Cultural Moment Benchmark: Evaluating Video Cultural Reasoning and Grounding in Southeast Asia

Anonymous ACL submission

Abstract

Cultural understanding in video means more than recognizing what is visible; it requires grasping the symbolic and temporal significance of cultural concepts. We decompose this into three abilities: naming what a concept symbolizes, visually recognizing it on video, and locating its sub-events in time. Existing video-cultural benchmarks tend to test what is seen, collapsing these three abilities into a single score that hides the bottleneck. We introduce the **Cultural Moment Benchmark** (CMB): 306 expert-curated concepts from seven Southeast Asia countries across five categories. We evaluate each concept through three stages, one per ability. Given a description, **Stage 1** (S1) selects from four candidate concept names, **Stage 2** (S2) selects from four candidate video moments, and **Stage 3** (S3) predicts the start and end times of the moment in a video. To keep each stage focused on a distinct ability, we use three design choices: semantic-similarity distractors (S1, S2), unlabeled video moments (S2), and free-form localization on a different example video (S3). Across six vision-language models, the failure mode varies by ability and modality. i) Even the strongest closed-source models score below 30% when all three stages must be correct; ii) Across stages, the three abilities do not fully cascade: naming a concept correctly helps half the models recognize it on video, but recognizing it has little effect on locating the sub-event in time; iii) Audio is complementary, redundant, or distracting depending on the concept, most often distracting in countries with non-Latin scripts; removing both audio and subtitles hurts Games and Music the most. Separately, our 14-rater human study shows that even *Expert* raters score below chance on concepts from a neighboring country, indicating that CMB requires country-specific cultural knowledge. CMB acts as a diagnostic harness, attributing failures to a specific ability or modality rather than aggregating them into a single score.

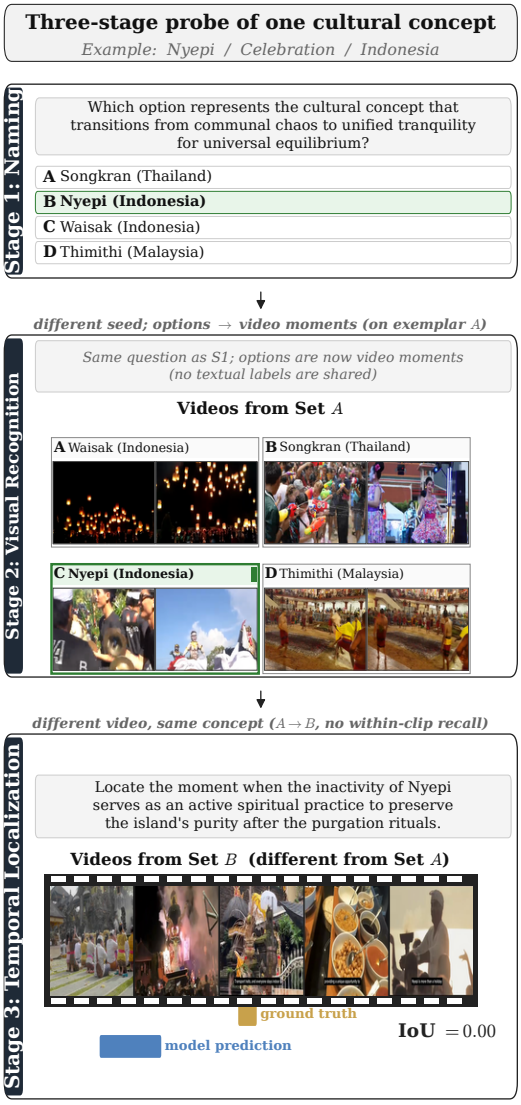

Figure 1: CMB three-stage probe. Stage 1 picks among four concept names. Stage 2 keeps the same stem but replaces the options with four video moments from example video A; it uses an independent random seed, so the option positions differ from S1. Stage 3 predicts a free-form $[t_{\text{start}}, t_{\text{end}}]$ span for a sub-event of the concept on a different example video B. The shift from A to B prevents within-clip memorization in Stage 3.

1 Introduction

Vision-language models (VLMs) have made substantial progress on video question answering and temporal grounding (Fu et al., 2024; Li et al., 2024a). Yet their ability to understand cultural content beyond what is visible remains underexplored. Existing video-cultural benchmarks (Shafique et al., 2025; Chen et al., 2025; Singh et al., 2026) tend to ask questions like "Which country is this clothing from?" or "Where is the dish popular?". Such questions can be answered by identifying what is seen, without engaging what it symbolizes. The same problem extends to cultural reasoning questions: even when a question appears to require cultural reasoning, the model can often answer it by matching keywords. A question about "a Burmese water-throwing celebration" is solved by finding a water frame without needing to know what Thingyan symbolizes (Satar et al., 2025). Our benchmark addresses this gap by asking what each concept symbolizes, such as "the washing away of misfortunes" for Thingyan, instead of just naming visible cues, as Figure 1 illustrates.

A single accuracy number conflates naming, visual recognition, and temporal localization, hiding which ability is failing. To diagnose this, we treat these three abilities as distinct and score each separately, providing a per-model failure profile rather than a single aggregate score. We instantiate this approach in the Cultural Moment Benchmark (CMB), focused on Southeast Asia (SEA). SEA cultures are diverse and underrepresented in current video-cultural benchmarks. Neighboring traditions often share enough surface features to demand fine-grained discrimination. CMB contains 306 expert-curated concepts from seven SEA countries across five categories. Four use the Latin script: Indonesia, Malaysia, the Philippines, and Vietnam. Three use non-Latin scripts: Cambodia, Myanmar, and Thailand.

