

Culturally-Aware Alignment for Multimodal Models via Group Reward Optimization for Southeast Asia

Longyin Zhang[†], Shuo Sun, Yingxu He,
Won Cheng Yi Lewis, Bowei Zou, and Ai Ti Aw
Institute of Advanced Intelligence and Computing, A*STAR, Singapore
zhang_longyin@a-star.edu.sg

Abstract

Multimodal large language models (MLLMs) perform well across many tasks but often exhibit Western bias, limiting their effectiveness in regions such as Southeast Asia (SEA). In this work, we study how both evaluation and model design can be better adapted to the linguistic and cultural characteristics of this region. We present the **SEA-Omni Benchmark Suite**, which brings together a set of SEA-focused image and video datasets capturing phenomena such as intra-sentential code-switching and culturally grounded multimodal cues. Alongside these benchmarks, we develop **MERaLiON-Omni-10B**, an omni-perception model trained with a progressive, data-efficient pipeline that moves from vision-language alignment to more advanced multimodal fusion. To better account for region-specific cultural signals, we study text-mediated cultural alignment in SEA multimodal settings using GRPO to surface culturally distinctive responses that are often suppressed by generic decoding strategies. We introduce a fact-anchored **dual-reward mechanism** to balance cultural richness with multimodal consistency in which cultural density is measured via localized knowledge entities of generated texts, and multimodal coherence is assessed against silver textual descriptions of the multimodal input. Experimental results indicate that the resulting model performs strongly on SEA benchmarks while maintaining reasonable general capabilities.

1 Introduction

The rapid development of Multimodal Large Language Models (MLLMs) aims to perceive, transcribe, and reason by bridging the modality gap between perception and language (Zhu et al., 2023; Bai et al., 2023; Yin et al., 2024; Caffagni et al.,

[†]This work was conducted at the Institute for Infocomm Research (I2R), A*STAR, which is now known as the Institute for Advanced Intelligence and Computing (IAIC).

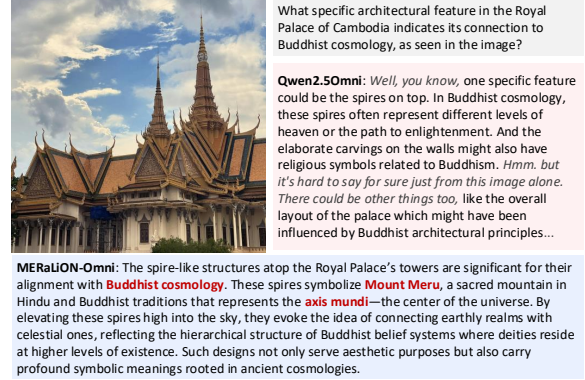


Figure 1: Culturally-aware regional reasoning. Compared with the size-matched Qwen2.5Omni model, which tends to produce generic responses, MERaLiON-Omni retrieves region-specific cultural concepts (e.g., Mount Meru) under the proposed alignment setting.

2024; Chen et al., 2025; Xu et al., 2025b). However, current MLLMs exhibit a significant cultural and geographical bias, often rendering them less effective in diverse global contexts (Pouget et al., 2024; Kadiyala et al., 2025). This limitation is especially pronounced in Southeast Asia (SEA), which comprises a diverse set of languages, cultural practices, and socio-visual environments. Existing models often struggle with region-specific concepts, multilingual interactions involving code-switching, and contextual cues embedded in local images and videos, limiting their effectiveness for a large population worldwide (Liu et al., 2025). Importantly, the challenge in SEA is not merely linguistic but multimodal. Many cultural signals in the region are conveyed through visual and acoustic channels, such as the appearance of local cuisine (e.g., “Kuih”), traditional attire, or the acoustic patterns associated with intra-sentential code-switching (e.g., “Singlish” expressions that combine English with Malay, Chinese, or Tamil vocabulary) (Ong and Limkonchotiwat, 2023; Nguyen et al., 2024). Without access to aligned visual and

audio context, these cues are difficult to interpret reliably. As a result, general-purpose MLLMs frequently misinterpret SEA visual inputs by projecting Western-centric concepts, leading to culturally inconsistent or hallucinatory outputs (Ramaswamy et al., 2023; Shumailov et al., 2024).

To address these challenges, we propose the following technical and data innovations:

- We present the SEA-Omni Benchmark Suite, an omni-perception benchmark suite specifically designed for the SEA context. It features authentic, region-specific content and is annotated to handle multiple languages and prevalent code-switching. As a foundational step, it currently targets five languages: English, Chinese, Malay, Indonesian, and Thai.
- We present MERaLiON-Omni-10B, a specialized MLLM engineered for omni-perception and text generation via a progressive training strategy. This work uncovers a key geometric property: visual alignment occurs within the orthogonal subspace of audio parameters, supporting the integration of cultural visual cues while retaining strong audio performance on many tasks.
- We study how culturally-aware post-training interacts with multimodal regional specialization in Southeast Asia. To support this investigation, we introduce a text-mediated dual-reward alignment framework based on Group Relative Policy Optimization (GRPO) that balances localized cultural expression with factual consistency against reference descriptions. The resulting model achieves leading performance on our curated regional benchmarks while retaining reasonable general-domain capabilities.

2 Related Work

Multilingual foundation models and regional adaptation. Recent work in large language models (LLMs) has increasingly turned from high-resource languages to multilingual and region-specific modeling. A comprehensive survey (Qin et al., 2025) of multilingual LLMs argues that while polyglot models exist, challenges remain in data imbalance, tokenization inefficiencies (Petrov et al., 2023), and alignment across languages. In the Southeast Asian region specifically, the SeaLion (Ong and Limkonchotiwat, 2023) and SeaLLMs (Zhang et al., 2025) series present great efforts in covering over a dozen languages (English,

Chinese, Indonesian, Vietnamese, Thai, Tagalog, Malay, Khmer, Lao, etc.), using instruction-tuning and culturally-aware safety mechanisms. Similarly, the CompassLLM model is oriented to the SEA market and emphasizes curriculum-style multilingual pre-training followed by Direct Preference Optimization fine-tuning to enhance low-resource language performance (Maria, 2024). These efforts reveal key insights that are relevant to our work: (i) region-specific vocabulary matters and (ii) multilingual instruction data must reflect local cultural and linguistic phenomena.

The landscape of SEA-centric multimodal models.

The transition from uni-modal language models to multimodal and ultimately omni-modal systems is now well documented (Han et al., 2024; Xu et al., 2025a,b). Within this space, more recent work has aimed to build unified backbones across modalities and to extend reasoning and instruction-following across modality boundaries (Yin et al., 2024). However, most existing multimodal and omni-modal work remains focused on high-resource languages and modalities, with less attention to region-specific multilingual and multimodal integration. When relating to the SEA context, although the majority of SEA-centric work has focused on text (Ong and Limkonchotiwat, 2023; Zhang et al., 2025), there are notable efforts to integrate speech and other modalities in the region. For example, He et al. (2025) explore building multimodal (speech and text) foundation models for SEA languages and dialects. Most recently, Cahyawijaya et al. (2025) present SEA-VL, a large-scale, open-source image-language dataset tailored for Southeast Asia. They collected around 1.28 million culturally relevant image-caption pairs via three methods: crowdsourcing, web crawling, and synthetic generation. This work fills a critical gap in under-representation of SEA cultures in vision-language research and offers a scalable framework for inclusive dataset construction. Nevertheless, existing region-targeted models remain text-centric or speech-text only, and few yet integrate image, video, audio, and culturally-rich contextual signals in one unified model. This motivates our exploration toward an omni model for the SEA region.

Culturally-aware training. Beyond regional language coverage, recent work directly targets cultural bias in model training. CulFiT (Feng et al., 2025) synthesizes diverse culturally relevant ques-

Task	Category & Fine-grained Focus	SEA-Specific Example
Captioning	General Description (Cultural Context)	A vendor selling Satay at a busy Pasar Malam .
Summarization	Contextual Summary (Background Knowledge)	Locals gathered at a Kopitiam , drinking Kopi-O , ...
Open-ended QA	Attribute Knowledge (Color, Shape, Material, Texture)	Is the Songkok (cap) black or dark blue?
	Spatial Relations (Absolute Position, Relative Position)	Is the Sambal to the left or right of the rice?
	Counting & Existence (Counting, Existence Judgment)	How many Tuk-tuks are visible in the traffic?
	Actions & States (Action Recognition, State Judgment)	What traditional dance is the performer doing?
	Scene Understanding (Time, Weather, Location, Environment)	Is the atmosphere at this Wet Market chaotic?
SEA-OCR (synthetic overlay images)	Multilingual Text (Latin&Non-Latin Scripts)	Image with: อาหารริมทางในกรุงเทพฯ [Trans: Street food in Bangkok]
Multimodal MCQA	Reasoning (Discriminative Analysis)	Which festival is depicted? (A) Songkran (Correct) (B) Holi (C) Christmas (D) Chinese Lunar New Year

Table 1: Overview of tasks in the SEA-Omni Image Benchmark. Region-specific concepts with cultural relevance (e.g., *Songkok*) are highlighted.

tions, constructs critique data in culturally relevant languages, and employs fine-grained rewards that decompose cultural texts into verifiable knowledge units for interpretable evaluation, thereby improving cultural alignment in LLMs. Our work differs in two respects: (i) we operate on omni-modal inputs (image, video, audio, and text) rather than text alone, and (ii) our alignment is grounded in a textual description of the multimodal input; cultural density is rewarded only when it is coherent with fact-anchored descriptions of the visual input (Subsection 4.2), rather than relying on text-only critique signals.

