adaptation has focused more on remote sensing (Nyborg et al., 2022; Wang et al., 2025), which results in limited understanding of a region's unique cultural and anthropomorphic characteristics. Cahyawijaya et al. (Cahyawijaya et al., 2025b) report that existing multilingual vision-language models such as MAYA-8B (Alam et al., 2024), PaliGemma-2-10B (Steiner, 2024), Pangea-7B(Yue et al., 2024), Qwen2-VL-7B (Wang et al., 2024b), Gemma-3 (Team et al., 2025) fail in generating culturally-relevant responses when benchmarked against regional-specific cultural benchmarks such as SEAVQA (Urailertprasert et al., 2024), CVQA (Mogrovejo et al., 2024) and World Cuisines (Winata et al., 2025). On the other hand, recent development of regional-specific vision-language models such as VIOLET (Mohamed et al., 2023) and Baseer (Hennara et al., 2025), VARCO-VISION (Ju et al., 2024), and SEA-LION-VL (Singapore, 2024), have strong performance on regional-specific benchmark, but fall short on a broader and more general global context. Motivated by these limitations, we present a generic framework to systematically assess the gap of anthropogenic regional alignment in different models and introduce a simple-yet-effective approach, GG-

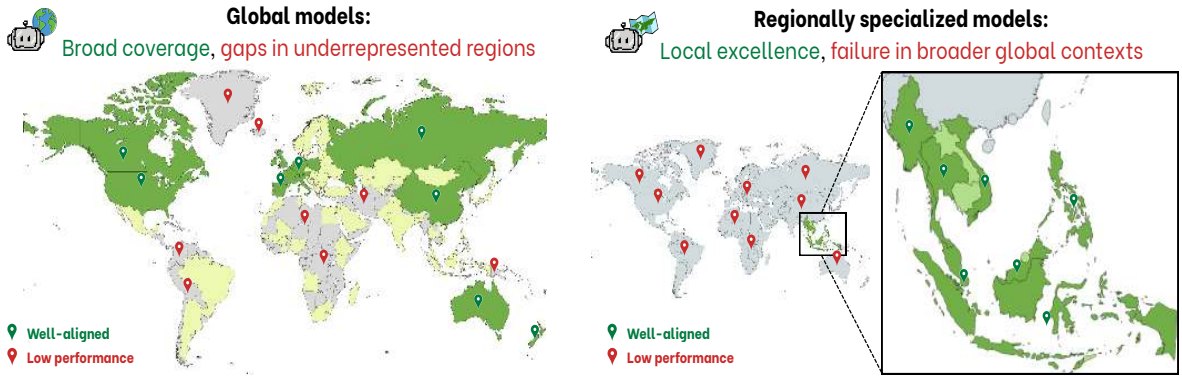


Figure 1: Through anthropogenic regional adaptation, we identify two primary model archetypes: **(left)** Global model with strong overall global performance, but struggle to represent certain regions appropriately, and **(right)** Regional-specific model that has a strong representation towards certain regions, but fall short on the global context.

EZ, to reduce the regional gap of existing multimodal vision-language models, while maintaining its performance on the broader global contexts.[2]

3 Anthropogenic Regional Adaptation

3.1 Overview

Let $\mathcal{R}^{\text{glob}}$ denote the comprehensive spatial domain representing our global context of interest. Within this universal space, we establish a discrete partitioning into meaningful geographical or conceptual units through the definition of $\mathcal{R}$, a finite collection of k distinct regions. Formally, $\mathcal{R} = \{R_1, R_2, \ldots, R_k\}$ where each individual region R_i constitutes a measurable subset of $\mathcal{R}^{\text{glob}}$. This decomposition enables systematic analysis of adaptation strategies across heterogeneous spatial contexts, recognizing that the world is rarely homogeneous in its characteristics or response patterns to anthropogenic influences.

Building upon this foundational regional partitioning, we introduce a critical distinction between regions based on strategic importance. The target region $\mathcal{R}^{\text{reg}}$ is defined as a specific, non-empty subset of $\mathcal{R}^{\text{glob}}$ representing our primary focus region where adaptation efforts should be maximized. Complementarily, the remaining regions are collectively designated as $\mathcal{R}^{\text{others}}$, defined through set difference as $\mathcal{R}^{\text{others}} = \mathcal{R}^{\text{glob}} \setminus \mathcal{R}^{\text{reg}}$. This ensures complete domain coverage while maintaining clear analytical separation between priority and non-priority regions, with each point in $\mathcal{R}^{\text{glob}}$ belonging to exactly one region.

Building upon this theoretical framework, as depicted in Figure 1, existing vision-language models predominantly follow two archetypal approaches. Global models capture broad, universal patterns across $\mathcal{R}^{\text{glob}}$ but sacrifice nuanced regional specificity. Conversely, regional-specific models focus on particular $\mathcal{R}^{\text{reg}}$ subsets with high local performance but reduced global coherence. This dichotomy reflects a fundamental trade-off in model design between universal applicability and specialized excellence. Most existing models, commit to one of the paradigm: global optimization for comprehensive coverage or regional specialization for enhanced regional performance.

3.2 Global-Regional Parity (GRP) Optimization

For each region R_i, we have a set of evaluation metrics $E^{R_i} = \{e_1^{R_i}, e_2^{R_i}, \ldots, e_m^{R_i}\}$. We define a quality metric $q_j^{R_i}$ for each evaluation metric $e_j^{R_i}$, and the collection of these quality metrics forms the set $Q^{R_i} = \{q_1^{R_i}, q_2^{R_i}, \ldots, q_m^{R_i}\}$. This multi-dimensional approach captures the complexity of regional characteristics influencing adaptation outcomes. The evaluation framework extends to our partitioned structure, yielding aggregated evaluation sets $Q^{\mathcal{R}^{\text{glob}}}$ and $Q^{\mathcal{R}^{\text{reg}}}$ that consolidate metrics across regions while maintaining connections to individual characteristics. To balance competing objectives, we introduce a globalization factor $\alpha \in [0, 1]$ that explicitly controls the trade-off between optimizing the target region and maintaining global generalization. The formal optimization objective is constructed as a weighted combination:

$$\max_{\theta} \left[\alpha \cdot Q^{\mathcal{R}^{\text{glob}}} + (1 - \alpha) \cdot Q^{\mathcal{R}^{\text{reg}}} \right]$$

This formulation creates a single scalar objective

[2]We provide comparison of our SEA-VLM models with other models in Appendix D.

that can be optimized using standard techniques, with α serves as a critical control mechanism that can be adjusted based on the specific focus and characteristics of each region. For regions with strong global connections or dependencies, α can be set closer to 1 to emphasize maintenance of $\mathcal{R}^{\text{glob}}$ conditions, while for regions where local impacts dominate, α can approach 0 to prioritize $\mathcal{R}^{\text{reg}}$ optimization. The solution achieves "best parity" between regions, representing the most favorable equilibrium where neither aspect is disproportionately favored according to our specified weighting scheme.

4 Geographical Generalization Made Easy (GG-EZ)

Geographical-generalization-made-easy (GG-EZ) is a general framework for regional adaptation of global multimodal models that achieves effective region-specific performance while preserving original generalization capabilities. As illustrated in Figure 2, the core mechanism of GG-EZ consists of two phases: **regional quality filtering** and **global-regional refinement**. GG-EZ not only enables adaptation across local textual and imagery contexts while maintaining base model capabilities, but also better yet being architecture-agnostic, allowing effective regional adaptation across different vision-language architectures.

4.1 Regional Quality Filtering

The regional quality filtering phase refines the training dataset to focus on regionally relevant examples while preserving general knowledge. Let $D = \{D_{\text{reg}}, D_{\text{gen}}\}$ represent the complete dataset, where D_{reg} contains region-specific examples and D_{gen} contains general-domain examples.

Regional Filter (F_{rf}): We define a boolean regional filter function that selects examples from the target region:

$$F_{\text{rf}}(x, r) = \begin{cases} 1 & \text{if region}(x) = r \\ 0 & \text{otherwise} \end{cases}$$

where r denotes the target region and $\text{region}(x)$ returns the geographical region associated with example x.