We score each ability separately by routing every concept through three stages. **Stage 1** (S1) tests naming: given a description, the model picks a concept name from four candidates, asking what the concept symbolizes. **Stage 2** (S2) tests visual recognition: from the same description, the model picks among four candidate video moments, asking how it appears on video. **Stage 3** (S3) tests temporal localization: on a different example video, the model predicts the start and end times, asking when its sub-events unfold. Three design choices keep each stage focused on a distinct ability. First, S1 and S2 distractors are semantically similar to the correct answer, so models cannot rely on surface-form keyword matching. Second, S2 video moments are presented without titles, captions, or filenames, so the model must judge from visual content alone. Third, S3 uses a different example video and a free-form span prediction, so models cannot reuse cues seen earlier in the cascade. Beyond ability isolation, we evaluate each stage under three modes that vary the prior-stage context: Reset (no prior context), Carry (the model's own previous answer), and Feedback (the previous answer plus the ground truth). Together, the three stages and three modes form a 3×3 evaluation framework.

With this setup, we evaluate six state-of-the-art VLMs on CMB. Across these six models, the failure mode varies by ability and modality, confirming that a single score hides the bottleneck. The following patterns stand out. i) Single-stage scores understate the gap; end-to-end success reveals it. Even Gemini 3.1 Pro and GPT-5.4, the two strongest closed-source models, get fewer than 30% of concepts right across all three stages. The four open-source models score in single digits, and InternVL-14B is near zero. ii) Across stages, the three abilities do not fully cascade. Naming a concept correctly helps half the models recognize it on video (S1→S2). However, recognizing it has little effect on locating the sub-event in time on a different video (S2→S3). iii) Modality roles also vary: audio is not uniformly useful. Depending on the concept, it can be complementary, redundant, or distracting. The distracting role is more frequent in countries with non-Latin scripts. Removing audio and subtitles together hurts Games and Music the most, where commentary and lyrics align with the sub-event; in many traditional concepts, audio is ambient backing that holds throughout the clip. Beyond modality, cultural knowledge is country-specific, not regional. In a 14-rater human study, even *Expert* raters score below chance on concepts from a neighboring country, indicating that CMB requires country-specific cultural knowledge.

CMB acts as a diagnostic harness, attributing failures to a specific ability or modality instead of collapsing them into a single score. We contribute: i) CMB, an expert-curated video benchmark dedicated to SEA cultures, grounded in real-world traditions; ii) A 3-stage × 3-mode evaluation framework with per-stage scores, conditional metrics (c-S2, c-mIoU) for cascade analysis, and a Joint score

Benchmark	SEA focus	Multi-stage cascade	Cond. metrics	Free-form temp. loc.	Avg. dur. (min)	Human eval	Cultures covered
SCB^{image} (Satar et al., 2025)	✓	✓	✓	✗	n/a	✗	7
AVMeme Exam (Jiang et al., 2026)	✗	✗	✗	✗	≤0.5	✓ (closed)	5+
ChineseVideoBench (Nie et al., 2025)	✗	✗	✗	✗	≈1.0	✗	1
GIMMICK (Schneider et al., 2025)	partial	✗	✗	✗	0.17	✗	139
MINERVA-Cultural (Singh et al., 2026)	partial	✗	✓	✗	12.62	✓	18
VideoNorms (Varimalla et al., 2025)	✗	✗	✗	✗	0.25	✗	2
VideoVista-CulturalLingo (Chen et al., 2025)	✗	✗	✗	✗	4.05	✗	3
ViMUL-Bench (Shafique et al., 2025)	✗	✗	✗	✗	2.92	partial	14
CMB (ours)	✓	✓	✓	✓	6.2*	✓ (closed)	7

Table 1: Comparison of CMB to recent multicultural video benchmarks. SCB^{image} is the image-based precursor; all other rows are video benchmarks. **Multi-stage cascade**: distinct ability stages scored sequentially with conditional dependency. **Cond. metrics**: conditional metrics that report cross-stage success-given-success (e.g., $P(\text{S2} \mid \text{S1 correct})$). **Free-form temp. loc.**: model predicts an unconstrained $[\hat{t}_{\text{start}}, \hat{t}_{\text{end}}]$ span rather than selecting from candidate timestamps or orderings. **Human eval**: ✓indicates humans recruited separately from annotators were evaluated on the same task; *closed* adds an explicit prohibition on web search or external aids. *: S3 videos.

for end-to-end success; iii) Failure profiles for six VLMs to pinpoint each model's bottleneck stage, suggesting model-specific intervention priorities.