3 The SEA-Omni Dataset

The **SEA-Omni Benchmark Suite** comprises two complementary components: (1) a multitask SEA image dataset designed for foundational training and broad-spectrum evaluation and (2) a manually curated video dataset focused on advanced spatiotemporal and audio-visual reasoning evaluation.

3.1 The SEA-Omni Image Dataset

The first component of our suite is the SEA-Omni Image Benchmark (**SEA-Image**), a large-scale collection for robust visual evaluation, built upon the culturally-rich repository of SEA-VL (Cahyawijaya et al., 2025). On this basis, we propose a systematic, semi-automated data curation pipeline for generating diverse, high-quality evaluation data at scale. Our pipeline employs a two-stage propose-

and-refine framework. On one hand, a specialized SEA visual analyzer¹ (SVLer), adapted from a SEA Audio-LLM (He et al., 2025) and the pre-trained Vision Transformer (Xu et al., 2025a), under the in-context-supervision of gold-standard image captions to generate candidate annotations grounded in regional context. On the other, a powerful teacher LLM acts as a refiner, systematically enhancing these candidates for logical consistency and factual accuracy, utilizing GPT-4o for the test set and Qwen3-235B (Yang et al., 2025) for training.

As shown in Table 1, the final dataset is structured into five carefully designed task categories:

- **Captioning.** We directly use the gold-standard image-caption pairs from the SEA-VL corpora.
- **Summary.** SVLer generates an initial summary, which is then post-edited by the teacher LLM for factual and logical consistency with the gold-standard caption as a strong reference.
- **QA.** We use the teacher LLM to generate questions from summaries. SVLer provides initial answers from visual context, and the teacher LLM performs a final strict filtering pass to remove pairs whose answers are uncertain or whose content is inconsistent with the ground-truth captions, finally retaining 58.52% of instances.

¹Sourced from WebVid-10M and Koala-36M-v1, the SVLer training data is constructed via knowledge distillation. We leverage Qwen2.5-Omni-7B to generate descriptive summaries for videos and frames, yielding a silver dataset that bootstraps SVLer’s general visual perception capabilities.

- **OCR.** A multilingual text recognition task designed to evaluate the model’s ability to decode region-specific text rendered within images, covering *English, Chinese, Malay, Indonesian, and Thai*. We overlay image captions as text onto the source images and apply varied visual transformations to increase OCR difficulty.
- **MCQA.** The teacher LLM converts the filtered QA pairs into a multiple-choice format, complete with plausible distractors to test discriminative reasoning.

We validate this pipeline’s fidelity through human evaluation by two expert annotators with extensive living experience in Southeast Asia. For the three semi-automatically generated tasks (Summary, QA, and MCQA), we observe excellent inter-rater reliability (ICC2 0.915) and an average quality score of 0.704. These results demonstrate that the generated annotations are both consistent and informative, establishing the dataset as a reliable benchmark. Although caption quality is treated as an upper bound, validation was image-grounded, given that SVLer conditions directly on the image (Appendix C). Beyond annotation fidelity, we also enforce strict source-level separation. As shown in Table 2, training data is drawn exclusively from the web-crawled SEA-VL subset, while evaluation relies solely on the crowdsourced subset to avoid any data overlap.

3.2 The SEA-Omni Video Dataset

To extend evaluation beyond static images, we introduce the SEA-Omni Video Benchmark (**SEA-Video-v1**), a manually curated collection of 3,142 video clips aimed at evaluating reasoning over dynamic multi-sensory inputs. The benchmark requires joint modeling of vision, audio, and text, and reflects linguistic diversity across five languages: English, Mandarin, Indonesian, Thai, and Malay. It also includes realistic instances of code-switching, which pose additional challenges for robust language understanding in SEA settings.

Our data annotation follows a *Human-AI-Human* workflow. Domain experts first produce video summaries, which are then used to prompt GPT-4o to generate culturally grounded questions. The questions are subsequently reviewed and refined by human annotators. To ensure annotation quality, we apply strict cross-verification procedures, resulting in a 90% agreement rate on the manual summaries. For evaluation, we enforce a 14.3% discard rate

Task	SEA-I _{train}	SEA-I _{test}	SEA-V _{test}
Captioning	940K	2,049	–
Summary	940K	2,049	675
Visual QA	940K	1,199	668
Audio QA	–	–	641
OCR	940K	1,885	–
MCQA	574K	1,175	1,158

Table 2: Statistics of our SEA-Omni benchmarks.

for MCQs to remove ambiguity, while adopting a *repair-over-reject* strategy for open-ended questions (3.3% discard). This method rectifies flawed but culturally significant samples to preserve long-tail scenarios. In particular, the MCQA experiment on Qwen3Omni-30B (Xu et al., 2025b) shows that text-only input achieves 46.03% (acc), compared to 83.07% with full multimodal inputs. This performance gap (37.04%) indicates that SEA-Video-v1 substantially benefits from multimodal evidence beyond text-only prompting.

4 Method

This work builds on MERaLiON-AudioLLM (He et al., 2025), a recent Audio-LLM built on Gemma-2-9B-IT (Gemma Team, 2024) and tailored for Southeast Asia.² To evolve this model into an omni architecture, we integrate a vision processing pipeline, resulting in the model we term MERaLiON-Omni. While recent native omni models achieve unity, they are prohibitively expensive for regional adaptation. We argue for a modular, data-efficient approach that leverages state-of-the-art encoders to construct a culturally adaptive omni-system without training from scratch.

As shown in Figure 2, the integrated model is capable of processing text, audio, and visual inputs. It employs a Vision Transformer Encoder (Dosovitskiy et al., 2021) and the pre-trained Whisper Speech Encoder (Radford et al., 2023) to independently extract feature representations from their respective modalities. These features are then projected into a shared embedding space through modality-specific adapters to align them with text representations. Finally, the aligned multimodal embeddings are processed by a Gemma-2-based decoder (Gemma Team, 2024), which autoregressively generates the text output. This unified 10.8B omni model serves as our backbone model.

²We use the updated checkpoint of MERaLiON-2-10B (He et al., 2025) on Hugging Face as our backbone model.

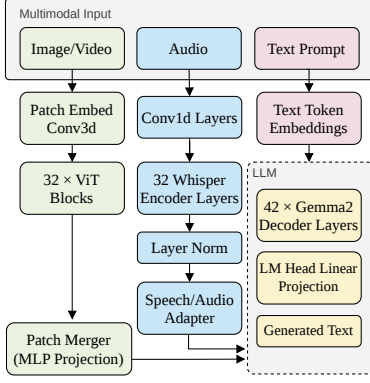


Figure 2: The unified architecture of MERaLiON-Omni.

4.1 Three-Stage Progressive Training

Let our model parameters be denoted by Θ , comprising the vision encoder and adapter parameters Θ_V , the audio encoder and adapter parameters Θ_A , and the LLM parameters Θ_L . Our training strategy optimizes subsets of these parameters in three distinct stages.

Efficient vision-language alignment. The initial stage aims to establish a semantic bridge between the vision and language modalities. We introduce low-rank adaptation (Hu et al., 2021) matrices $\Delta\Theta_V$ and $\Delta\Theta_L$ for the vision encoder and LLM, respectively. The training objective is to minimize the standard cross-entropy loss $\mathcal{L}_{CE}$ for next-token prediction on image-text pairs (v, t) :

$$\min_{\Delta\Theta_V, \Delta\Theta_L} \mathbb{E}_{(v,t) \sim \mathcal{D}_{g_img}} [\mathcal{L}_{CE}(t|v; \Theta_V + \Delta\Theta_V, \Theta_L + \Delta\Theta_L)], \quad (1)$$

where $\mathcal{D}_{g_img}$ refers to a general visual-language dataset. We analyze the geometric alignment between the sequential update vectors, $\Delta\Theta_{AL}$ ³ and $\Delta\Theta_L$. Despite a substantial parameter overlap of 93.86%, the updates exhibit near-perfect orthogonality with a global cosine similarity of around -10^{-4} . This checkpoint-difference analysis is consistent with near-orthogonal audio-adaptation and visual-alignment update directions in the shared LLM LoRA parameters under our setup. Together with the final-model audio results, it suggests partial retention of audio behavior after visual alignment, rather than establishing causal

³Let A , A' , A'' denote base (*gemma-2-9b-it*), audio-adapted (MERaLiON-2), and visually aligned checkpoints after LoRA merging. We compute two task vectors $\Delta\Theta_{AL} = A' - A$ and $\Delta\Theta_L = A'' - A'$, and measure cosine similarity only at parameter positions where both exhibit non-trivial changes ($> \epsilon$).

non-interference or lossless retention. To distinguish modality-specific structure from a high-dimensional baseline, we add a same-modality control: independently trained audio directions are positively correlated, whereas the original visual-audio direction is near zero and continued ASR training moves against the visual direction. See Appendix F for more details.

Deep unimodal visual knowledge ingestion. In the second stage, we focus on enriching the vision encoder with domain-specific knowledge. We merge the LoRA weights from Stage 1 into the base model and freeze all parameters except for the vision encoder and visual adapter, jointly denoted by Θ_V . The model is trained via full-parameter fine-tuning on the multi-task SEA-image dataset. Intuitively, this stage aims to force the vision encoder to become a domain expert on SEA-specific visual concepts by learning from large-scale SEA images while preserving the LLM’s original integrity. The objective is to optimize Θ_V to minimize the same cross-entropy loss:

$$\min_{\Theta_V} \mathbb{E}_{(v,t) \sim \mathcal{D}_{s_img}} [\mathcal{L}_{CE}(t|v; \Theta_V, \Theta_L)], \quad (2)$$

where $\mathcal{D}_{s_img}$ refers to our curated SEA-image visual-language training corpus in Subsec. 3.1.