Multilingual Reward (F_{rm}): To further refine data filtering, we employ reward models that scores data quality and relevance on a continuous scale:

$$F_{\text{rm}}(x) = \theta_{rwd}(x)$$

where θ_{rwd} is the weight of the reward model and $\theta_{rwd}(x)$ measure the quality of the data x using the reward model θ_{rwd}. The filtered regional-specific dataset is then obtained by applying a threshold τ:

$$D_{\text{filt}} = \{x \in D_{\text{reg}} \mid F_{\text{rf}}(x, r) = 1 \text{ and } F_{\text{rm}}(x) > \tau\}$$

Language Translation Augmentation: Beyond regional filtering of existing cultural-relevant corpora, a further data enrichment is done through translating English high-quality datasets to multiple target regional languages. Let T represent the translation function mapping English to regional languages L_r:

$$D_{\text{trans}} = \{T(x, \text{English} \rightarrow l) \mid x \in D_{\text{filt}}, l \in L_r\}$$

4.2 Global-Regional Refinement

Using a pre-trained global model and the resulting corpora from the regional quality filtering phase, two steps of global-regional refinements are incorporated: 1) supervised fine-tuning, which turns the global model into a highlt performant region-specific model, and 2) model merging to ensure the preservation of the original global generalization capabilities. We first trained the global model θ_{global} into a quality regional-specific model θ_{reg} using $D_{\text{sft}} \subseteq \{D_{\text{filt}} \cup D_{\text{trans}}\}$. After supervised fine-tuning on D_{sft}, a linear model merging technique is applied to combine the global model with the region-specific model. The merged model parameters θ_{merged} are defined as:

$$\theta_{\text{mrg}}(\beta) = \beta \cdot \theta_{\text{reg}} + (1 - \beta) \cdot \theta_{\text{glob}}$$

where β controls the interpolation between the region-specific and the original global model. For beta interpolation, we first select a certain value for the global-regional parity factor α (e.g., $\alpha = 0.65$) to balance region-specific performance and general capabilities. The optimal β value is selected based on evaluation metrics:

$$\beta^* = \underset{\alpha \in [0.0, 1.0]}{\text{argmax}} \left[\alpha \cdot Q^{\mathcal{R}^{\text{reg}}}_{\theta_{mrg}(\beta)} + (1 - \alpha) \cdot Q^{\mathcal{R}^{\text{glob}}}_{\theta_{mrg}(\beta)}\right]$$

where $Q^{\mathcal{R}^{\text{reg}}}_{\theta_{\text{mrg}}(\beta)}$ measures regional-specific performance from the evaluation sets $E^{\mathcal{R}^{\text{reg}}}$ and $Q^{\mathcal{R}^{\text{glob}}}_{\theta_{\text{mrg}}(\beta)}$ measures global performance from the evaluation sets $E^{\mathcal{R}^{\text{glob}}}$. This approach ensures minimal degradation of base model capabilities while achieving strong regional adaptation.

5 Case Study on Southeast Asian (SEA) Adaptation

We explore GG-EZ on 3 distinct multimodal architectures: a large-scale Vision-Language Model

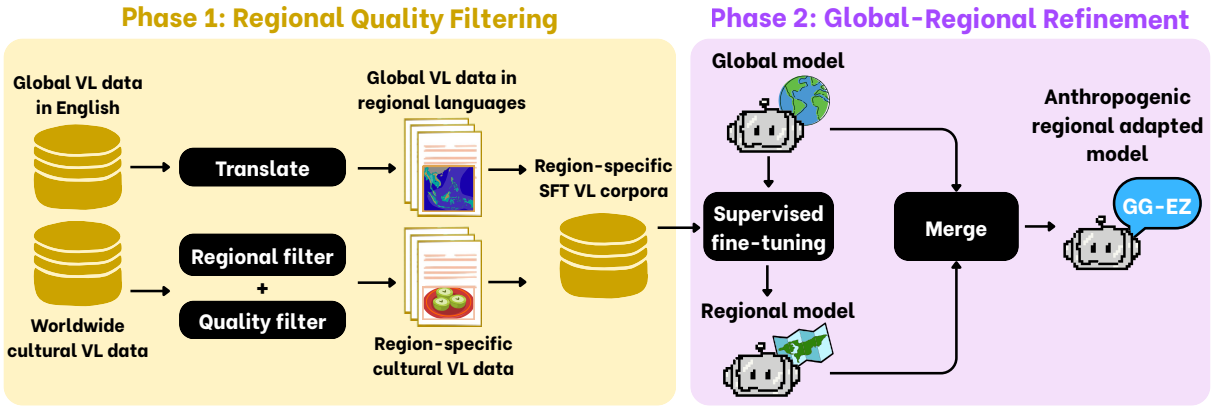


Figure 2: Overview of our Geographical-generalization-made-easy (GG-EZ) framework. Our framework consist of 3 constituents: (1) High-quality regional data filtering pipeline; (2) supervised fine-tuning to create a high quality regional-specific model; and (3) model merging to capture the best combination between regional-specific and global represention while also maintaining the generalization capabilities of the model.

Model	GRP	Global			SEA-Specific			
		Avg.	WC	CVQA	Avg.	SEAVQA	WC	CVQA
Google Gemma-3	59.4	63.5	59.8	67.2	56.3	41.0	60.1	67.8
SEA-Gemma-3 5%	64.0	64.3	**60.0**	68.7	63.7	61.2	**60.3**	**69.5**
SEA-Gemma-3 10%	**64.1**	**64.4**	**60.0**	**68.8**	**63.8**	**61.7**	60.2	**69.5**
SEA-Gemma-3 50%	57.3	56.7	51.6	61.8	57.8	59.5	51.4	62.6
SEA-Gemma-3 70%	56.1	56.3	51.9	60.6	56.0	54.0	52.6	61.3
SEA-Gemma-3 (w/o merge)	42.2	42.1	48.5	35.6	42.2	41.9	48.6	36.2

Table 1: SEA-VLM evaluation results on CVQA, SEA-VQA, and WorldCuisine (WC). <MODEL> (X%) denotes a merged model where X% is the weight β use in the linear merging process of the region-specific model.

Model	Overall			Language Breakdown				
	GRP	Global	SEA	fil	ind	tha	vie	zsm
Google Gemma-3	2.29	**2.54**	2.09	1.69	2.15	2.17	2.37	2.07
SEA-Gemma-3 10%	**2.31**	2.42	2.22	1.88	2.07	2.29	**2.61**	**2.25**
SEA-Gemma-3 (w/o merge)	1.74	1.18	**2.23**	**2.75**	**2.29**	**2.33**	1.76	2.00

Table 2: SEA-VLM human evaluation results. We report the average rank across different models (Higher is better). <MODEL> (X%) denotes a merge with X% as the weight β of the region-specific model.

(VLM) with 27B parameters, a contrastive Vision-Language Embedding Model (VL-Embed) with 1B parameters, and a Contextualized Diffusion Model (SEA-ImageGen) with 3.5B parameters. Our experiments focus on the SEA region, encompassing 11 countries—Singapore, Indonesia, Malaysia, Brunei, Thailand, Philippines, Vietnam, Myanmar, Cambodia, Laos, and East Timor—with a total population of approximately 700 million. In this section, we provide detailed configurations for applying GG-EZ and the evaluation strategy used to assess its effectiveness.

5.1 Data Curation

For data curation, we leveraged prefiltered SEA-specific content from SEA-VL, culturally relevant imagery from CulturalGround (Nyandwi et al., 2025), and translated instruction data from MAmmoTH-VL (Guo et al., 2024) (5 shards converted to major SEA languages, including Indonesian, Malaysian, Thai, Vietnamese, Filipino, Khmer, Lao, Chinese, and Tamil). To ensure high-quality translation on the underrepresented languages like Khmer and Lao, we ablate different translation models for each language and select the one with optimal quality and cost[3]. The data filtering process applied a regional filter to isolate SEA-specific examples, utilized the UnifiedReward (Wang et al., 2026) model for quality assessment with a threshold set at 3 or above[4], and combined regional and translated instruction data

[3] See Appendix A for more detail.

[4] We pick the best reward model based on our ablations on a human-annotated SEA-specific human preference dataset. The comparison of different reward model quality on the SEA regional test set is shown on Appendix B.

to create the final fine-tuning dataset. This dataset curation is crucial to enable adaptation of a general global model into a strong regional-specific model as explained in §6.2.