2 Related Work

2.1 Multicultural Image Benchmarks

Most multicultural benchmarks for VLMs use a single image as input and evaluate cultural understanding via either textual or visual answer choices. The textual-choice family includes CVQA (Mogrovejo et al., 2024), CulturalVQA (Nayak et al., 2024), SEA-VQA (Urailertprasert et al., 2024), K-Viscuit (Baek et al., 2024), CultureVerse (Liu et al., 2025), the image subset of GIMMICK (Schneider et al., 2025), and ALM-Bench (Vayani et al., 2025). CultureVerse and parts of GIMMICK additionally test cultural reasoning. MaRVL (Liu et al., 2021) checks truth-values of image captions across five languages. Region- or language-specific image benchmarks include Rice-VL (Pranav et al., 2025) and MMA-ASIA (Zheng et al., 2025) for Asia, IndicVisionBench (Faraz et al., 2026) for India, TaiwanVQA (Hsieh et al., 2025) for Taiwan, Afri-MCQA (Tonja et al., 2026) for Africa, and BanglaProtha (Fahim et al., 2026) for Bengali culture. FoodieQA (Li et al., 2024b) introduces visual options for Chinese cuisine. Most relevant to our work, SCB (Satar et al., 2025) adapts the visual-option design to SEA cultural reasoning and pairs it with cultural-artifact segmentation at the image level. CMB carries this design into video, adding a naming stage on the same description (S1), video-moment options (S2), and free-form temporal localization on a different video (S3). Image-only setups, including SCB, do not isolate naming from visual recognition and cannot locate when a cultural sub-event unfolds within a video.

2.2 Multicultural Video Benchmarks

Cultural understanding has recently been extended to video, where the challenge is harder because cultural concepts unfold across time, audio, and visual modalities. VideoVista-CulturalLingo (Chen et al., 2025) covers three cultural regions (Chinese, North American, European) in MCQ design across 1,389 videos. ViMUL-Bench (Shafique et al., 2025) is a multilingual, culturally diverse VideoQA benchmark spanning 14 languages and cultures via multimodal QA in 337 videos. MINERVA-Cultural (Singh et al., 2026) emphasizes long-form video reasoning across 18 locales with human-authored reasoning traces. AVMeme Exam (Jiang et al., 2026) probes cultural and contextual knowledge in 1,032 audio-visual memes spanning speech, songs, music, and sound effects. VideoNorms (Varimalla et al., 2025) targets cultural norms across 384 concepts in American and Chinese contexts. ChineseVideoBench (Nie et al., 2025) is a mono-cultural Chinese video QA benchmark with 1,625 videos across eight task categories. The video subset of GIMMICK (Schneider et al., 2025) also includes short video moments from 144 cultures, accompanied by AI-generated open-ended questions. Adjacent video work includes MAVEN (Li et al., 2026) for multicultural text-to-video generation and MusiQAl (Christodoulou et al., 2025) for audio-video music question answering, which test generation and music-specific understanding

rather than cultural video QA. Even within video, existing benchmarks share three design gaps that CMB closes. First, none decomposes performance by ability. Second, none uses video moments as answer options for cultural recognition. Third, none includes free-form temporal localization on a different example video. Table 1 contrasts CMB against these benchmarks across these dimensions.

3 Cultural Moment Benchmark

CMB is built on one design principle: humans produce all visual ground truth; LLMs scaffold only textual artifacts. The benchmark is constructed in three phases: i) concept curation and video collection, ii) stage construction with semantic-similarity distractors, and iii) quality validation through LLM auditing, an Outsider Filter, and S3 sub-event triage.

Annotators and Reviewers. We separate three distinct roles by qualifications and tasks. *Local Annotators* are native speakers of the country's primary language; they validate the concepts proposed by GPT-5 and filter the candidate videos. *Cultural Annotators* are also native speakers and additionally have research or co-curricular involvement in their country's cultural traditions; they curate the symbolic S1/S2 descriptions and annotate S2 moments and S3 spans. *Reviewers* are non-local researchers with documented SEA cultural exposure, distinct from Annotators; they run the Outsider Filter on every concept and triage flagged S3 sub-events. Two Cultural Annotators are co-authors (Vietnam, Malaysia); the remaining five were recruited and compensated.

3.1 Concepts and Videos

CMB concepts span the cultural diversity of SEA, organized hierarchically as `country/category/concept`. The seven countries span the diversity of SEA, covering both countries with Latin script and those with non-Latin scripts. The five categories are Celebration, Dance, Game, Music, and Wedding. For each `(country, category)` pair, GPT-5 (OpenAI, 2025) proposes candidate concepts. Each candidate is reviewed by two Local Annotators per country (three for the Philippines and Myanmar), all of whom are native speakers of the country's primary language. We retain only candidates on which all Annotators reach a unanimous consensus that the candidate names a genuine cultural tradition rather than a stereotype or an overly broad descriptor. This yields a final pool of 306 concepts. For each concept, we record four name variants along two axes: common (colloquial) vs. official (formal), and Latin transliteration vs. local script. See Appendix. For each concept, we query YouTube and Google Video Search by concept and country, e.g. "Thingyan festival, Myanmar", collecting up to 20 candidates per concept and about 6,120 in total. Local Annotators filter for real-world depiction, non-duplication, and visual clarity. The final video pool comprises 624 unique videos: Set A (one per concept, 306 videos, average 2.0 min) for S2, and Set B (318 videos, average 6.2 min) for S3, with multi-event traditions contributing more than one Set B video.

3.2 Stage Construction

Symbolic concept descriptions. Each S1 and S2 question begins with a textual description of the concept that captures its symbolic meaning, not its visual appearance. Two LLMs from different families, DeepSeek-V3.1 (DeepSeek-AI et al., 2025) and Qwen-SEA-LION-v4 (Ng et al., 2025), each draft an independent description (e.g., "the washing away of misfortunes" for Thingyan); the Cultural Annotator selects one and rewrites it if it lists only visible objects or actions. This enforces the symbolic constraint: a question that can be answered by generic object recognition ("find the water frame") is rejected at this stage.