Advanced multimodal fusion and reasoning. The final stage integrates all modalities, including audio and video. We re-introduce LoRA parameters for all components ($\Delta\Theta_V$, $\Delta\Theta_A$, and $\Delta\Theta_L$) to facilitate a holistic, yet efficient, end-to-end fine-tuning. The objective is as follows:

$$\min_{\Delta\Theta} \mathbb{E}_{(v,a,t) \sim \mathcal{D}_{mm}} [\mathcal{L}_{CE}(t|v, a; \Theta + \Delta\Theta)], \quad (3)$$

where $\mathcal{D}_{mm}$ consists of multimodal instruction-following fine-tuning data. We intentionally exclude SEA-specific videos from training to evaluate the model’s capability to generalize cultural knowledge from static images (Stage 2) to dynamic visual contexts. Consequently, our contributed SEA-Omni Video Benchmark serves as a strictly held-out test set to assess this zero-shot cultural transfer. In general, the three stages use 158K, 4.23M, and 41.6K training instances respectively (Appendix B).

4.2 Cultural Alignment via GRPO

We build upon Group Relative Policy Optimization (Shao et al., 2024) to study culturally-aware

post-training for SEA multimodal settings. To encourage localized concepts without unsupported content, we introduce a **dual-reward mechanism** that combines a cultural-density reward with a text-mediated coherence reward. The configuration uses batch-wise advantage normalization across rollout groups and a best-of-group positive-advantage safeguard.

Since reliable automatic metrics for cultural grounding from pixels or audio are unavailable, both rewards use textual proxies. The cultural density reward evaluates localized knowledge entities in generated texts, while the text-mediated multimodal coherence reward is judged by an LLM against SVLer summaries (Subsection 3.1; hereafter “DESC”). Our human validation supports the proxies: SVLer summaries achieve ICC 0.915 with human annotators (Appendix C), and the coherence judge reaches 95% agreement on 100 samples (Appendix K), suggesting that text-mediated rewards offer promising and cost-effective training signals for culturally situated alignment.

Reward formulation. Formally, given a prompt x and its corresponding response $y \sim \pi(\cdot|x)$, we formulate the composite reward function as:

$$R(y, x) = \lambda \rho_c(y) + (1 - \lambda) s_c(y, x). \quad (4)$$

(1) Cultural Density Reward— $\rho_c(y)$: We utilize Qwen3-235B (Yang et al., 2025) as an automated judge, prompting it to extract Localized Knowledge entities (LKs) from each response y as a measure of cultural relevance. To assess the reliability of this procedure, we manually annotate a validation corpus by identifying LKs from 1,024 SEA captions (Appendix J). Performance evaluation shows that Qwen3 delivers competitive performance (90.11-P, 86.23-R, 86.90-F1) against GPT-4o (89.38-P, 89.59-R, 89.11-F1). With these reliable LKs now available, we define the cultural density reward as:

$$\rho_c(y) = \frac{\min(|K_{LK}(y)|, K_{\max})}{K_{\max}}. \quad (5)$$

Here, $|K_{LK}(y)|$ denotes the number of unique localized terms, and $K_{\max} = 20$ serves as the empirical ceiling based on the training data distribution. This formulation aims to encourage cultural density while avoiding keyword stuffing.

(2) Multimodal Coherence Reward— $s_c(y, x)$: Standard Likert-scale scoring often yields high variance. Instead, we implement a *fact-anchored deductive scoring rubric* to ensure fidelity. For each

training sample, we use the silver summary generated by our SVLer pipeline (Subsection 3.1) as the DESC of the corresponding multimodal input. Here, DESC serves as the reference context against which the text-only judge (Qwen3-235B) evaluates the model’s response. Starting with a perfect score of $S_{full} = 5.0$, the judge applies deductive penalties based on five rigid criteria: (A) Common Sense Errors, (B) Factual Mistakes that contradict the reference DESC, (C) Hallucinations unsupported by the reference description, (D) Logical Inconsistency, and (E) Question-Answer Mismatch. The final score is normalized by: $s_c(y, x) = s_{judge}/S_{full}$.

Optimization objective. We employ the GRPO framework to maximize expected rewards subject to a KL penalty by minimizing the loss $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = -\mathbb{E} \left[\min(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon) A_t) \right] + \beta D_{KL}(\pi_\theta \parallel \pi_{\text{ref}}), \quad (6)$$

where $A_t = (R_t - \mu_B)/(\sigma_B + \xi)$ is the advantage computed from batch-level reward statistics μ_B and σ_B , and r_t denotes the probability ratio between the current policy and the old policy. Unlike the vanilla GRPO, our normalization strategy uses statistics computed across groups, which better emphasizes high-reward prompts and leads to improved performance (Table 11). In addition, we enforce a minimal positive advantage for the top-ranked candidate in each group. This acts as a dynamic baseline to prevent gradient vanishing, ensuring the reinforcement of locally best responses without training a separate critic model.

5 Experiments

Experimental setups, hyperparameters, and full details of our training data are presented in Appendices A and B.

5.1 Main Results

This subsection presents evaluation results on the SEA-Omni benchmarks. In Table 3, we begin with an ablation study on the initialization of the visual module, where “w/o Pretrain” denotes a baseline in which the visual module is trained from scratch. In contrast, MERaLiON-Omni-VL initializes the visual module by cloning parameters from Qwen2.5Omni ViT. The subsequent model variants, “-Image”, “-MM”, and “-GRPO”, are built on this pre-trained initialization.

Method	Captioning	QA		Summ	MCQA	OCR		
	BERTScore↑	BERTScore↑	F1↑	BERTScore↑	Acc↑	CER↓	WER↓	WER*↓
SeaLion4VL-27B (2025)	77.64	85.16	14.71	82.92	83.32	0.40	0.45	0.094
Qwen3Omni-30B (2025b)	78.90	84.63	14.51	82.86	86.21	0.29	0.31	0.102
Qwen2.5Omni-7B (2025a)	80.63	83.58	12.79	81.86	72.26	0.36	0.41	0.237
InternVL3-9B (2025)	82.34	85.78	16.63	83.36	82.13	1.12	1.19	0.103
LLaVA-NeXT-Video-7B (2024)	81.30	84.98	14.51	82.72	55.06	3.94	4.01	0.362
Phi-3.5-Vision (2024)	82.05	85.37	14.56	82.88	77.53	0.51	0.58	0.240
MERaLiON-Omni-VL w/o Pretrain	79.22	83.30	12.95	80.39	65.62	1.03	1.10	0.994
MERaLiON-Omni-VL	80.41	83.28	12.93	81.54	72.32	0.98	1.03	0.675
MERaLiON-Omni-Image	83.32	83.19	12.93	83.60	85.11	0.11	0.19	0.129
MERaLiON-Omni-MM	84.98	85.65	18.78	84.13	<u>88.68</u>	0.07	0.12	0.071
+ Density Reward	<u>85.21</u>	<u>85.92</u>	17.76	<u>84.39</u>	88.69	0.04	0.09	<u>0.070</u>
+ Coherence Reward	84.72	85.69	<u>19.26</u>	84.22	88.10	<u>0.05</u>	<u>0.10</u>	0.072
+ Dual (GRPO)	85.31	86.56	20.05	85.04	87.83	<u>0.05</u>	0.09	0.069

Table 3: Performance on the SEA-Image benchmark. We report the standard Character/Word Error Rate (C/WER) and an adjusted WER* as OCR metrics. Best results are in bold; second best are underlined.

Method	Summ	AQA	VQA	MCQA
Qwen3Omni-30B	84.75	18.75	17.02	83.07
Qwen2.5Omni-7B	85.13	16.82	16.50	49.83
MO-VL	<u>85.22</u>	17.99	17.07	24.35
MO-Image	77.32	11.33	11.67	10.19
MO-MM	85.10	<u>24.10</u>	<u>23.37</u>	<u>84.89</u>
MO-GRPO	86.89	26.97	26.99	86.01

Table 4: Performance on the SEA-Video-v1 benchmark, with Summary measured by BERTScore, AQA and VQA by F1-score, and MCQA by accuracy.

Preliminary ablation results indicate that initializing from pre-trained parameters significantly improves MCQA and OCR performance over a cold start. Building upon this, our MERaLiON-Omni-GRPO leads every reported SEA-Image metric among the evaluated baselines. Notably, our model achieves an accuracy of 87.83% in MCQA, surpassing SeaLion4VL⁴ (83.32%) and Qwen3Omni (86.21%), and significantly outperforming the size-equivalent Qwen2.5Omni (72.26%). The ranking persists with three non-Qwen VLMs where MO-GRPO leads every SEA-Image metric. The impact of our region-specific training is particularly nuanced in the synthetic overlay OCR tasks. We observe a significant performance gain in WER, where our model achieves 0.09 compared to 0.31 for Qwen3Omni. We also refer to the simplified WER* metric⁵, which focuses purely on substitutions and deletions while ignoring insertion errors. Under this lens, our model maintains its lead

⁴SeaLion4VL is a 27B VL model tailored for Southeast Asia. Currently, it only supports static image-text input.

⁵WER = (s + d + i) / N, WER* = (s + d) / N, where s, d, and i are the numbers of substitutions, deletions, and insertions, and N is the total number of words in the reference.