5.2 Regional Supervised Fine-tuning

All models underwent supervised fine-tuning using the AdamW optimizer (Loshchilov and Hutter, 2017) with linear learning rate decay. The SEA-adapted VLM (SEA-VLM) was trained from the Gemma-3 model as the initial checkpoint with a batch size of 64, learning rate of 2e-5, weight decay of 0.01, and 3 epochs. Before the supervised fine-tuning, we perform continuous pre-training on the Gemma-3 checkpoint using SEA-VL, XM3600, and Flickr30k, where we translated these datasets into SEA languages similar to Section 5.1. For the contextualized diffusion model (SEA-ImageGen), the model was fine-tuned from SDXL with a batch size of 32, learning rate of 1e-5, weight decay of 0.01, and 4 epochs. We only fine-tuned the UNet module the model while keeping the VAE module as is. For the VL embedding model (SEA-VL Embed), the model was fully fine-tuned from SigLIP2-SO400m with a batch size of 128, learning rate of 5e-6, and 2 epochs. The resulting regionally-adapted models are denoted as SEA-Gemma-3, SEA-SDXL, and SEA-SigLIP2 for the VLM, diffusion, and embedding architectures respectively.

5.3 Global-Regional Merging

After fine-tuning, we performed linear model merging to combine region-specific adaptations with the original base models. For VL-Embed and SEA-ImageGen, we explored interpolation weights β in the range [0.25, 0.5, 0.75], while for the SEA-VLM we tested a broader range of β values [0.05, 0.10, 0.5, 0.7]. The optimal β was determined by maximizing a weighted combination of regional and general performance metrics, where RegionAcc measures performance on SEA-specific validation data and GeneralAcc evaluates performance on broad-domain benchmarks. This formulation prioritizes regional adaptation while ensuring minimal degradation of foundational capabilities. We denote the resulting merged model as "<MODEL> (X%)" – e.g., SEA-Gemma-3 10% – where X% denotes the weight percentage β of the regionally-specific model used in the linear merging.

5.4 Evaluation

For the SEA-VLM model, evaluation included the SEAVQA regional visual question answering benchmark, CVQA for culturally nuanced queries, WorldCuisine for food identification tasks, and human evaluation to assess cultural relevance and accuracy.[5]. The SEA-VL Embed was evaluated on SEAVQA and CVQA to measure regional and cultural performance in embedding space. The SEA-ImageGen model underwent assessment using a DPGBench benchmark (Hu et al., 2024) as a proxy for the global quality of the models and diffusion model human evaluation set from SEA-VL (Cahyawijaya et al., 2025c) to gauge cultural relevance and visual quality on the SEA region. All evaluations were conducted on held-out test sets to ensure fair comparison across architectures and adaptation strategies. For human evaluation we use a 3-level likert scoring with native speaker annotators for each of the evaluated regional language. Finally, to compute a fair and practical global-regional parity (GRP) score from both of the global and region-specific evaluations, we set the globalization factor $\alpha = 0.43$[6] derived from 2023 KOF Globalization Index (Haelg, 2019; Gygli et al., 2019) which measures the level of globalization across 190 countries.

6 Result and Discussion

6.1 Overall Results

6.1.1 GG-EZ on SEA-VLM

We observe two major trends in SEA-Gemma-3 performance (Table 1). First, without model merging, SEA-Gemma-3 generally underperforms Gemma-3 on CVQA and World-Cuisine benchmarks, likely due to reduced task generalization after regional fine-tuning. However, it exceeds Gemma-3 on the SEA-specific SEA-VQA evaluation. This suggests that pure supervised fine-tuning for regional adaptation may not be optimal, as pre-trained VLMs already contain diverse data that can be compromised by smaller regional datasets. With model merging, performance improves significantly across all evaluations. A 50:50

[5]Due to limited evaluation resources for the region, we develop two human evaluation sets: (1) SEA-AYA a human translated AyaVisionBench (Dash et al., 2026); and (2) SEA-VL VQA, a human annotated VQA prompts from SEA imagery. Details in Appendix C.

[6]$\alpha = 0.43$ indicates a moderate tendency toward SEA-focused quality $Q^{\mathcal{R}^{\text{reg}}}$ with a fair amount of consideration for global quality $Q^{\mathcal{R}^{\text{glob}}}$.

Model	Overall Score	Aspects				
		Attribute	Relation	Entity	Other	Global
StabilityAI SDXL	73.75	79.18	86.38	81.07	60.00	83.89
SEA-SDXL 25%	**74.75**	80.43	86.34	**81.86**	61.60	85.41
SEA-SDXL 50%	74.61	**80.93**	**87.12**	81.15	60.00	**86.02**
SEA-SDXL 75%	74.39	79.98	85.96	81.36	63.20	82.67
SEA-SDXL (w/o merge)	74.32	80.79	86.92	81.57	**65.20**	81.16

Table 3: SEA-SDXL Performance on Image Generation Benchmark DPGBench (Higher is better). <MODEL> (X%) denotes a linear merge where X% is the weight β of the region-specific model.

Model	Correctness				Naturalness			
	Overall	T	L	C	Overall	T	L	C
StabilityAI SDXL	1.491	1.470	1.636	1.387	1.675	1.436	2.023	1.613
SEA-SDXL 25%	**1.569**	**1.587**	**1.729**	**1.413**	**1.767**	**1.473**	**2.124**	**1.753**
SEA-SDXL (w/o merge)	1.431	1.473	1.527	1.307	1.557	1.340	1.806	1.560

Table 4: SEA-SDXL human evaluation results on 3 distinct cultural aspects. **"T"** denotes Tradition, **"L"** denotes Landmark, and **"C"** denotes Cuisine

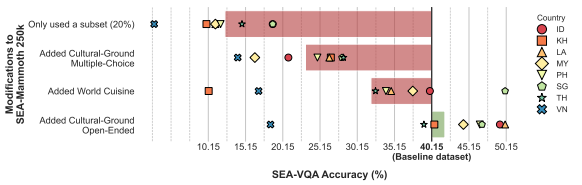


Figure 3: Impact of regional-specific data curation strategy on SEA-Gemma-3.

mix (SEA-Gemma-3 50%) achieves a turning point where global performance slightly decreases while SEA-specific performance surpasses Gemma-3. Increasing the merging weight to 10% (SEA-Gemma-3 10%) further improves results, outperforming Gemma-3 on all evaluation sets with a 1% global performance gain and 7.5% SEA-specific improvement, yielding the highest GRP. This demonstrates GG-EZ's ability to create regionally-adapted models strong in both specificity and generalization.

Human evaluation (Table 2) confirms these trends. SEA-Gemma-3 (w/o merge) shows superior regional specialization with the best average ranks across Southeast Asian languages: Filipino (2.75), Indonesian (2.29), Thai (2.33), and SEA region (2.23). Gemma-3 maintains global dominance (2.54), while SEA-Gemma-3 10% achieves the best Vietnamese (2.61) and Malaysian (2.25) scores with minimal global degradation (2.42 vs. 2.54), yielding the highest GRP of 2.31. This confirms that even modest SEA integration provides measurable regional benefits (Appendix E and F).

6.1.2 GG-EZ on SEA-ImageGen

The SEA-SDXL models demonstrate strong regional adaptation while maintaining robust global performance on DPGBench (Table 3). The fully fine-tuned SEA-SDXL achieves 74.32 overall, slightly improving upon SDXL's 73.75. GG-EZ's linear merging further enhances performance, with SEA-SDXL 25% and 50% models reaching 74.75 and 74.61 respectively, surpassing both the general model and regional-specific SEA-SDXL.