Stage 1 (Naming). Given the description, the model selects the correct concept name from four candidates. The three distractors are drawn via a Semantic Similarity Matrix (Reimers and Gurevych, 2019) over description embeddings, as shown in Figure 2. One distractor is from the same country as the correct answer, forcing fine-grained intra-country discrimination, and two are from different countries with semantically similar concepts, preventing reliance on broad regional priors. To prevent any single concept from dominating the distractor set across items, the candidate pool for each category is dynamically capped at 10% of the category's cultural concepts. The full sampling algorithm is given in the Appendix.

Stage 2 (Visual recognition). The model sees the same description and four candidate video moments that correspond to the same concept names as S1. Each moment is a sub-event sourced from the corresponding concept's video in Set A. Gemini 2.5 Flash (Comanici et al., 2025) proposes a small

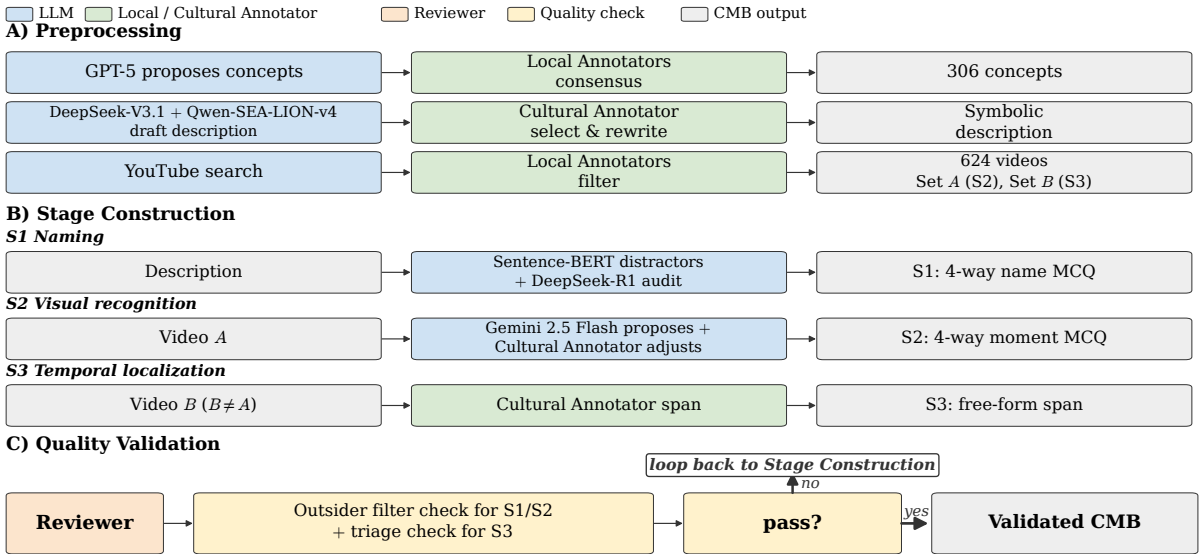

Figure 2: CMB construction pipeline. A) Preprocessing yields 306 concepts (proposed by GPT-5, validated by Local Annotators) and 624 source videos (Set A for S2, Set B for S3); descriptions are drafted by DeepSeek-V3.1 and Qwen-SEA-LION-v4 and rewritten by a Cultural Annotator. B) Each concept is materialized into the S1 and S2 questions (four-way MCQ with semantic-similarity distractors) and the S3 question (free-form span on a different video); flagged S3 sub-events are also audited by a Reviewer. C) A Reviewer runs an Outsider filter check on S1/S2 descriptions (does the Reviewer answer correctly from regional cues alone?) and a triage check on flagged S3 spans. Concepts that pass both enter the Validated CMB; failures loop back to Stage Construction.

set of candidate moments per video, highlighting cultural cues; a Cultural Annotator then retains, adjusts the boundary timestamps of, or rejects each candidate, and only Annotator-confirmed moments enter the S2 pool. Within each stage, the correct answer letter is balanced across A, B, C, and D so that no letter is correct more often than another across the item pool. S1 and S2 use independent random seeds, so the position of the correct answer differs between the two stages.

Stage 3 (Temporal localization). Given the concept name and a textual description of a target sub-event, the model predicts a free-form start and end time on a different *example video B*. Using a separate video from S2 forces the model to apply cultural knowledge to a fresh visual context rather than memorize within-clip cues. Each S3 span is annotated by a Cultural Annotator, who also verifies that the target sub-event occurs exactly once in the source video; videos containing multiple matching sub-events are excluded.

Modality handling at construction and evaluation. Cultural Annotators viewed each video with the original audio and with subtitles, if any. Reviewers performing the Outsider Filter for S1 and S2 saw only the textual description and the four name options, with no access to video, audio, or subtitles; this limits the filter to country-specific symbolic knowledge. For S3 evaluation, models receive the full video with native audio, subtitle tracks, and on-screen text intact by default.

3.3 Quality Validation

LLM auditing. DeepSeek-R1 audits the four S1 name options for mutual exclusivity and logical distinctness ("Are these options mutually exclusive and logically distinct?"), and rates each S3 sub-event description for well-definedness; low-rated items are surfaced to Reviewers for rewrite or discard. DeepSeek-R1 never sees video, audio, or frames; all S2 moments and S3 spans are produced and verified by human Annotators and Reviewers.