(0.069) over SeaLion4VL (0.094). Table 3 further ablates each reward component: density-only raises BERTScore but lowers QA-F1, indicating that optimizing cultural-density rewards alone does not improve all task metrics uniformly. The coherence-only reward improves QA but lowers Captioning and MCQA, and the dual formulation yields the best overall balance, with only a slight MCQA trade-off (−0.85%) attributable to GRPO favoring open-ended generation. Crucially, the 4K GRPO prompts are a subset of the Stage-3 training data already seen during SFT. Therefore, the gains from “-GRPO” over “-MM” cannot be attributed to exposure to additional training data, but instead reflect the effect of reward-shaped culturally-aware post-training on previously seen data. For significance, item-level tests confirm OCR-WER and QA-F1 gains against all comparators ($p < 0.001$), whereas the +1.6-point MCQA gap to Qwen3Omni is not significant ($p = 0.21$; Appendix G).

On the SEA-Omni Video Benchmark (Table 4), among the evaluated audio-capable models, MO-GRPO obtains the highest reported Summary, AQA, and VQA scores. While Qwen3Omni-30B performs competitively on the MCQA task, it shows clear limitations in handling region-specific audio-visual details, with scores below 19% on both AQA and VQA. In comparison, our model attains higher performance on these tasks, reaching 26.97% and 26.99%, respectively. These results suggest that cultural alignment enables more effective modeling of local acoustic and visual cues. From a training perspective, we observe a performance drop at the MO-Image stage, likely due to missing temporal information. This gap is mitigated during the MO-MM stage, restoring perfor-

Model	PB-Read		Topic-Read		Daily Conv.		Code-Switch		Themed Conv.		Service Call	
	WER	WER*	WER	WER*	WER	WER*	WER	WER*	WER	WER*	WER	WER*
Qwen3Omni-30B	0.062	0.053	0.324	0.242	<u>0.278</u>	<u>0.221</u>	0.440	0.368	0.187	0.151	0.147	0.122
Qwen2.5Omni-7B	0.291	0.237	0.792	0.598	0.820	0.744	0.886	0.851	0.751	0.655	0.783	0.666
MERaLiON-AL	<u>0.059</u>	0.048	0.144	0.130	0.263	0.210	0.342	0.268	0.199	0.154	0.160	0.123
MERaLiON-Omni-VL	0.108	0.081	0.262	0.234	0.559	0.490	0.696	0.648	0.504	0.444	0.426	0.348
+ $\mathcal{D}_{s_img}$	0.108	0.081	0.257	0.229	0.545	0.476	0.703	0.653	0.518	0.451	0.430	0.349
+ $\mathcal{D}_{imm}$	0.055	<u>0.046</u>	0.112	0.101	0.284	0.239	0.338	<u>0.279</u>	0.199	0.160	<u>0.156</u>	0.128
+ GRPO	0.055	0.045	<u>0.113</u>	<u>0.102</u>	0.283	0.239	<u>0.341</u>	0.281	<u>0.196</u>	0.159	0.147	0.122

Table 5: Performance comparison of different models on six SEA-ASR datasets (Appendix B). Promising results on 35 AudioBench tasks further demonstrate the effectiveness of leveraging this orthogonality characteristic (Table 14).

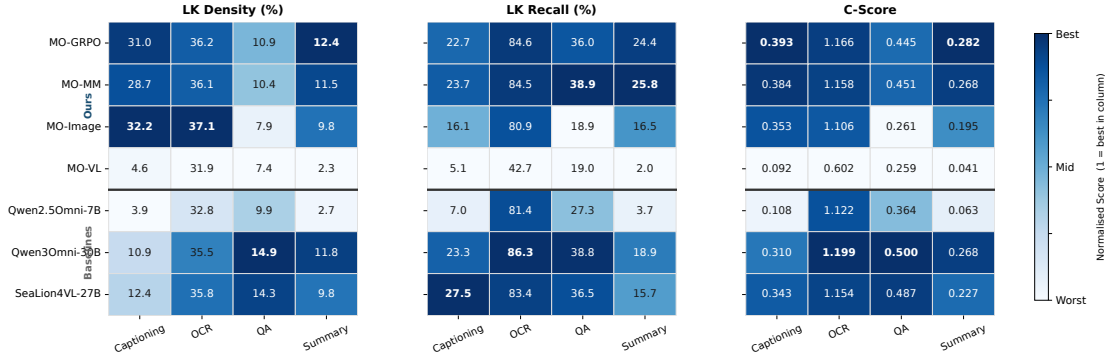


Figure 3: Per-task cultural alignment analysis. Color intensity reflects column-normalized performance (darker = better). C-Score definition is presented in Appendix J; full metrics in Figure 6.

mance and supporting the final GRPO refinement. Crucially, as SEA videos are excluded from training, these **zero-shot results** demonstrate effective transfer of cultural knowledge from static images to dynamic multimodal reasoning tasks.

To holistically assess cultural alignment beyond individual metrics, we adopt a composite Cultural Alignment Score (C-Score; Appendix J) that rewards dense and accurate local-knowledge generation while penalizing hallucination. As shown in Figure 3, MO-GRPO achieves the highest C-Score across most tasks, outperforming baselines with up to $3\times$ more parameters, confirming the effectiveness of our dual-reward GRPO in balancing cultural depth with factual precision.

5.2 Performance on ASR Datasets

We evaluate the models on six diverse SEA Automatic Speech Recognition (ASR) datasets (Wang et al., 2025b), reporting both WER and WER*. As presented in Table 5, MERaLiON-Omni demonstrates superior audio understanding ability compared to general-purpose models. For instance, on Code-Switch, our MM variant achieves a WER of 0.338, significantly lower than Qwen2.5Omni-7B (0.886). Even compared to the much larger

Qwen3Omni-30B (0.440), our model remains highly competitive while being more efficient. Ignoring insertion errors, our final model still achieves an average WER* of 0.158, surpassing Qwen models’ 0.625 and 0.193. The results demonstrate that our progressive training strategy supports final-model recoverability after multimodal integration. The progression from MERaLiON-AL (He et al., 2025) to MERaLiON-Omni-VL and its variants shows that MO-VL incurs a transient audio dip, but Stage 3 restores performance (e.g., Code-Switch WER 0.696→0.338). This indicates that orthogonality supports final-model recoverability, not lossless intermediate updates. Additional results across 35 AudioBench tasks in Appendix I further support this argument.

5.3 Performance on General Benchmarks

We further evaluate our trained models on six general-purpose benchmark metrics: Video-MME, Video-MME-sub (Fu et al., 2025), MVBench (Li et al., 2024), MME (Fu et al., 2025), MMB-V11 (Duan et al., 2024), and MMVet (Yu et al., 2024). Results in Table 6 show that MO-Image exhibits a temporary dip on several general benchmarks, whereas MO-MM and MO-GRPO recover

Model	Video Benchmarks			Image Benchmarks		
	Video-MME	Video-MME-sub	MVBench	MME	MMB-V11	MMVet
Qwen3Omni-30B	61.48	66.22	52.11	2,043	86.18	73.90
Qwen2.5Omni-7B	54.19	58.63	53.96	1,756	81.97	61.10
MERaLiON-Omni-VL	49.26	55.33	32.36	1,434	53.25	30.40
MERaLiON-Omni-Image	44.37	51.26	27.71	1,326	47.06	30.70
MERaLiON-Omni-MM	52.52	56.67	36.98	1,669	62.03	34.10
MERaLiON-Omni-GRPO	53.11	57.26	38.06	1,673	63.22	38.50

Table 6: Comprehensive performance comparison of the proposed MERaLiON-Omni models against strong baselines (e.g., Qwen-Omni) on general-purpose video and image multimodal benchmarks.

Metric	MO-MM	MO-GRPO
Avg-LK / response	1.27	0.56
Injection rate (%)	24.0	16.1
Sum-LK	260	121

Table 7: Judge-extracted localized-entity prevalence on external MMVet images. *Avg-LK / response*: the average number of unique LKs per successfully parsed response. *Injection rate*: the fraction of successfully parsed responses containing at least one such entity. *Sum-LK*: the sum of these per-response counts.

and exceed the vision-language-aligned MO-VL baseline on all six metrics. For example, MME increases from 1,434 to 1,673 and Video-MME from 49.26 to 53.11. Thus, cultural post-training does not reduce general-benchmark performance relative to MO-VL. Relative to Qwen2.5Omni, MO-GRPO is close on MME (1,673 vs. 1,756) and Video-MME (53.11 vs. 54.19), while lower on Video-MME-sub, MVBench, MMB-V11, and MMVet. As this work prioritizes SEA localization, broader visual alignment remains constrained by the scale and diversity of general visual data, and expanding this coverage is an important direction for future work.

5.4 Entity Over-Injection on External Images

A potential concern is that culturally-aware post-training may encourage models to introduce unnecessary localized entities when describing inputs outside the SEA-Image benchmark. We therefore compare MO-MM and MO-GRPO on 218 MMVet images using the neutral prompt “Describe this image in detail.” For both models, we apply the same LK judge used by the cultural-density reward to count judge-extracted entities in each response.

As shown in Table 7, GRPO reduces the mean number of extracted entities from 1.27 to 0.56 per response and lowers the response-level injection rate from 24.0% to 16.1%, reducing the total count

from 260 to 121. MO-MM has 14 judge parse failures, and treating these as zero-entity responses will change its mean to 1.19 and its injection rate to 22.5%, both still above the MO-GRPO values. Thus, within this external-image analysis, GRPO does not increase the entity over-injection signal relative to MO-MM; instead, it yields substantially fewer entity outputs.