Human evaluation results (Table 4) confirm the merging strategy's effectiveness for regional context. The SEA-SDXL 25% model outperforms both SDXL and fully fine-tuned SEA-SDXL across all cultural aspects (Tradition, Landmark, Cuisine) in both correctness and naturalness metrics. This indicates the merging process successfully recovers generalization capability lost during regional specialization, enabling the models to maintain or enhance global performance while excelling in regional context evaluations.[7]

6.1.3 GG-EZ on SEA-VL Embed

The SEA-SigLIP2 model outperforms the original Google SigLIP2 on both SEA and non-SEA regions in CVQA and SEA-VQA evaluations (Table 5). Merging SEA-SigLIP2 with the original SigLIP2 further enhances performance. The SEA-SigLIP2 50% model achieves the highest global CVQA score (27.52) and strongest non-SEA performance (27.97), excelling in general knowledge transfer. The SEA-SigLIP2 75% model demonstrates superior regional specialization with the highest SEA-VQA overall score (29.66), peak performance for Indonesia (30.05) and Vietnam (28.75), and maintains strong global performance (27.12), yielding the highest GRP score (27.96) compared to the original SigLIP2 (25.17) and fine-tuned SEA-SigLIP2

[7]GRP score was not measured for image generation due to lack of comparable test sets, though the model clearly improves both regional-specific and general image generation scores.

Model	GRP	CVQA			SEAVQA			
		Global	SEA	Others	SEA	Khm	Ind	Vie
Google SigLIP2	25.17	25.51	24.02	25.84	25.81	28.95	25.00	25.56
SEA-SigLIP2 25%	24.96	24.38	24.44	24.36	26.36	27.63	25.40	23.96
SEA-SigLIP2 50%	27.10	**27.52**	25.50	**27.97**	28.06	**31.25**	26.99	27.16
SEA-SigLIP2 75%	**27.96**	27.12	**27.51**	27.03	**29.66**	28.62	**30.05**	**28.75**
SEA-SigLIP2 (w/o merge)	26.96	26.75	25.29	27.07	28.96	28.29	28.59	29.07

Table 5: SEA-VL Embedding Zero-Shot Performance on Cultural VQA Benchmarks. <MODEL> (X%) denotes a linear merge where X% is the weight β of the region-specific model. For SEA-VQA, we provide additional breakdown of 3 of the biggest country data in the dataset with "Khm"=Cambodia, "Ind"=Indonesia, "Vie"=Vietnam.

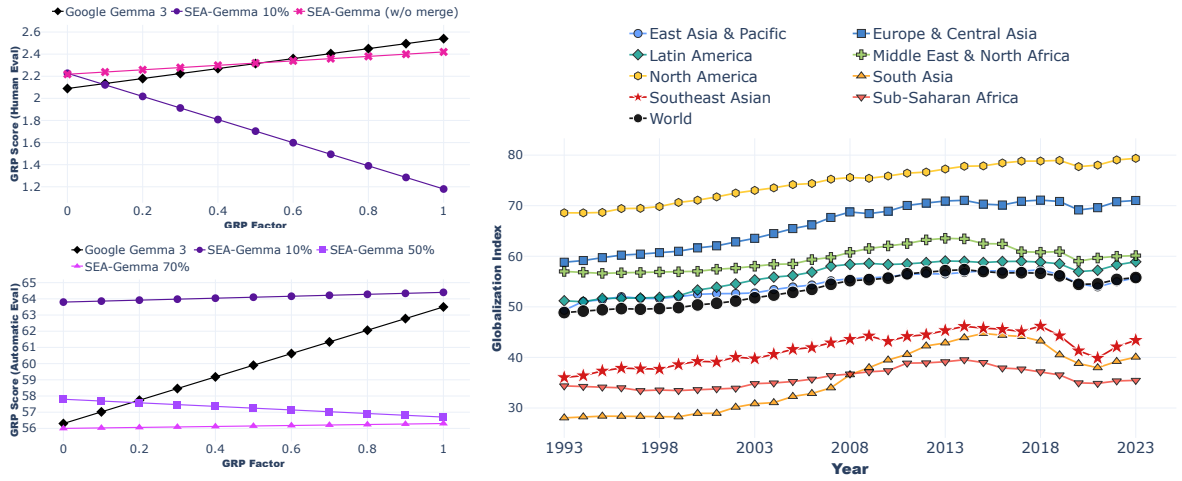


Figure 4: **(left)** Impact of globalization factor $alpha$ to GRP across different models. Optimizing on a misaligned α can lead to suboptimal performance. **(right)** We derive α from the KOF globalization index (Haelg, 2019; Gygli et al., 2019) to better reflect the degree of globalization across regions. The globalization index is distinct across regions and evolves over time.

(26.96). The merging approach effectively balances regional adaptation and generalization, preserving and enhancing SigLIP2's broad capabilities while incorporating specialized SEA knowledge.[8]

6.2 Impact of Regional Quality Filtering

Data quality and relevance critically impact region-specific model adaptation performance (Cahyawijaya et al., 2023b, 2024). SEA-Gemma-3 ablation studies show naive augmentation does not guarantee gains; outcomes depend on data structure, scope, and provenance. As in Figure 3, data volume is a bottleneck: a 20% subset of the SEA-Mammoth 250k dataset (translated to 10 languages) causes a 70% performance drop vs. full data, stressing the need for sufficient scale.

Integrating the SEA-filtered Cultural-Ground dataset (Nyandwi et al., 2025) shows performance depends on VQA format. Open-ended VQA boosts regional understanding to 41.9% (baseline) but is sensitive to prompts; multiple-choice degrades to 21.6%. The SEA-filtered WorldCuisine dataset (Winata et al., 2024) reduces performance by 42%. We hypothesize its regional cuisine focus narrows the model, lacking comprehensive cultural knowledge (e.g., landmarks, history, clothing). Thus, strategic curation is vital; indiscriminate dataset addition without format- and domain-aware evaluation harms generalizability.

6.3 Globalization Factor in Anthropogenic Regional Adaptation

The globalization factor α governs GRP optimization for Anthropogenic Regional Adaptation, determining how models should balance global and regional characteristics. An appropriate α reflects actual regional interaction patterns, whether societies prefer unified global perspectives or localized regional views. Misalignment causes undesirable model behavior (Figure 4, left). We derive α from the KOF-GI (Konjunkturforschungsstelle Globalization Index) (Haelg, 2019; Gygli et al., 2019), using the de facto interpersonal component that

[8] Detailed evaluation breakdowns are in Appendix E.

measures person-to-person globalization intensity (Martens and Raza, 2009; Axel et al., 2010; Figge and Martens, 2014). This provides a standardized, quantifiable metric by calculating KOF-GI expectations across target region countries $\mathcal{R}^{\text{reg}}$.

The KOF-GI-based α offers two key advantages: 1) dynamic adaptation as globalization evolves over time (Figure 4, right), and 2) an objective, external metric reducing subjectivity in regional importance determination. This establishes a fair global-regional parity baseline that bridges the gap between purely global models (lacking regional specificity) and purely regional models (lacking global coherence) (Figure 1). Models optimized with KOF-GI-derived α adaptively balance focus: emphasizing regional performance in less globalized areas while maintaining global consistency in highly interconnected regions. This creates a system that simultaneously satisfies universal applicability and specialized excellence demands, enabling robust performance across heterogeneous regional contexts.[9]

7 Conclusion

We propose the Anthropogenic Regional Adaptation framework to address cultural insensitivity and reduced performance in underrepresented regions, establishing a foundational paradigm for human-centric alignment in vision-language systems. Our GG-EZ approach implements this framework, demonstrating 5-15% gains in cultural relevance metrics across SEA contexts (Table 1) while maintaining or improving generalization capabilities across three architectures—Gemma-3 27B, SDXL, and SigLIP-2—enabling equitable deployment without sacrificing core performance. This work highlights GRP's importance in human-centric alignment, providing data-driven and dynamically adjusted optimization.

8 Limitation

Our study demonstrates promising gains in regional cultural relevance, but several constraints limit the breadth and depth of the findings. First, the experiments are confined to relatively modest model sizes (Gemma-3 27B, SigLIP-2, and SDXL), which may not capture the full spectrum of performance improvements achievable with larger architectures; scaling to billion-plus parameter models could reveal different trade-offs between regional adaptation and global generalization. Second, while we target eleven Southeast Asian languages, the coverage is uneven: major languages such as Indonesian, Thai, and Vietnamese receive substantially more data and evaluation than smaller languages like Khmer and Lao, potentially biasing the results and leaving gaps in truly low-resource linguistic contexts. Third, the exploration of model merging strategies is limited to linear interpolation with a few discrete weight points; more sophisticated merging techniques remain underexplored, and could yield better parity or higher regional gains without sacrificing global performance.