Outsider Filter. For every concept, a Reviewer answers the S1/S2 description from regional priors alone, without consulting country-specific sources. The Reviewer is called an *Outsider* because they have general exposure to SEA but no familiarity with the concept's country of origin. If the Outsider picks the correct name, the description has leaked something a non-local could guess from regional priors: it is either rewritten to be more symbolic (less dependent on visible surface features) or, if no rewrite restores difficulty, the concept is discarded. Because S1 and S2 share the description as their stem, the filter covers both stages.

S3 Reviewer triage. For each S3 candidate, the

textual rating produced by DeepSeek-R1 for the sub-event description flags items that may be underspecified or ambiguously locatable. A Reviewer inspects each flagged item, watches the candidate span on video B, and either rewrites the sub-event description, adjusts the span boundaries, or discards the candidate. Unflagged items pass without Reviewer review. This triage focuses Reviewer effort on the highest-risk spans rather than re-auditing all 306 items end-to-end.

4 Experiments

4.1 Implementation Details

We evaluate six vision-language models. The two closed-source models are Gemini 3.1 Pro and GPT-5.4. The other four are open-source: Qwen3-VL-32B-Instruct (Bai et al., 2025), Qwen3.5-27B (Qwen Team, 2026), and InternVL3.5-14B (Wang et al., 2025) evaluated in two configurations, (t) with inference-time reasoning and (n) without. Gemini, GPT, and Qwen read video at one frame per second; InternVL uses its native twelve uniformly spaced frames.

Each model runs each stage three times across three modes, varying what it sees from earlier stages, as shown in Table 2. Under *Reset*, the model sees nothing from earlier stages. Under *Carry*, the model's previous answer is carried over to the next stage. Under *Feedback*, the model's previous answer is carried over along with the ground truth.

S1 and S2 are scored as four-option multiple-choice accuracy. S3 uses the mean Intersection over Union (mIoU) between the predicted span $[\hat{t}_{\text{start}}, \hat{t}_{\text{end}}]$ and the ground-truth span $[t_{\text{start}}, t_{\text{end}}]$, averaged over the N=306 concepts.

4.2 Overall Results

How well do VLMs perform on CMB overall? We summarize end-to-end success with the Joint score, which credits a concept only when all three abilities line up: Joint $= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\text{S1}_i\}\cdot\mathbf{1}\{\text{S2}_i\}\cdot \text{IoU}_i$, where i indexes the N=306 concepts and $\mathbf{1}\{\cdot\}$ is the indicator function. Overall, Joint scores are low across all six models. Even Gemini and GPT, the two strongest models, achieve fewer than 30% Joint score, as shown in Table 2. The four open-source models do much worse: the two Qwen variants score in the single digits, and InternVL sits close to zero. Model scale does not explain the spread: Qwen3.5-27B beats InternVL on Joint by an order of magnitude, despite being only twice its size.

Reset / Carry / Feedback: effect of prior context. Table 2 shows that prior context helps the strongest models and hurts the weakest. Under *Carry*, Gemini's S2 accuracy rises by about ten points over *Reset*. Under *Feedback*, it rises by another thirteen points. GPT follows the same pattern. The Qwen models show no consistent change across modes, and the InternVL configurations lose accuracy under *Carry*. Prior context confirms the reasoning when the model can already perform the task; otherwise, it serves as a distraction.

4.3 Cascade Analysis

We test whether correct performance on one stage improves performance on the next, using two conditional metrics. c-S2 $= P(\text{S2} \mid \text{S1 correct})$ measures how correct naming relates to visual recognition, and c-mIoU $= \mathbb{E}[\text{IoU} \mid \text{S2 correct}]$ measures how correct visual recognition relates to temporal localization. Both are computed under *Carry*, where the model sees its own previous answer. Table 3 reports both, with significance markers against the unconditional S2 and mIoU baselines.

S1 to S2: correct naming helps the strongest models. For Gemini and GPT, once they correctly name a concept, their chance of picking the right video moment increases by about 18 points. The improvement shrinks across the open-source models: Qwen3.5-27B picks up about a third of it; Qwen3-VL-32B and both InternVL configurations show none. The smallest model, InternVL-14B (t), inverts the trend: c-S2 is below the unconditional S2 rate. **S2 to S3: correct recognition rarely improves localization.** For five models, choosing the right S2 moment provides no measurable benefit when the model must predict a start-end span on a fresh clip. Temporal localization quality is essentially the same regardless of whether S2 was correct. The one exception is InternVL with inference-time reasoning, with c-mIoU two points above its baseline, but the baseline is so low that the effect is not practically meaningful.

4.4 Modality Dependence Is Concept-Specific

We categorize each modality into three *roles* based on the performance change Δ after removing the modality: *Complementary* if $\Delta \leq -5$ (indicating a removed modality plays a complementary role); *Distracting* if $\Delta \geq +5$ (indicating a modality negatively impacts performance); *Redundant* if $|\Delta| < 5$ (indicating little or no impact on performance). Fig-

Model	Stage 1 (%)	Stage 2 (%)			Stage 3 (%)			Joint (%)		
		Reset	Carry	Feedback	Reset	Carry	Feedback	Reset	Carry	Feedback
Gemini 3.1 Pro	77.8	62.7	72.9	85.9	31.9	33.4	33.3	20.6	27.0	27.8
GPT-5.4	70.9	52.6	68.3	81.7	29.5	31.0	30.3	14.5	21.7	21.5
Qwen3-VL-32B	51.0	39.9	31.4	28.4	22.0	22.3	30.0	6.4	4.3	4.6
Qwen3.5-27B	53.6	39.9	40.5	39.9	23.2	24.1	30.2	6.7	8.1	7.7
InternVL3.5-14B (n)	46.7	26.5	24.5	9.8	2.1	2.5	12.5	0.4	0.4	0.7
InternVL3.5-14B (t)	41.2	26.5	26.8	16.7	3.2	3.0	17.8	0.8	0.6	0.8

Table 2: Overall results from the 3-stage $\times$ 3-mode analysis. S1 and S2 scores refer to accuracy, and S3 score refers to mIoU. Midrules separate closed-source models (top), 27–32B open-source models (middle), and the 14B InternVL configurations (bottom). S1 has a single column because there is no prior context to carry.