6 Conclusion

We proposed MERaLiON-Omni, along with two carefully built benchmarks for Southeast Asia. Through progressive training and culturally-aware post-training with text-mediated dual rewards, the model achieves strong regional performance while improving general-benchmark results relative to its VL-aligned base model. Our results further suggest that effective regional specialization can be attained through data-efficient and culturally attentive training strategies. We hope this study can serve as a useful reference for developing more equitable and culture-aware AI systems in underrepresented regions with similar linguistic diversity. A broader research preview of our MERaLiON-Omni exploration is available in our technical report (Zhang et al., 2026).

The model weights and both benchmarks will be released upon publication to support reproducibility and community use. Additional details will be provided at <https://huggingface.co/zzlynxSG>.

7 Limitations

We summarize six limitations of our work. (1) Southeast Asia encompasses a very large number of dialects, many of which remain underrepresented in our current evaluation benchmarks. (2) Uncovering latent cultural concepts through dynamic sampling inevitably involves trade-offs between computational efficiency and training stability. Moreover, the reliance on recursively generated data poses

a theoretical risk of model collapse (Shumailov et al., 2024), where subtle, long-tail cultural nuances could be homogenized over successive training iterations. (3) Our video reasoning is currently limited to short contexts due to computational constraints, potentially missing long-term temporal dependencies in extended footage. (4) Our cultural alignment rewards operate through textual proxies: cultural density is measured from generated text, and multimodal coherence is assessed against textual summaries rather than raw visual or audio signals. While we validate that these proxies correlate well with human judgment, developing reward functions that directly evaluate culturally grounded responses against raw multimodal evidence remains an important direction for future work. (5) Our OCR evaluation focuses on recognition of multilingual text synthetically overlaid on images, covering Latin and non-Latin scripts prevalent in Southeast Asia. While this setup effectively tests the model’s multilingual visual decoding capability under controlled transformations, it does not directly evaluate performance on natural scene text such as street signage, handwritten documents, or in-the-wild text captured in photographs. (6) The 100-sample judge validation does not eliminate residual reward error, and native-speaker head-to-head evaluation remains an important direction for future work.

8 Ethics Statement

Data privacy and safety. In this research, the introduced SEA-Omni Benchmark Suite is constructed from a combination of open-source datasets and publicly available internet content. We have implemented strict filtering pipelines to remove content containing excessive violence, hate speech, or explicit material to ensure the safety of the model and benchmark. For the SEA-Video-v1 benchmark involving human subjects, we prioritize videos from YouTube (Creative Commons) to minimize privacy risks regarding personally identifiable information.

Annotator compensation and fair work. Our manual annotation and evaluation processes (e.g., SEA-Video-v1) involve a diverse group of experts, including native speakers from Southeast Asia and long-term residents with at least three years of immersive experience in the region. We ensure that all annotators are compensated fairly, with hourly rates exceeding the local minimum wage standards

in their respective countries. All participants are informed about the nature of the task and the intended use of the data for research purposes.

9 Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore. The computational work for this article was fully performed on resources of the National Supercomputing Centre (NSCC), Singapore (<https://www.nsc.sg>).

References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. *Preprint*, arXiv:2308.12966.
- Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618.
- Samuel Cahyawijaya, Holy Lovenia, et al. 2025. *Crowd-source, 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.
- Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. *Multimodal Language Models See Better When They Look Shallower*. In *Proceedings of the 2025 Conference on EMNLP*, pages 6677–6695, Suzhou, China.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. *An Image is*

- Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021*.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.
- Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. *CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. OneLLM: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595.
- Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, and Aiti Aw. 2025. MERAiON-AudioLLM: Advancing Speech and Language Understanding for Singapore. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 22–30.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. Preprint, arXiv:2106.09685.
- Ram Mohan Rao Kadiyala, Siddhant Gupta, Jebish Purbey, Srishti Yadav, Suman Debnath, Alejandro R. Salamanca, and Desmond Elliott. 2025. *Uncovering Cultural Representation Disparities in Vision-Language Models*. In *Proceedings of IJCNLP-AACL 2025*, pages 2087–2117, Mumbai, India.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. CultureVLM: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.
- Sophia Maria. 2024. *Compass: Large Multilingual Language Model for South-east Asia*. *arXiv preprint arXiv:2404.09220*.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, et al. 2024. SeaLLMs - Large Language Models for South-east Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304.
- David Ong and Peerat Limkonchotiawat. 2023. *SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models*. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.
- Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.
- Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. *A survey of multilingual large language models*. *Patterns*, 6(1):101118.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. GeoDE: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

- Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. [AudioBench: A Universal Benchmark for Audio Large Language Models](#). In *Proceedings of NAACL 2025*, pages 4297–4316, Albuquerque, New Mexico.
- Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F. Chen, and AiTi Aw. 2025b. Advancing Singlish Understanding: Bridging the Gap with Datasets and Multimodal Models. *arXiv preprint arXiv:2501.01034*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025a. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. 2025b. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57730–57754.
- Longyin Zhang, Shuo Sun, Yingxu He, Won Cheng Yi Lewis, Muhammad Huzaifah Bin Md Shahrin, Hardik Bhupendra Sailor, Heng Meng Jeremy Wong, Tarun Kumar Vangani, Yi Ma, Qiongqiong Wang, et al. 2026. [Unlocking cognitive capabilities and analyzing the perception-logic trade-off](#). *arXiv preprint arXiv:2602.23730*.
- Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2025. [SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025) — System Demonstrations*, pages 96–105.
- Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. [Llava-next: A strong zero-shot video understanding model](#).
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models](#). *Preprint*, arXiv:2304.10592.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Experimental Setup

Settings. We trained the unified architecture (Figure 2) through a three-stage progressive training strategy. For Stage 2, we trained the model on large-scale SEA images, with all parameters frozen except for the ViT module and the visual adapter. For the other two stages, we applied LoRA fine-tuning to the following modules: *attn.q*, *attn.k*, *attn.v*, *attn.proj*, *mlp.0*, *mlp.2*, *gate_proj*, *up_proj*, *down_proj*, *q_proj*, *k_proj*, *v_proj*, and *o_proj*. Notably, these module names strictly adhere to the naming conventions of the two pre-trained models (MERaLiON-v2 and Qwen2.5Omni-ViT). For the supervised fine-tuning tasks, we utilized standard Hugging Face training scripts with default hyperparameters.

For reinforcement training, we optimize GRPO using a PPO-style clipped surrogate objective. To ensure training stability, we implemented a best-of-group guarantee: if the normalized advantage of the highest-scoring response in a group is non-positive, it is overridden to a small constant $\delta = 1e-6$. This ensures the policy always receives a positive learning signal for the best relative generation. Table 8

Hyperparameter	Value
Optimizer	Adam
Learning Rate	3e-6
Max Gradient Norm	1.0
PPO Clip Epsilon (ϵ)	0.2
Normalization Stability (ξ)	1e-8
Gradient Accumulation Steps	4
Min. Advantage Guarantee (δ)	1e-6
Target KL Divergence	0.04
Initial KL Weight	0.02
LoRA Rank (r)	8
LoRA Alpha (α)	32
GRPO Group Size	2

Table 8: Hyperparameters for GRPO implementation.

details the configuration used to balance training stability and efficiency. For the hyperparameter λ in Eq. (4), empirical testing suggested a value of 0.2, which helps limit keyword stuffing. We also explored an equal weighting scheme ($\lambda = 0.5$) during ablation. However, this setup caused instability in the reinforcement training and degraded overall performance.

All training is conducted on NVIDIA H100 GPUs, using 8 to 48 GPUs depending on the training stage, with a total training time of approximately 10 to 15 days. Unless otherwise stated, all model training and automatic evaluation results are reported from a single final run due to the computational cost of training omni-perception models.

B Training Data

Three-stage training. Our progressive training strategy utilizes a data-efficient pipeline comprising three distinct categories of datasets. By curating specific data for each stage, we minimize the computational cost while maximizing cultural alignment. The details are as follows:

- **Efficient VL alignment.** We use a set of open-source datasets (TextVQA⁶, DocVQA⁷, MathVista⁸, DVQA⁹, WebVid-10M¹⁰, Koala-36M-v1¹¹), containing approximately 158K general-domain image/video-text QA pairs, where the image data contains ground-truth while for the video data we create the training data using Qwen2.5Omni as a teacher model for knowledge distillation following a procedure of: (1) Obtaining video description using the teacher model, (2) forming a natural video question based on the description using GPT-4, and (3) asking the teacher model to answer the question. The objective of this stage is purely structural: to align the vision encoder (ViT) with the LLM’s embedding space, enabling the model to recognize basic objects and scenes without overfitting to specific cultural domains.

⁶https://huggingface.co/datasets/redactable-llm/TextVQA/blob/main/TextVQA_0.5.1_train.json

⁷<https://huggingface.co/datasets/lmms-lab/DocVQA>

⁸<https://huggingface.co/datasets/AI4Math/MathVista>

⁹<https://huggingface.co/datasets/DavidNguyen/DVQA>

¹⁰<https://huggingface.co/datasets/TempoFunk/webvid-10M>

¹¹<https://huggingface.co/datasets/Koala-36M/Koala-36M-v1>

- **Deep cultural knowledge ingestion.** This is the core stage for instilling regional knowledge. We employ our curated SEA-Image training set (Subsec. 3), which consists of 4.23M culturally grounded image-text pairs. Unlike Stage 1, we perform full-parameter fine-tuning on the vision encoder here. This allows the model to learn fine-grained visual features specific to Southeast Asia that are typically absent in general web-crawled datasets.
- **Advanced multimodal fusion.** To enable cross-modal reasoning, we incorporate a cross-domain multimodal instruction-tuning dataset. This includes 21.2K multi-task image data from sea-vl_crowdsourcing¹² (no intersection with the training set), 3K image-audio instances from OmniInstruct_v1, 5.4K video-audio instances from Koala-36M-v1, and a small ASR subset with 12K instances from the training set of Multitask-National-Speech-Corpus-v1.