[9]For detailed KOF-GI breakdown, see Appendix G.

9 Ethical Statement

All human annotations were conducted on a voluntary basis by the authors and members of relevant regional communities, with fair compensation provided for their contributions. Annotations underwent rigorous double-checking to ensure quality, safety, and cultural relevance, adhering to ethical guidelines for data collection in underrepresented regions. The evaluation design incorporated a fair scoring system to minimize bias, and annotators were selected from diverse linguistic and cultural backgrounds within Southeast Asia to ensure contextual accuracy. No sensitive or private data was used, and all processes aligned with principles of informed consent and equitable participation. During writing of the paper, we also use AI assistant to help with paraphrasing and post-editing.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'Souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Shruti Rijhwani, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, Tao Sheng, Sujith Ravi, and Dan Roth. 2025. Aligning LLMs for multilingual consistency in enterprise applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language*

Processing: Industry Track, pages 117–137, Suzhou (China). Association for Computational Linguistics.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, S M Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth. S, Snehanshu Mukherjee, and Alham Fikri Aji. 2024. Maya: An instruction finetuned multilingual multimodal model. *Preprint*, arXiv:2412.07112.

David Anugraha, Patrick Amadeus Irawan, Anshul Singh, En-Shiun Annie Lee, and Genta Indra Winata. 2025. M4-rag: A massive-scale multilingual multi-cultural multimodal rag. *arXiv preprint arXiv:2512.05959*.

Dreher Axel, Gaston Noel, Martens Pim, and Van Boxem Lotte. 2010. Measuring globalization opening the black box. a critical analysis of globalization indices. *Journal of Globalization Studies*, 1(1):166–185.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Samuel Cahyawijaya. 2024. Llm for everyone: Representing the underrepresented in large language models. *Preprint*, arXiv:2409.13897.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025a. High-dimension human value representation in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5330, Albuquerque, New Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, and 29 others. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, Dea Annisayanti Putri, Bryan Wilie, Genta Indra Winata, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025b. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, and 1 others. 2025c. Sea-vl: A multicultural vision-language dataset for southeast asia. *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 18685–18717.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vladyslav Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Harvey Richemond, Acyr Locatelli, Nicholas Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2026. Aya vision: Advancing the frontier of multilingual multimodality.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Lukas Figge and Pim Martens. 2014. Globalisation continues: The maastricht globalisation index revisited and updated. *Globalizations*, 11(6):875–893.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale.

Savina Gygli, Florian Haelg, Niklas Potrafke, and Jan-Egbert Sturm. 2019. The kof globalisation index – revisited. *The Review of International Organizations*, 14(3):543–574.

Florian Haelg. 2019. The kof globalisation index – a multidimensional approach to globalisation. *Jahrbücher für Nationalökonomie und Statistik*, 240(5):691–696.

Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Baseer: A vision-language model for arabic document-to-markdown ocr. *Preprint*, arXiv:2509.18174.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *Preprint*, arXiv:2403.05135.

Anupam Jain and 1 others. 2024. Ai4bharat indicbert: A monolingual bert model for indian languages. pages 10685–10706.

Jeongho Ju, Daeyoung Kim, SunYoung Park, and Youngjune Kim. 2024. Varco-vision: Expanding frontiers in korean vision-language models. *Preprint*, arXiv:2411.19103.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Fahim Khan and 1 others. 2024. Ai4bharat indicbert: A monolingual bert model for indian languages. pages 10685–10706.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation*, pages 789–799, Suzhou, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Cohere Labs. 2025. Aya vision benchmark.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Pim Martens and Mohsin Raza. 2009. *The maastricht globalisation index: An update*, pages 279–309. Nova Science Publishers, United States.

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, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for Arabic image captioning with gemini decoder. In *Proceedings of ArabicNLP 2023*, pages 1–11, Singapore (Hybrid). Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Huang Yuli, Tai Ngee Chia, Leong Wai Yi, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Tan Choon Meng, Brandon Ong, and 11 others. 2025. SEA-LION: Southeast Asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 512–526, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Tan Sang Nguyen, Muhammad Reza Qorib, and Hwee Tou Ng. 2026. Openseal: Good, fast, and cheap construction of an open-source southeast asian llm via parallel data. *Preprint*, arXiv:2602.02266.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of ACL*, pages 294–304.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.

Joachim Nyborg, Charlotte Pelletier, Sébastien Lefèvre, and Ira Assent. 2022. Timematch: Unsupervised cross-region adaptation by temporal shift estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:301–313.

Hitesh Laxmichand Patel, Amit Agarwal, Arion Das, Bhargava Kumar, Srikant Panda, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, and Dong-Kyu Chae. 2025. Sweeval: Do llms really swear? a safety benchmark for testing limits for enterprise use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 558–582.

AI Singapore. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Andreas et al. Steiner. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 24 others. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 170 others. 2025. Gemma 3 technical report. arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. arXiv:2207.04672.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khanuja, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. *arXiv preprint arXiv:2411.02538*.

Hengbin Wang, Yu Yao, Junyi Liu, Xindan Zhang, Yuanyuan Zhao, Shaoming Li, Zhe Liu, Xiaodong Zhang, and Yelu Zeng. 2025. Unsupervised cross-regional and cross-year adaptation by climate indicator discrepancy for crop classification. *Journal of Remote Sensing*, 5.

Jiayi Wang, David Ifeoluwa Adelani, and 1 others. 2024a. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. pages 5997–6023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2026. Unified reward model for multimodal understanding and generation. *Preprint*, arXiv:2503.05236.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264.

Genta Indra Winata and 1 others. 2024. Worldcuisines: A benchmark dataset for multilingual and multicultural image classification. *arXiv preprint arXiv:2405.14133*.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *Preprint*, arXiv:2306.09341.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, and 3 others. 2026. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *Preprint*, arXiv:2412.21059.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

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

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

A Assessment for SEA languages Translation Quality

Language	Model	Grammar	Naturalness
Burmese (mya)	NLLB	4.6 ± 0.8	4.1 ± 1.1
	Gemma-3-27b	4.1 ± 1.2	3.6 ± 1.2
	Gemini-2.5-flash	4.8 ± 0.6	4.6 ± 0.8
	Gemini-2.5-pro	$\mathbf{4.9 \pm 0.4}$	$\mathbf{4.9 \pm 0.5}$
Thai (tha)	NLLB	3.3 ± 1.1	3.0 ± 1.2
	Gemma-3-27b	4.2 ± 0.9	4.1 ± 0.9
	Gemini-2.5-flash	$\mathbf{4.8 \pm 0.5}$	$\mathbf{4.7 \pm 0.5}$
	Gemini-2.5-pro	4.7 ± 0.5	4.6 ± 0.5
Filipino (fil)	NLLB	4.4 ± 0.6	4.3 ± 0.9
	Gemma-3-27b	$\mathbf{4.8 \pm 0.6}$	$\mathbf{4.6 \pm 0.8}$
	Gemini-2.5-flash	4.7 ± 0.4	4.4 ± 0.7
	Gemini-2.5-pro	$\mathbf{4.8 \pm 0.4}$	4.5 ± 0.6
Indonesian (ind)	NLLB	4.3 ± 0.9	3.8 ± 1.1
	Gemma-3-27b	$\mathbf{4.6 \pm 0.6}$	4.2 ± 0.9
	Gemini-2.5-flash	4.5 ± 0.6	4.1 ± 0.9
	Gemini-2.5-pro	$\mathbf{4.6 \pm 0.5}$	$\mathbf{4.3 \pm 0.8}$
Vietnamese (vie)	NLLB	4.7 ± 0.5	4.4 ± 0.8
	Gemma-3-27b	$\mathbf{5 \pm 0.1}$	4.9 ± 0.3
	Gemini-2.5-flash	$\mathbf{5 \pm 0.1}$	4.9 ± 0.2
	Gemini-2.5-pro	$\mathbf{5 \pm 0.1}$	$\mathbf{5 \pm 0.1}$

Table 6: Human evaluation of English→5 SEA languages translation quality using a five-point scale (mean ± standard deviation)