	S1 $\rightarrow$ S2 (%)		S2 $\rightarrow$ S3 (%)	
Model	S2	*c-S2*	mIoU	*c-mIoU*
Gemini 3.1 Pro	72.9	**89.5*****	33.4	33.0
GPT-5.4	68.3	**87.6*****	31.0	29.7
Qwen3.5-27B	40.5	**46.3***	24.1	23.4
Qwen3-VL-32B	31.4	30.8	22.3	23.1
InternVL-14B (t)	26.8	22.2	3.0	**5.0***
InternVL-14B (n)	24.5	21.7	2.5	3.9

Table 3: Cascade analysis. The S2 and mIoU columns report unconditional performance; c-S2 and c-mIoU are the conditional metrics. Bold marks conditional metrics that differ significantly from their unconditional counterparts. * $p<.05$, ** $p<.01$, *** $p<.001$.

ure 3 shows the results when audio, subtitle, and both audio and subtitle are removed, respectively. The audio and subtitle modalities are redundant for temporal localization in more than 60% of videos, indicating that removing them has little or no effect on performance. However, these modalities are complementary in approximately 17%–21% of the videos, where their removal degrades localization performance. In contrast, for about 13%–17% of the videos, removing these modalities improves performance, suggesting that they introduce distracting or noisy information. Next, we provide further insights into modality dependency.

Removing both audio and subtitles is most damaging to Games and Music. Figure 4 shows the per-category proportions of videos in which these modalities play complementary or distracting roles, respectively. As shown in the top five rows, incorporating both modalities improves performance for slightly more than 30% of the videos in the Game and Music categories. In the Game category, subtitles and audio cues such as commentary, callouts, and on-screen score updates provide strong temporal signals that help localize relevant moments more precisely. Similarly, in the Music category, the alignment between audio content (e.g., songs or background music) and on-screen textual cues such as lyrics, captions, and titles substantially improves temporal localization accuracy. In these cases, removing either modality often leads to noticeable performance degradation, indicating that the modalities provide complementary information. In contrast, for other video categories, visual modality alone is distinctive enough for most videos. In some cases, audio and subtitle modalities are only weakly correlated with the target moments. For example, videos in the Wedding and Celebration categories are often accompanied by ambient background noise, music, or unrelated conversations that provide limited temporal cues and may even negatively affect localization performance. Similarly, in the Dance category, the accompanying music often persists throughout the entire video and provides limited cues for temporal localization.

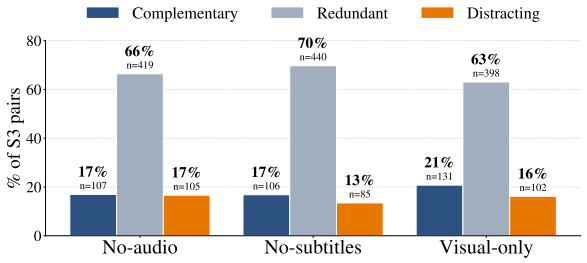

Figure 3: Share of S3 video pairs (n=631) in each modality role (Complementary, Redundant, Distracting; defined in §4.4) for the three ablations. The mIoU change Δ is averaged across three models before binning. Complementary and Distracting shares are roughly balanced and cancel in the mean mIoU.

Non-Latin-script countries show a higher Distracting share for audio. 29% of the videos from Cambodia, Myanmar, and Thailand have audio rated as *Distracting*, compared with 14% for Indonesia, Malaysia, the Philippines, and Vietnam,

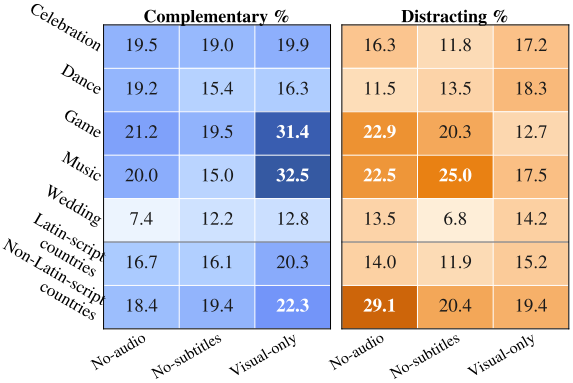

Figure 4: Modality roles across all S3 video pairs, segmented by category (top 5 rows) and country writing-system clusters (bottom 2 rows).

as shown in Figure 4. Subtitles follow the same pattern with a weaker effect. The *Distracting* cases are dominated by clips featuring traditional solo instruments that play continuously across the sub-event. Thus, the audio lacks a reliable temporal cue for the moment the question asks about, acting more as noise than as signal.

4.5 Human Evaluation

	S1 (%)		S2 (%)	
Rater	Native	Non-native	Native	Non-native
Expert	58.4	21.3	51.9	28.7
Non-expert	39.2	27.9	45.2	31.5

Table 4: Human evaluation accuracy (%) on S1 and S2, by rater group (Expert vs Non-expert) and role (Native vs Non-native). Each cell averages over the seven country-level mean scores. More in the Appendix.