RL training. We randomly sample 4K instances from the Stage 3 training split ($\mathcal{D}_{mm}$) to support the dual-reward GRPO training mechanism.

National speech corpus. The National Speech Corpus (Wang et al., 2025b) comprises approximately 10,600 hours of SEA recordings. It is categorized into six subsets (referenced in Table 5 as headers):

- **Phonetically Balanced Readings:** 3,000 hours of prompted readings derived from phonetically balanced scripts, ensuring comprehensive coverage of phonemes in the local accent.
- **Topic-Specific Readings:** 3,000 hours of prompted speech, this subset focuses on specific entities and topics, including sentences related to people, food, locations, and brands.
- **Daily Conversations:** It contains 900 hours of spontaneous dialogue data, capturing interactions centered around everyday conversations and gaming scenarios.
- **Code-Switching Conversations:** A 900-hour subset consisting of conversations in which speakers switch between Singlish and their native languages (Chinese, Malay, or Tamil).

¹²https://huggingface.co/datasets/SEACrowd/sea-vl_crowdsourcing

- **Thematic & Emotional Conversations:** 1,500 hours of dialogue focus on 4 themes: debate, finance, positive emotions, and negative emotions.
- **Simulated Service Calls:** A total of 1,300 hours of simulated telephone calls cover three domains: tourism and hospitality, commercial services, as well as government services.

C Inter-Annotator Reliability

Evaluation methodology. To assess the quality of our curated SEA-Omni Image dataset, we conduct a manual evaluation of 200 random samples across all five tasks. For this task, two annotators with years of experience living in Southeast Asia participate in the evaluation process. Specifically, the annotators are asked to evaluate the quality of each sample and assign a score from 0 to 1. To ensure objectivity and factual accuracy, they are allowed to use external resources such as image search engines to verify entities with specific cultural significance. Given the continuous nature of the scores, we report the Intraclass Correlation Coefficient (the two-way random effects, absolute agreement, single rater model (ICC2)) to measure inter-rater reliability.

Results and discussion. Table 9 shows the results across all five tasks using 200 samples, while Table 10 shows the focused results for the 106 samples from the three semi-automatically generated tasks (i.e., Summary, QA, and MCQA). Results in Table 10 show that manual evaluation yields an ICC2 score of 0.915, which indicates excellent agreement between the two annotators. Moreover, the average quality score across the evaluated samples is 0.704, indicating useful but imperfect quality in the semi-automatically generated tasks. Together with the high ICC2 score, this supports the consistency of the evaluated annotation subset while motivating cautious interpretation of residual annotation noise.

Annotation instructions. Annotators were instructed to evaluate the overall quality of each sampled Summary, QA, and MCQA instance. For each sample, they examined whether the generated text was factually consistent with the corresponding image or video context, whether the question and answer were clear and answerable, and whether the sample preserved the intended Southeast Asian cultural context when such context was present.

Type	ICC	F	df1	df2	p-value	95% CI
ICC1	0.938	31.455	199	200	< .001	[0.92, 0.95]
ICC2	0.938	32.244	199	199	< .001	[0.92, 0.95]
ICC3	0.940	32.244	199	199	< .001	[0.92, 0.95]
ICC1k	0.968	31.455	199	200	< .001	[0.96, 0.98]
ICC2k	0.968	32.244	199	199	< .001	[0.96, 0.98]
ICC3k	0.969	32.244	199	199	< .001	[0.96, 0.98]

Table 9: Inter-rater Reliability (ICC) scores across the five tasks in the SEA-Image Benchmark.

Type	ICC	F	df1	df2	p-value	95% CI
ICC1	0.915	22.410	105	106	< .001	[0.88, 0.94]
ICC2	0.915	23.499	105	105	< .001	[0.87, 0.94]
ICC3	0.918	23.499	105	105	< .001	[0.88, 0.94]
ICC1k	0.955	22.410	105	106	< .001	[0.93, 0.97]
ICC2k	0.955	23.499	105	105	< .001	[0.93, 0.97]
ICC3k	0.957	23.499	105	105	< .001	[0.94, 0.97]

Table 10: Focused Inter-rater Reliability (ICC) analysis for the three semi-automatically generated tasks, i.e., Summary, QA, and MCQA.

For MCQA items, annotators additionally checked whether the correct answer was supported by the visual or audiovisual context and whether the distractors were plausible but not correct. Each sample was assigned a continuous quality score from 0 to 1, where higher scores indicate better factual consistency, clarity, and cultural appropriateness. To improve objectivity and factual accuracy, annotators were allowed to consult external resources, such as image search engines, when verifying culturally specific entities.

D Dual-Reward Prompt Engineering

Our reward function is composed of two heads: Cultural Density, $\rho_c(y)$, and Multimodal Coherence, $s_c(y, x)$. To quantify cultural relevance, we use the powerful Qwen3-235B model as an automated judge to extract Localized Knowledge (LK) entities from the model’s response. The prompt used to guide the LLM judge is as follows:

Prompt for cultural density judge

You are a highly intelligent linguistic and cultural analyst specializing in Southeast Asian studies. Your task is to extract two types of keywords from a given sentence: General Knowledge (GK) and Localized Knowledge (LK). Here are the definitions you must follow:

- **General Knowledge (GK):** These are common, high-level, or generic terms that a general audience would understand without specific cultural

context. Examples: ‘temple’, ‘food’, ‘noodles’, ‘building’.

- **Localized Knowledge (LK):** These are specific, culturally rich proper nouns or terms that require domain-specific knowledge of Southeast Asia. Examples: ‘Wat Arun’, ‘Nasi Lemak’, ‘Batik’, ‘Pho Bo’.

Now, analyze the following sentence. First, think step-by-step about which words or phrases fall into each category. Then, provide your final answer ONLY in a valid JSON format. The JSON object must contain two keys: “gk_keywords” and “lk_keywords”, with their values being lists of the extracted keywords. If a category has no keywords, return an empty list. The sentence to analyze is: {sentence} Analyze the sentence and provide the JSON output.

To ensure factual grounding, we employ a fact-anchored deductive scoring rubric, also executed by Qwen3-235B. The prompt for the coherence judge is:

Prompt for multimodal coherence judge

You are an expert AI evaluator, acting as a response quality grader. You will receive a user question, a textual Description (desc) of a multimodal input, and a model’s response. Your task is to score the model’s response based on a deductive rubric. Start with a total score of 5 and subtract points for each flaw identified. Scoring Rules:

- **(A) Common Sense Errors:** Contradicts common knowledge. Deduct 1 point.
- **(B) Factual Mistakes:** Contradicts statements in the provided desc. Deduct 1 point.
- **(C) Hallucination:** Detail is implausible or directly contradicts desc. Deduct 1 point.
- **(D) Inconsistent Response:** Self-contradictory or logical leaps. Deduct 1 point.
- **(E) Question-Answer Mismatch:** Fails to answer the user’s question. Deduct 1 point.

Output Format: Your output MUST be a single JSON object:
{"justification": "<briefly explain>",
"deductions": [<codes>], "score":
<final_score>}
=====

Your Task:

Question: {instruction}
Multimodal Description (desc):
{mm_desc_RL}
Model Response: {answer_one}
Your judgment:

E Dynamic Temperature Sampling + Dual Reward Strategy

A key challenge in RL-based alignment is generating a diverse yet high-quality set of responses for the GRPO ranking step. We observe a strong correlation between the generation temperature and the prevalence of local keywords in the omni model output. As shown in Figure 4, lower temperatures tend to produce more generic responses, while higher temperatures encourage more diverse and culturally specific (but potentially less coherent) text.

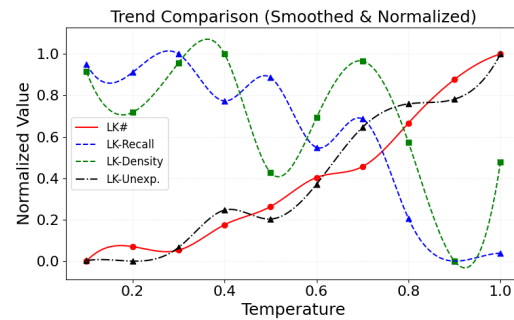


Figure 4: Higher temperatures expose LKs but increase error rates, motivating our dual-reward mechanism.

To generate diverse and effective model responses for GRPO training, we vary the sampling temperature across the rollouts within each group. Specifically, we use an incremental temperature strategy, starting at 0.1 and increasing by 0.1 to search for a high-reward candidate. This approach ensures the inclusion of at least one high-reward candidate within each group, which provides a stable optimization signal while encouraging the exploration of diverse concepts through higher temperatures. We also evaluated fixed temperatures of 0.1 and 0.9; both degraded performance compared to dynamic sampling, confirming that adaptive temperature control is beneficial for balancing diversity and coherence.

Table 11 compares two different normalization strategies. Batch-wise normalization yields consistent gains across all tasks, as it stabilizes advantage estimation under the high reward variance introduced by our dual-reward formulation.