We conducted a systematic evaluation on 100 randomly selected human-generated English captions from the SEA-VL (Cahyawijaya et al., 2025b) dataset for Burmese (mya), Thai (tha), Filipino (fil), Indonesian (ind), and Vietnamese (vie). The evaluation employed a five-point scale (1–5) assessing grammatical accuracy and naturalness, using NLLB (Team et al., 2022), Gemma-3-27b (Team et al., 2025), Gemini-2.5-flash, and Gemini-2.5-pro (Team et al., 2023) models. Table 6 presents the results of the translation quality evaluation in five Southeast Asian languages using a five-point scale (mean ± standard deviation). Overall, the Gemini-2.5 models consistently achieve the highest scores for Burmese (mya) and Thai (tha), with Gemini-2.5-Pro obtaining the best performance for Burmese and Gemini-2.5-Flash performing slightly better for Thai. For Filipino (fil), Indonesian (ind), and Vietnamese (vie), Gemma-3-27B achieves competitive or top performance, often matching or exceeding the Gemini models in grammatical accuracy while maintaining high naturalness scores. Based on these findings, Gemma-3-27b was used for Indonesian (ind), Vietnamese (vie), Standard Malay (zsm), Filipino (fil), and Chinese (zho), where it demonstrated strong and stable performance. Gemini-2.5-flash was used for Thai (tha), Burmese (mya), Lao (lao), Khmer (khm), and Tamil (tam), as the Gemini family showed superior robustness and higher translation quality for these languages.

B Reward Model Ablation

We evaluated four reward models: HPSv2 (Xu et al., 2023), HPSv2 (Wu et al., 2023), VisionReward-Image (Xu et al., 2026), and UnifiedReward (Wang et al., 2026). In order to pick a reward model to use for data curation as described in section 5.1, we measured how much the reward models agreed with human evaluations from SEA-VL (Cahyawijaya et al., 2025c). For each image I_n, we normalized the scores via min-max normalization and then took the average across categories to get a human annotator score $s^n_{human} \in [0, 1]$. Notably, the absolute score range of ImageReward-style models varies across datasets, making direct normalization and comparison with human ground-truth scores unreliable. Therefore, instead of comparing absolute scores, we reformulated the evaluation as a pairwise preference task. Specifically, We randomly sampled 500 pairs of images I_a and I_b such that $s^a_{human} > s^b_{human}$. For each reward model, we then compute the rate at which it gives I_a a higher score than I_b i.e. $s^a_{rm} > s^b_{rm}$. Results are summarized in table 7. Although VisionReward-Image achieved the highest pairwise agreement with human judgments, we ultimately selected UnifiedReward by considering both computational efficiency and predictive performance. While its accuracy is slightly lower, UnifiedReward offers a more favorable trade-off overall, as it supports efficient batched inference through vLLM (Kwon et al., 2023), substantially improving throughput for large-scale data filtering.

Reward Model	Rate where $s^a_{rm} > s^b_{rm}$
ImageReward	0.394
HPSv2	0.384
VisionReward-Image	**0.466**
UnifiedReward	0.442

Table 7: Rate at which $s^a_{rm} > s^b_{rm}$ for each reward model under consideration.

C Annotated Human Evaluation Test Sets

We conduct human evaluation on two datasets: SEA-VL VQA and SEA AYA.

C.1 SEA-VL VQA

SEA-VL VQA contains ∼1.1k visual questions spanning 9 official languages in Southeast Asia: Malay (zsm), Vietnamese (vie), Thai (tha), Indonesian (ind), Filipino (fil), Tamil (tam), Khmer (khm), Chinese (cmn), and Burmese (mya). The culturally grounded images were originally collected via crowdsourcing by (Cahyawijaya et al., 2025c). For each image, a native speaker authored a visual question in the target language, which was subsequently reviewed by two additional native speakers, with revisions made when needed to ensure quality.

To promote consistency and cultural relevance, we provided annotators with detailed instructions for writing high-quality visual questions. Specifically, questions were required to: (1) be directly grounded in the visual content of the image; (2) go beyond low-level perception and instead target culturally salient elements, such as traditional attire, local cuisine, religious practices, architectural styles, or social activities; (3) be clear, grammatically correct, and contextually appropriate; and (4) be answerable solely from the image.

We also specified several categories of questions to avoid. These included: (1) overly generic questions (e.g., "What is this?" or "What color is the shirt?"); (2) purely factual questions lacking cultural grounding (e.g., "How many people are in the image?"); (3) questions that cannot be answered from the image alone (e.g., "What is her name?" unless explicitly shown); (4) questions that fail to reflect the Southeast Asian context without clear justification; and (5) questions that can be answered without the visual context.

C.2 SEA AYA

We construct SEA AYA by manually translating 135 visual questions from the Aya Vision Benchmark (Labs, 2025) into 6 official Southeast Asian languages: Thai (tha), Malay (zsm), Filipino (fil), Tamil (tam), Chinese (cmn), and Burmese (mya). Combined with the English (eng) and Indonesian (ind) subsets already available in Aya Vision Benchmark, the resulting benchmark contains ∼1.2k instances covering 9 vision-language tasks: image captioning, chart and figure understanding, image difference detection, general VQA, OCR, document understanding, text transcription, visual reasoning, and screenshot-to-code generation. For each example, annotators were first given a rough machine-translated draft as a starting point. They were then asked to revise it to ensure that the final translation is fluent, culturally appropriate, and semantically faithful to the English source.

For quality and consistency, our annotation guidelines emphasized several principles: (1) preserve the meaning, tone, and intent of the original English prompt; (2) produce text that is natural and fluent for native speakers of the target language; (3) use grammar, vocabulary, and syntax that conform to standard usage in the target language; (4) match the level of formality of the source; (5) respect language-specific cultural and pragmatic norms, including idiomatic usage and politeness conventions; and (6) appropriately translate, retain, or normalize named entities and technical terms when necessary.

We also requested annotators to avoid common translation errors, including semantic drift, omission of essential information, overly literal adherence to machine-translated phrasing, literal translation of idiomatic expressions, over-translation or under-translation, awkward or unidiomatic wording, and inconsistent terminology, particularly across multi-sentence examples.

D Comparison of SEA-Gemma-3 and Other VLMs

We provide the full comparison per language on CVQA and per region on SEA-VQA in Table 8 and Table 9, respectively. The merged global-regional model SEA-Gemma-3 10% achieves strongest performance on SEA regions and other regions on CVQA, producing a solid performance even hgher compared to other smaller and similar sized models like Llama-3.2 Vision (Grattafiori et al., 2024), Pixtral (Agrawal et al., 2024), Gemma-3 (Team et al., 2025), and even outperform larger models such as Aya-Vision (32B) (Dash et al., 2026), Qwen-3 VL (35B) (Bai et al., 2025a), Qwen-2.5-VL (72B) (Bai et al., 2025b), and Molmo (72B) (Deitke et al., 2024). Similarly on SEA-VQA, SEA-Gemma-3 10% also achieves the best average regional score, although on several regions it is still outperformed by large models especially Aya-Vision (32B) (Dash et al., 2026) which showcase strong performance on multiple SEA regions but falls short on Indonesia, Thailand, and Laos.