We test whether CMB requires country-specific cultural knowledge, not generic regional cues, with a 14-rater human study (two raters per country). All raters are native speakers of their country's primary language and fluent in English at university level. They are categorized into two groups: *Expert*, who are actively involved in cultural studies through research or co-curricular work, and *Non-expert*, who are not. Each rater answered S1 and S2 in two roles. As *Native*, they evaluated concepts from their own country. As *Non-native*, they evaluated concepts from a neighboring country. The seven countries form a cyclic order. One rater per country received concepts from the next country; the other, from the country after that. About 20% of items were double-rated for inter-rater reliability. The study covers all 306 concepts. We focus on two contrasts that we report below.

Native vs Non-native: cultural knowledge is country-specific. Both rater groups perform far worse on a neighboring country's concepts than on their own, as shown in Table 4: *Expert* raters drop by 37 points on S1 and 23 on S2, while *Non-expert* raters drop by 11 and 14, respectively, from a lower starting point. The sharpest case is that *Expert* raters on a neighboring country score below the 25% chance level on S1, meaning the semantic-similarity distractors actively mislead even *Expert* raters. This indicates that CMB tests for country-specific knowledge rather than regional cues.

Expert vs Non-expert: the expertise effect is country-conditional. On their own country's concepts, *Expert* raters outperform *Non-expert* raters by 19 points on S1 and 7 points on S2, as shown in Table 4. However, the ordering flips for a neighboring country's concepts: *Expert* raters score slightly lower than *Non-expert* raters on both S1 and S2. The cultural knowledge that helps *Expert* raters at home misleads them abroad. We also report the model mean across the six VLMs as a human-ceiling reference. On S1, the model mean is 57%, matching *Expert* raters and beating *Non-expert* raters by 18 points. On S2 (Carry mode), it drops to 44%, level with *Non-expert* raters and 8 points below *Expert* raters. The model's average performance is therefore stage-specific: Expert-level at naming, Non-expert-level at visual recognition, consistent with the partial S1→S2 cascade.

5 Conclusion

Cultural video understanding is not a single capability but three, and they do not compose in current VLMs. CMB's cascade shows that correctly naming a concept helps only some models recognize it on video, and correctly recognizing it rarely helps any locate its sub-events in time. The bottleneck lies not in any single ability but in the connections between them, and which connection breaks depends on the model. A 14-rater human study confirms that the test measures country-specific cultural knowledge rather than regional cues, with model performance reaching Expert level at naming but only Non-expert level at visual recognition. CMB therefore acts as a diagnostic harness, pointing to model-specific priorities: cultural knowledge for the 14B InternVL configurations, visual recognition for the Qwen variants, and temporal localization for all six models.

Limitations

Cultural scope. CMB covers seven SEA countries and five categories (Celebration, Dance, Game, Music, Wedding); other SEA countries (Brunei, Laos, Timor-Leste, Singapore) and traditions outside these categories are out of scope, reflecting the availability of qualified Local Annotators and of high-quality source videos. We hope the CMB framework serves as a template for broadening cultural assessment to other underrepresented regions (e.g., Africa, South America, Oceania).

Scale vs. depth. CMB prioritizes expert curation (306 concepts, 624 source videos) over scale, making it suitable for benchmark evaluation but not for VLM training or fine-tuning. Further scaling would require semi-automated pipelines or expanded annotator capacity.

Long-tailed distribution. Three distributional imbalances should temper per-category and per-country comparisons. i) Concept counts are uneven across categories (Celebration: 91 vs. Wedding: 11) and across countries (Vietnam: 55 vs. Cambodia: 28), reflecting the natural breadth of each category and the volume of online video content in each country. ii) The Set B S3 pool is shaped differently from the concept pool: concepts with many ritual sub-events (Wedding, $\sim$13 spans per concept) generate many S3 questions, while object-like concepts (Music instruments, $<$1 span per concept) generate few. iii) Sub-national diversity (minority languages, regional variants, diaspora communities) is also under-represented; per-category aggregate scores should therefore be read as averages over uneven concept and span counts, not as uniform coverage of the underlying tradition space.

Modality ablation scope. The modality ablation reruns three of the six evaluated models (Gemini 3.1 Pro, Qwen3-VL-32B, Qwen3.5-27B) due to API and compute-budget constraints; the Complementary / Redundant / Distracting role labels are derived from these three models and may differ for GPT-5.4 or InternVL.

Static-bias diagnostic. CMB does not include a single-frame baseline that would isolate how much S2 accuracy depends on a static cue versus the temporal motion in the moment. The low S2$\rightarrow$S3 transfer (§4.3) provides indirect evidence that models rely on cues beyond concept identity, but a direct single-frame comparison is left for future work.

Source-video reproducibility. CMB references source videos by YouTube URL. If an uploader later removes their video, it becomes inaccessible to downstream users. We maintain versioned releases of CMB so each evaluation can be tied to a specific snapshot, but the underlying video itself cannot be recovered once it is removed from YouTube. See Ethical Considerations for the full release and opt-out protocol.