Norm.	Cap.	QA	Summ.	MCQA	OCR-WER
Group-wise	84.70	85.18	83.80	87.74	0.054
Batch-wise	85.31	86.56	85.04	87.83	0.045

Table 11: Effect of normalization strategy on GRPO.

F Parameter Orthogonality Analysis

To investigate potential interference between modalities, we examined the cosine similarity between the parameter updates from the audio tuning stage ($\Delta\Theta_{AL}$) and the visual alignment stage ($\Delta\Theta_L$ in Stage 1). As illustrated in Figure 5, the similarity scores across layers are extremely small (10^{-3}). This suggests that the optimization directions for the two tasks are nearly orthogonal. It should be noted that this observation aligns with the strong ASR performance shown in Table 5. This observation is consistent with the final model retaining strong audio performance on many tasks, while degradation remains on some speech translation and instruction-following tasks.

Let A , A' , A'' denote the backbone model (google/gemma-2-9b-it), the audio-adapted model (MERaLiON-2), and the Stage 1 visually-aligned model (all after LoRA merging). For each weight tensor indexed by layer l and module m , we compute two task vectors: $\vec{b}^{(l,m)} = A'^{(l,m)} - A^{(l,m)}$ (audio) and $\vec{c}^{(l,m)} = A''^{(l,m)} - A'^{(l,m)}$ (visual). To exclude trivially zero positions, we restrict the analysis to an intersection mask $\mathcal{M} = \{i : |\vec{b}_i| > \epsilon \wedge |\vec{c}_i| > \epsilon\}$ with $\epsilon = 10^{-5}$, covering 93.86% of all non-trivially updated positions. We then compute per-module cosine similarity within $\mathcal{M}$ (Figure 5). The reported global score ($\approx -10^{-4}$) is the arithmetic mean across all 42 layers $\times$ 7 module types.

Same-modality control. Under the identical recipe (294 module tensors; $\epsilon = 10^{-5}$), we add a second audio direction $\Delta\Theta_{AL2} = A''' - A''$ from ASR continued training on A'' . The results in Table 12 indicates that initial audio adaptation and visual alignment are near-orthogonal but this local relationship is not invariant under sequential training. After visual alignment, continued ASR optimization exhibits a consistent component opposite to the prior visual update, consistent with partial recovery of audio-relevant behavior in shared parameters.

G Significance of Headline Results

Main results are obtained with greedy decoding. We assess statistical significance using McNemar exact tests for MCQA and paired bootstrap (10,000 resamples) for QA-F1 and OCR-WER. Across all comparators, improvements in QA-F1 (+0.034 to +0.073) and OCR-WER (-0.48 to -3.84) are significant (all $p < 0.001$). For MCQA, the model

Comparison	Mean cos	Mean cos
visual vs. audio (cross)	-1.5×10^{-4}	3.6×10^{-4}
audio vs. audio2 (<i>same</i>)	$+6.3 \times 10^{-4}$	7.6×10^{-4}
visual vs. audio2 (cross)	-9.5×10^{-3}	9.7×10^{-3}
permutation null	≈ 0	2.7×10^{-5}

Table 12: Orthogonality control with real task vectors.

Metric	MO-GRPO	Qwen3-Omni
Peak VRAM	32.3 GB	71.6 GB
Latency / image	10.8 s	45.7 s
Latency / video	15.4 s	41.6 s
Decode speed	16.6 tok/s	11.8 tok/s

Table 13: Single-GPU inference efficiency.

significantly outperforms all comparators (+5.70 to +32.77, $p < 1e-6$) except for Qwen3Omni, where the gap is (+1.62, 95% CI [-0.77, +4.00], $p = 0.209$).

H Inference Efficiency

We use a single H200 GPU with bf16 precision, FlashAttention-2, greedy decoding, and max_new_tokens set to 256. Details are presented in Table 13.

I Extended Performance Analysis

As presented in Table 14, we compare the audio-only model (MERaLiON-AL), our vision-integrated model (MERaLiON-Omni), and the size-matched baseline Qwen2.5Omni (7B). The results demonstrate that MERaLiON-Omni, despite the introduction of the visual modality, preserves or improves audio performance on many QA, recognition, and dialogue tasks relative to MERaLiON-AL, while some degradation remains on speech translation and instruction-following benchmarks. The best results among the three models are bolded.

J Cultural Alignment Metrics

We report six LK metrics per model-task pair (Figure 6): LK Count, LK Density (LK-D), LK Recall (LK-Re), Reasonable Extrapolation (LK-Rs), and Visual Error Rate (LK-E). To aggregate these into a single score that balances cultural coverage against hallucination, we define:

$$\text{C-Score} = (\text{LK-D} + \text{LK-Re}) \times (1 - \text{LK-E}), \quad (7)$$

where LK-D, LK-Re, and LK-E are each in $[0, 1]$, yielding a C-Score in $[0, 2]$. The multiplicative

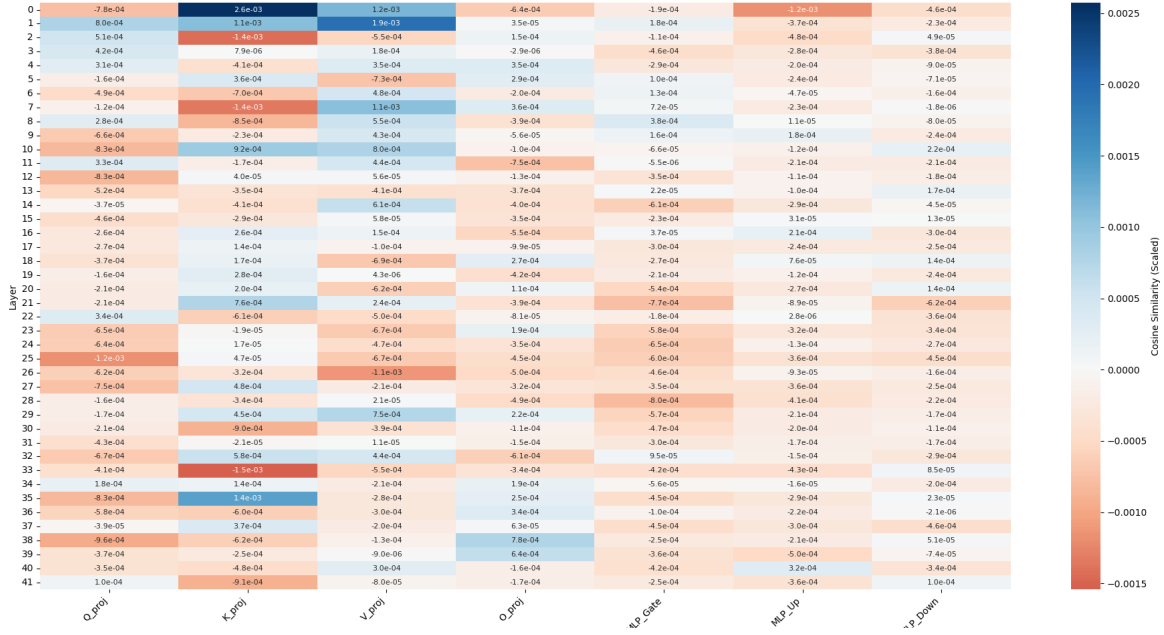


Figure 5: Layer-wise update orthogonality. Each cell shows the cosine similarity between the audio task vector ($\vec{b} = A' - A$) and the visual task vector ($\vec{c} = A'' - A'$) within a specific LoRA module of each transformer layer, computed only over parameter positions where both updates exceed $\epsilon = 10^{-5}$. Rows correspond to the 42 transformer layers; columns correspond to seven LoRA target modules (Q/K/V/O projections and MLP projections).

penalty discounts coverage gains by the error rate, so only models that are both culturally rich and factually grounded score highly.

LK quality verification. We manually verify the reliability of our automated judge (i.e., Qwen3-235B) in LK extraction. Figure 7 presents some qualitative examples comparing the model’s extracted keywords against manually detected ones. The results show that Qwen3-235B performs well in three aspects: (1) It correctly extracts entities across various languages, including Thai, Indonesian, and Chinese; (2) It can identify specific cultural elements, such as traditional desserts (*Es Doger*), deities (*Lord Murugan*), and landmarks; (3) The model’s outputs closely match human annotations, which supports its use in calculating the Cultural Density Reward (ρ_c).

For the LK evaluation in Figure 3, all our reported results are based on the evaluation of model-generated LKs against the ground-truth LKs. It is worth mentioning that we separate the unexpected LKs into two groups: LK-Rs (i.e., unexpected but reasonable LKs) and LK-E (i.e., erroneous LKs). To distinguish between LK-Rs and LK-E, we prompt Qwen3-235B to judge whether a term belongs to LK-Rs or LK-E with reference to the ground-truth answers and LK lists.

K Multimodal Coherence Referees

As stated above, we use Qwen3-235B along with the fact-anchored deductive scoring rubric to assess the quality of each generated response against its corresponding reference description (from our SVLer pipeline in Subsec. 3.1). As detailed in Appendix D, the scoring process begins with a perfect score of 5.0, from which points are deducted according to the predefined errors. This appendix further provides two real cases in Figures 8 and 9 to illustrate the referee’s decision-making process.