E Results Breakdown

We provide the results breakdown in Table 10, Table 12, Table 11, and Table 15 for SEA-VLM; Ta-

Language	Llama-3.2 Vision (11B)	Pixtral (12B)	Gemma-3 (27B)	SEA-Gemma-3 (27B)	SEA-Gemma-3 10% (27B)	Aya-Vision (32B)	Qwen-3 VL (35B)	Qwen-2.5-VL (72B)	Molmo (72B)
Indonesian	56.31	62.86	66.50	21.60	66.26	67.88	66.26	66.50	63.83
Filipino	51.72	64.53	66.01	24.14	70.94	66.34	64.04	65.02	64.53
Malay	56.19	61.90	67.62	29.84	71.43	72.38	70.16	62.50	69.84
Chinese	62.80	68.93	72.66	40.73	75.14	77.48	78.28	80.98	79.29
Tamil	58.41	51.87	73.36	40.65	74.30	61.68	61.68	58.88	58.41
Javanese	47.81	51.18	62.96	22.22	63.64	55.56	54.55	55.22	54.88
Minangkabau	51.79	51.39	64.54	23.90	67.73	61.75	54.98	51.79	58.17
Sundanese	44.00	49.00	59.00	12.50	63.50	52.53	55.50	56.50	52.00
SEA Avg.	53.63	57.71	66.58	26.95	69.12	64.45	63.18	62.17	62.62
Amharic	39.32	32.91	62.82	26.07	62.39	29.18	42.74	36.48	45.30
Bengali	55.59	48.25	72.03	33.57	75.87	64.31	64.34	61.27	68.88
Breton	34.57	35.80	42.96	27.65	45.68	39.36	34.81	37.78	35.06
Bulgarian	49.06	22.91	62.26	34.23	64.42	56.49	54.45	61.99	57.68
Arabic	49.26	43.35	62.07	32.02	65.02	68.47	57.64	61.08	58.62
Hindi	68.16	30.85	80.60	42.79	81.09	78.11	75.62	75.12	78.11
Igbo	44.00	41.50	43.50	25.50	47.00	38.00	34.50	36.55	41.50
Irish	53.99	57.67	64.11	31.29	65.95	56.13	55.52	57.98	57.06
Japanese	50.74	49.26	53.20	29.06	53.69	59.11	58.62	58.62	57.14
Kinyarwanda	35.32	34.47	53.62	28.51	53.19	40.43	32.77	38.30	40.43
Korean	59.66	73.45	76.90	42.76	77.93	80.00	77.24	77.59	74.14
Marathi	48.02	31.19	74.75	34.16	77.23	66.17	63.86	61.39	68.81
Mongolian	39.42	39.74	50.32	23.40	50.64	36.01	44.23	39.10	47.76
Norwegian	54.52	64.21	66.89	38.80	69.57	66.22	61.20	68.56	69.90
Oromo	34.11	35.51	45.79	24.77	47.20	36.45	38.32	35.05	42.06
Portuguese	57.75	73.59	80.63	41.20	80.63	78.01	76.06	76.76	77.46
Romanian	58.94	67.88	75.17	40.73	76.49	74.17	70.20	75.83	70.20
Russian	66.50	37.00	80.50	45.50	81.50	80.00	76.50	79.00	84.00
Sinhala	48.00	28.44	65.33	29.78	65.78	39.56	43.56	45.50	45.78
Spanish	57.02	69.13	69.34	37.61	69.92	73.25	68.20	71.44	71.55
Swahili	53.11	60.07	76.19	36.26	80.95	66.18	60.07	55.31	67.77
Telugu	55.50	32.50	71.50	32.00	74.50	57.79	61.00	58.50	57.00
Urdu	55.98	48.02	73.39	36.93	75.69	66.69	66.99	64.92	69.50
Other Avg.	50.81	45.99	65.39	33.68	67.06	58.70	57.32	58.01	60.25
Global Avg.	51.53	49.01	65.69	31.94	67.59	60.18	58.83	59.08	60.86

Table 8: Comparison of our best model (SEA-Gemma-3 10%) againsts other competitor models and the regionally-adapted Gemma-3 (SEA-Gemma-3) model on CVQA.

Region	Gemma-3 (27B)	SEA-Gemma-3 (27B)	SEA-Gemma-3 10% (27B)	Aya-Vision (32B)	Qwen-3 VL (35B)
Cambodia	50.7	40.5	52.3	**56.25**	50.99
Indonesia	43.5	49.3	**69.0**	66.09	66.09
Laos	0.0	50.0	**68.1**	45.83	63.89
Malaysia	37.5	44.4	62.4	**69.31**	60.32
Philippines	63.4	46.6	58.8	**63.40**	56.21
Singapore	56.3	46.9	68.8	**71.88**	65.63
Thailand	51.6	39.1	52.2	49.46	**52.72**
Vietnam	18.9	18.5	57.5	**63.26**	53.04
SEA Avg.	40.2	41.9	**61.1**	60.68	58.61

Table 9: Comparison of our best model (SEA-Gemma-3 10%) againsts other competitor models and the regionally-adapted Gemma-3 (SEA-Gemma-3) model on SEA-VQA.

Model	Avg. Score	Country							
		Camb.	Indo.	Laos	Mala.	Phil.	Sing.	Thai.	Viet.
Google Gemma-3	41.0	50.7	43.5	0.0	37.5	**63.4**	56.3	51.6	18.9
SEA-Gemma-3 5%	61.2	50.3	68.8	66.7	**62.4**	59.5	65.6	51.1	**57.8**
SEA-Gemma-3 10%	**61.7**	**52.3**	**69.0**	**68.1**	**62.4**	58.8	**68.8**	52.2	57.5
SEA-Gemma-3 50%	59.5	52.0	67.3	62.5	57.1	56.9	53.1	**52.7**	55.0
SEA-Gemma-3 70%	54.0	51.3	67.3	63.9	56.6	58.2	53.1	51.6	20.5
SEA-Gemma-3 (w/o merge)	41.9	40.5	49.3	50.0	44.4	46.6	46.9	39.1	18.5

Table 10: SEA-VLM Zero-shot performance on SEA-VQA. The result is broken down per country. "Camb." = Cambodia, "Indo." = Indonesia, "Mala." = Malaysia, 'Phil.' = Phillipines, "Sing."=Singapore, "Thai." = Thailand, "Viet." = Vietname

Model	Overall Score	SEA Avg.	By Prompt Type			
			Task 1 (NC)	Task 2	Task 1 (C)	Task 1 (Adv.)
Google Gemma-3	59.8	60.1	**64.8**	**48.3**	78.5	47.5
SEA-Gemma-3 5%	**60.0**	**60.3**	64.5	48.1	**79.2**	**48.1**
SEA-Gemma-3 10%	**60.0**	60.2	**64.5**	48.0	79.1	**48.1**
SEA-Gemma-3 50%	51.6	51.4	53.7	42.0	67.0	43.6
SEA-Gemma-3 70%	51.9	52.6	53.5	43.5	67.1	43.4
SEA-Gemma-3 (w/o merge)	48.5	48.6	49.3	43.5	61.9	39.4

Table 11: SEA-VLM Zero-shot performance on World Cuisine. (NC) = No Context, (C) = Context, (Adv.) = Adversarial

ble 13 for SEA-ImageGen, and Table 14 fo SEA-VL Embedding.

F Sample Outputs of SEA-VLM and SEA-ImageGen

We provide samples generated by our best SEA-VLM and SEA-ImageGen compared to the baselines on Figure 5 and Figure 6, respectively. As shown in Figure 6, the SEA-SDXL 25% model produces the most accurate representation among

Model	Overall SEA		Prompt Type	
	Score	Avg.	English	Local Lang.
Gemma-3	67.2	67.83	67.96	66.44
SEA-Gemma-3 5%	68.72	**69.45**	69.39	68.05
SEA-Gemma-3 10%	**68.76**	**69.45**	**69.42**	**68.09**
SEA-Gemma-3 50%	61.84	62.62	62.9	60.79
SEA-Gemma-3 70%	60.58	61.27	61.51	59.66
SEA-Gemma-3 (w/o merge)	35.58	36.16	38.0	33.16

Table 12: SEA-VLM Zero-shot performance on CVQA.

Model	Attribute					Entity			Other		Relation	
	color	other	shape	size	texture	part	state	whole	count	text	non-spatial	spatial
StabilityAI SDXL	80.88	79.49	69.87	67.36	79.98	78.52	74.23	82.71	56.00	76.00	81.13	86.73
SEA-SDXL 25%	82.34	81.49	75.11	63.22	80.82	82.81	73.39	83.45	59.50	70.00	77.36	86.93
SEA-SDXL 50%	82.39	80.49	73.80	66.53	82.52	80.08	74.13	82.67	58.50	66.00	76.73	87.80
SEA-SDXL 75%	81.59	79.49	72.49	67.36	81.19	79.49	73.71	83.09	61.50	70.00	81.76	86.23
SEA-SDXL 100%	82.14	81.26	77.73	64.46	81.73	79.69	74.45	83.20	62.00	78.00	84.91	87.06

Table 13: Detailed per aspect breakdown (Attribute, Entity, Other, and Relation) of the DPGBench evaluation.