Channel	Dataset	MERaLiON		Baseline
		MERaLiON-AL	MERaLiON-Omni	Qwen2.5Omni
Spoken QA (English)	cn_college_listen_mcq	76.50	75.80	78.70
	slue_p2_sqa5	89.51	90.34	82.55
	dream_tts_mcq	71.90	74.80	72.00
	public_sg_speech_qa	73.28	78.90	65.87
	spoken_squad	89.04	87.70	77.46
Spoken QA (Singlish)	imda_part3_30s_sqa	58.40	60.60	45.00
	imda_part4_30s_sqa	58.80	61.20	47.60
	imda_part5_30s_sqa	68.40	72.40	60.40
	imda_part6_30s_sqa	70.80	73.80	63.00
Spoken Dialogue Summarization	imda_part3_30s_ds	51.20	52.00	35.40
	imda_part4_30s_ds	49.60	48.80	29.20
	imda_part5_30s_ds	57.80	57.60	47.00
	imda_part6_30s_ds	61.40	65.00	51.20
Gender Recognition	voxceleb_gender	97.30	99.30	17.60
	iemocap_gender	94.80	96.30	23.60
	imda_gr_sentence	53.80	56.80	1.80
	imda_gr_dialogue	70.40	74.30	3.70
Audio Scene QA	clotho_aqa	56.86	61.98	42.04
	wavcaps_qa	44.54	45.99	35.79
	audiocaps_qa	50.03	51.25	40.64
Emotion Recognition	iemocap_emotion	61.00	59.60	42.60
	meld_sentiment	64.50	47.60	33.40
	meld_emotion	54.50	40.70	35.60
Speech- Instruct	openhermes_audio	60.80	58.20	64.00
	alpaca_audio	72.00	65.60	67.60
	spoken-mqa_short	0.90	0.67	0.90
	spoken-mqa_long	0.50	0.35	0.48
	spoken-mqa_1-step	0.68	0.63	0.87
	spoken-mqa_m-step	0.36	0.28	0.77
Speech Translation	covost2_en_id	36.89	28.04	6.69
	covost2_en_zh	31.16	17.20	11.54
	covost2_en_ta	21.44	11.68	0.47
	covost2_id_en	46.32	28.46	10.20
	covost2_zh_en	20.18	14.94	10.76
	covost2_ta_en	3.16	1.57	0.05

Table 14: Detailed performance comparison on the standard AudioBench tasks (Wang et al., 2025a) to characterize audio retention after multimodal integration.

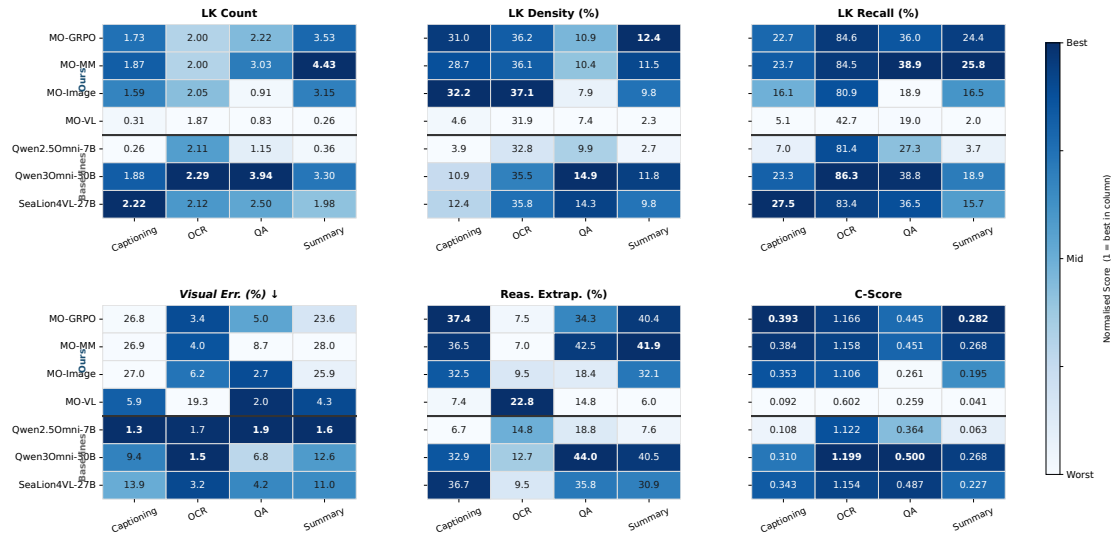


Figure 6: Complete per-task cultural alignment (6 metrics). Extends Figure 3 with LK Count, Visual Error Rate (lower is better), and Reasonable Extrapolation.

Input Caption (w/ Trans.)	Qwen3-235B Judge	Human
หิ้งพระสำหรับไหว้พระทำบุญ วัดสักระแห่งในเชียงใหม่ Trans: A shelf for Buddha images for worship and merit-making at a temple in Chiang Mai.	หิ้งพระ (Buddha shelf) ไหว้พระ (Worship) ทำบุญ (Merit-making) วัดสักระแห่ง (A temple) เชียงใหม่ (Chiang Mai)	หิ้งพระ (Buddha shelf) ไหว้พระ (Worship) ทำบุญ (Merit-making) วัดสักระแห่ง (A temple) เชียงใหม่ (Chiang Mai)
马来西亚猫山王品种的榴莲，价格不算贵，味道与泰国的差别不大。 Trans: Malaysian Musang King durian; the price is not expensive, and the taste is not much different from Thailand's.	马来西亚 (Malaysia) 猫山王 (Musang King) 泰国 (Thailand)	马来西亚 (Malaysia) 猫山王 (Musang King) 泰国 (Thailand)
बाटू गुफाएँ, मलेशिया – भगवान मुरुगन की विशाल स्वर्ण प्रतिमा और चूना पत्थर की गुफाओं तक जाने वाली रंगीन सीढ़ियों वाला प्रसिद्ध हिंदू मंदिर। Trans: Batu Caves, Malaysia – Famous Hindu temple with a huge golden statue of Lord Murugan and colorful stairs leading to limestone caves.	बाटू गुफाएँ (Batu Caves) मलेशिया (Malaysia) भगवान मुरुगन (Lord Murugan)	बाटू गुफाएँ (Batu Caves) मलेशिया (Malaysia) भगवान मुरुगन (Lord Murugan)
Ini es doger khas bandung, dengan tapai, cincau, alpukat, kacang, dan susu kental manis. Trans: This is Es Doger, typical of Bandung, with fermented cassava, grass jelly, avocado, peanuts, and sweetened condensed milk.	es doger (Es Doger) bandung (Bandung) tapai (fermented cassava) cincau (grass jelly) susu kental manis (condensed milk)	es doger (Es Doger) bandung (Bandung) tapai (fermented cassava) cincau (grass jelly)
The Esplanade theatre, also called the durian as it looks like one.	Esplanade durian	Esplanade

Figure 7: Qualitative validation of the local knowledge entity recognizer. We compare entities extracted by the automated judge against human annotations. Here, green text indicates valid LKs identified by the judge but overlooked by the annotators.



Instruction: “What color are the LED strings that light up the tree-lined pathways of Orchard Road, and what type of tropical plants are they draped across?”

Verified Context (Ground Truth): “The image captures a vibrant nighttime scene along Orchard Road... Towering **palm trees** line the streets, adorned with twinkling strings of lights—predominantly **white and colorful LEDs**—that cast a soft, reflective glow on wet pavement...”

Model’s Response: “The LED lights illuminating the trees along Orchard Road’s walkways appear **white** or **clear**... Regarding the types of tropical foliage... common choices include **palms** (e.g., coconut palm), **ferns**, **bromeliads**, **orchids**, and **bougainvillea**—all known for their resilience under humid conditions...”

LLM Judge’s Evaluation:

- **Final Score:** 4.0 / 5.0
- **Deduction(s):** ‘[B]’
- **Justification:** The response claims the LED lights are ‘white or clear’, which partially aligns with the context stating ‘predominantly white and colorful LEDs’. However, the response **omits the presence of colorful LEDs**, which is a key feature in the verified description. This constitutes a factual mistake (B) because it underrepresents the visual information. Regarding the plants, the response correctly identifies ‘palms’ but introduces other species not confirmed in the context. This is treated as plausible speculation rather than hallucination. The primary flaw is the incomplete characterization of the LED colors.

Figure 8: Example D.1: A case of a subtle factual mistake and omission. The model’s response is partially correct but omits a key detail “colorful LEDs” from the verified context, leading to a deduction for a factual mistake (B).



Instruction: “Based on the image of Sumber Maron, what type of structures are the small buildings near the waterfall?”

- [A] Religious shrines or prayer rooms;
- [B] Restrooms or public facilities;
- [C] Cafés or small restaurants;
- [D] Ticket booths or information centers;

Verified Context (Ground Truth): “A serene waterfall cascades over smooth, light-colored rocks... Nearby, several structures are visible... one appears to have blue walls... and **may serve as a shelter or refreshment spot**, though its exact use is unclear.”

Model’s Response: C

LLM Judge’s Evaluation:

- **Final Score:** 5.0 / 5.0
- **Deduction(s):** ‘[]’ (None)
- **Justification:** The model selects option [C] Cafés or small restaurants. The verified context states that a nearby structure ‘**may serve as a shelter or refreshment spot**’. This makes option [C] a plausible and reasonable inference. The response aligns with the context and does not contain any factual errors, hallucinations, or inconsistencies. It directly answers the multiple-choice question. Therefore, no deductions are warranted.

Figure 9: Example D.2: A case of high coherence. The model’s choice [C] is deemed correct because it is a reasonable inference from the verified context’s description of a building as a potential “refreshment spot”. This demonstrates the referee’s ability to validate plausible reasoning, awarding a full score.