Model	SEAVQA Zero-Shot Performance								
	Overall	Cambodia	Indonesia	Laos	Malaysia	Phillipines	Singapore	Thailand	Vietnam
Google SigLIP2	25.81	28.9	25.0	22.2	24.3	28.8	21.9	25.5	25.6
SEA-SigLIP2 25%	26.36	27.6	25.4	29.2	29.1	28.8	28.1	26.1	24.0
SEA-SigLIP2 50%	28.06	31.3	27.0	22.2	29.6	27.5	21.9	31.0	27.2
SEA-SigLIP2 75%	29.66	28.6	30.1	19.4	32.3	34.0	25.0	29.9	28.8
SEA-SigLIP2 100%	28.96	28.3	28.6	29.2	26.5	32.7	31.3	30.4	29.1

Table 14: Detailed per country zero-shot performance of the SEAVQA evaluation.

Model	Avg. Score	Country							
		Khm	Ind	Lao	Mys	Phl	Sgp	Tha	Vie
SEA-Mammoth 50k	13.0	9.9	18.8	11.1	11.1	11.8	18.8	14.67	2.88
SEA-Mammoth 250k	39.6	13.8	**52.9**	**52.8**	42.9	**47.1**	46.9	**44.0**	**20.8**
SEA-Mammoth 250k + CG OE	**41.9**	**40.5**	49.3	50.0	**44.4**	46.6	46.9	39.1	18.5
SEA-Mammoth 250k + CG MC	21.6	26.6	20.9	26.4	16.4	24.8	28.1	28.3	14.1
SEA-Mammoth 250k + WC	30.4	10.2	39.9	34.7	37.6	34.0	**50.0**	32.6	16.9

Table 15: SEA-VLM Zero-shot performance on SEA-VQA. The result is broken down per country. "Khm" = Cambodia, "Lao" = Laos, "Ind" = Indonesia, "Mys" = Malaysia, 'Phl' = Phillipines, "Sgp"=Singapore, "Th." = Thailand, "Vie" = Vietnam. "CG" denotes Cultural-Ground, "WC" denotes WorldCuisine, while "OE" and "MC" denote open-ended and multiple-choice, respectively.

others, while retaining the image naturalness and overall quality of the original SDXL. For example, while the baselines also capture the component of nasi lemask such as nasi putih (white rice) and telur (egg), the SEA-SDXL 25% model incorporates more detailed factors including timun (cucumber), kacang (peanut), ikan bilis (anchovy). Similarly for cultural and tradition categories, SEA-SDXL 25% captures a more accurate illustration of playing angklung and Hung king festival in Phu Tho, Vietnam.

G Globalization Index by Region by Year

We provide the detailed per year per region globalization index which we used for deciding the globalization factor α in GG-EZ in Table 16.

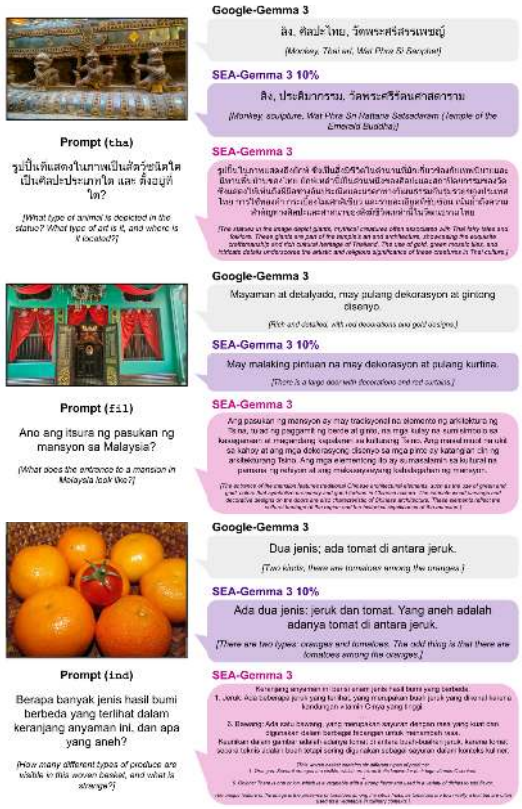


Figure 5: Generated responses using different model architypes. From left-to-right: global model (Gemma-3), our regional model (SEA-Gemma-3), our merged model (SEA-Gemma-3 10%) along with the prompts. Our model produces the most correct image among others, while retaining the image naturalness and overall quality of the original Gemma-3.



Figure 6: Generated image using different model architypes. From left-to-right: global model (SDXL), our regional model (SEA-SDXL), our merged model (SEA-SDXL 25%), and reference of natural images.

H Compute Resources

For all of our experiments, we a singel H100 compute node with 8 H100 GPUs. The biggest computes come from running CPT and SFT of VLM models which takes use of the full compute for

Year	EAP	ECA	LAC	MENA	NA	SA	SEA	SSA	World
1993	49.32	58.82	51.20	57.00	68.59	28.05	36.04	34.37	48.86
1994	51.07	59.15	51.00	56.85	68.58	28.25	36.40	34.26	49.15
1995	51.40	59.72	51.67	56.62	68.66	28.39	37.34	34.13	49.42
1996	52.02	60.23	51.67	56.74	69.43	28.37	37.89	33.99	49.64
1997	51.68	60.43	51.75	56.79	69.48	28.35	37.76	33.43	49.53
1998	51.63	60.71	51.86	56.89	69.83	28.33	37.70	33.55	49.66
1999	52.00	61.00	52.20	56.94	70.64	28.26	38.59	33.45	49.85
2000	52.48	61.65	53.33	57.03	71.09	28.94	39.23	33.63	50.38
2001	52.61	62.06	53.88	57.43	71.74	28.96	39.12	33.80	50.70
2002	52.63	62.83	54.53	57.66	72.50	30.16	40.06	33.92	51.15
2003	52.75	63.55	55.32	58.04	73.02	30.84	39.76	34.82	51.79
2004	53.41	64.48	55.88	58.42	73.52	31.06	40.63	34.95	52.33
2005	53.93	65.47	56.19	58.48	74.17	32.29	41.61	35.24	52.86
2006	54.32	66.19	56.86	59.34	74.39	32.87	41.96	35.70	53.46
2007	55.16	67.67	58.04	59.78	75.24	33.99	42.90	36.40	54.46
2008	55.55	68.76	58.41	60.78	75.57	36.58	43.59	36.72	55.17
2009	55.70	68.42	58.59	61.54	75.41	37.98	44.27	37.16	55.38
2010	55.94	68.92	58.37	62.05	75.87	39.49	43.19	37.37	55.68
2011	56.40	70.04	58.50	62.51	76.43	40.59	44.17	38.87	56.53
2012	56.55	70.52	58.75	63.24	76.64	42.27	44.50	38.92	56.88
2013	56.50	70.93	59.06	63.55	77.25	42.87	45.33	39.14	57.16
2014	56.77	71.07	59.00	63.44	77.80	43.90	46.15	39.54	57.36
2015	57.13	70.32	58.76	62.50	77.86	44.75	45.79	38.99	56.98
2016	57.09	70.13	58.97	62.44	78.44	44.38	45.58	37.91	56.70
2017	57.04	70.88	59.00	60.91	78.81	44.14	45.12	37.71	56.75
2018	57.34	71.10	58.85	60.79	78.82	43.25	46.21	37.15	56.65
2019	56.36	70.84	58.51	60.92	78.95	40.53	44.29	36.54	56.12
2020	54.54	69.16	56.97	59.02	77.72	38.80	41.33	34.97	54.45
2021	53.94	69.60	57.23	59.66	78.02	37.97	39.87	34.86	54.54
2022	54.96	70.78	58.30	60.00	79.04	39.20	42.11	35.36	55.42
2023	55.74	71.03	58.91	60.13	79.36	40.06	43.40	35.46	55.79

Table 16: Globalization index by region by yea from 1993 to 2023. “EAP” = East Asia and Pacific , ‘ECA” = Europe and Central Asia,“LAC” = Latin America and Caribbean, “MENA” Middle East and North Africa, “NA” = North America, “SA” = South Asia, “SEA” = Southeast Asian, “SSA” = Sub-Saharan Africa.

several days per experiment. while other experiments, i.e., SEA-ImageGen and SEA-VL Embed, uses only 2 H100 GPUs for each experiment.