Ethical Considerations

Annotator compensation. The two co-author Cultural Annotators (Vietnam and Malaysia) contributed without compensation. The remaining five Cultural Annotators were recruited through the authors' academic networks at Southeast Asian universities and compensated fairly, with rates exceeding the local minimum wage and living wage standards in their respective countries, paid in each annotator's local currency. Local Annotators received a flat per-concept honorarium for the consensus and video-filtering passes, calibrated to the same minimum-wage and living-wage standards. Reviewers are co-authors and received no separate compensation for the review pass. The 14 human-study raters were each paid a fixed honorarium calibrated to the same minimum-wage and living-wage standards as the Cultural Annotators. They gave informed consent before participating and could withdraw at any time; no personally identifying information was collected, and rater identifiers in the released metadata are pseudonymous.

Source videos, release format, and opt-out. All 624 source videos (Set A for S2 and Set B for S3) were publicly hosted on YouTube at the time of annotation. To respect the platform's terms of service and the uploader's copyright, CMB will be released on HuggingFace as a JSON file containing YouTube video IDs, S2 moment timestamps for video A, and S3 ground-truth spans for video B, accompanied by evaluation scripts; we do not redistribute raw video bytes. The CMB annotations and evaluation scripts are released under CC BY-NC-SA 4.0 for non-commercial research use. CMB contains no personally identifiable information: the released metadata covers only cultural concepts, timestamps, and annotation records. Users obtain the underlying videos directly from YouTube under YouTube's terms. Uploaders who wish to have a video removed from CMB may contact the authors at the project email listed in the GitHub README;

we will remove the corresponding entries within 7 days and reissue a versioned release.

Cultural sensitivity and integrity. The primary ethical concern in this work is the potential for misrepresentation or commodification of cultural heritage. To mitigate this, we use a human-in-the-loop curation process involving local experts from each Southeast Asia country. These experts verified that the questions and annotations accurately reflect the cultural semantics of the artifacts and rituals depicted, minimizing the risk of imposing external or Western-centric interpretations on the data. We caution users that culture is fluid and multifaceted: our annotations represent a valid interpretation, not necessarily the only interpretation.

References

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The K-Viscuit benchmark with human-VLM collaboration. *Preprint*, arXiv:2406.16469.

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. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Xinyu Chen, Yunxin Li, Haoyuan Shi, Baotian Hu, Wenhan Luo, Yaowei Wang, and Min Zhang. 2025. VideoVista-CulturalLingo: 360° horizons-bridging cultures, languages, and domains in video comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27102–27128, Vienna, Austria. Association for Computational Linguistics.

Anna-Maria Christodoulou, Kyrre Glette, Olivier Lartillot, and Alexander Refsum Jensenius. 2025. Musiqal: A dataset for music question–answering through audio–video fusion. *Transactions of the International Society for Music Information Retrieval*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv e-prints*, arXiv:2501.12948.

Md Fahim, Md Sakib Ul Rahman, Akm Moshiur Rahman, Md Farhan Ishmam, Md Tasmim Rahman, Fariha Tanjim Shifat, Fabiha Haider, and Md Farhad Alam Bhuiyan. 2026. Banglaprotha: Evaluating vision language models in underrepresented long-tail cultural contexts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1159–1169.

Ali Faraz, Akash, Shaharukh Khan, Raja Kolla, Akshat Patidar, Suranjan Goswami, Abhinav Ravi, Chandra Khatri, and Shubham Agarwal. 2026. Indicvisionbench: Benchmarking cultural and multilingual understanding in VLMs. In *The Fourteenth International Conference on Learning Representations*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.

Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. 2024. Trace: Temporal grounding video llm via causal event modeling. *ArXiv*, abs/2410.05643.

Hsin-Yi Hsieh, Shang Wei Liu, Chang Chih Meng, Shuo-Yueh Lin, Chen Chien-Hua, Hung-Ju Lin, Hen-Hsen Huang, and I-Chen Wu. 2025. TaiwanVQA: A benchmark for visual question answering for Taiwanese daily life. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 57–75, Abu Dhabi, UAE. Association for Computational Linguistics.

Xilin Jiang, Qiaolin Wang, Junkai Wu, Xiaomin He, Zhongweiyang Xu, Yinghao Ma, Minshuo Piao, Kaiyi Yang, Xiuwen Zheng, Riki Shimizu, Yicong Chen, Arsalan Firoozi, Gavin Mischler, Sukru Samet Dindar, Richard Antonello, Linyang He, Tsun-An Hsieh, Xulin Fan, Yulun Wu, and 14 others. 2026. Avmeme exam: A multimodal multilingual multicultural benchmark for llms' contextual and cultural knowledge and thinking. *Preprint*, arXiv:2601.17645.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Shuowei Li, Yuming Zhao, Parth Bhalerao, and Oana Ignat. 2026. Maven a multi-agent framework for multicultural text-to-video generation. *Preprint*, arXiv:2605.16716.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024b. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyu Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2025. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. *ArXiv*, abs/2505.20715.

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*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, 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, Wayne Lau, Choon Meng Tan, and 12 others. 2025. Sea-lion: Southeast asian languages in one network. *Preprint*, arXiv:2504.05747.

Yuxiang Nie, Han Wang, Yongjie Ye, Haiyang Yu, Weitao Jia, Tao Zeng, Hao Feng, Xiang Fei, Yang Li, Xiaohui Lv, Guozhi Tang, Jingqun Tang, Jinghui Lu, Zehui Dai, Jiacong Wang, Dingkang Yang, An-Lan Wang, and Can Huang. 2025. Chinesevideobench: Benchmarking multi-modal large models for chinese video question answering. *Preprint*, arXiv:2511.18399.

OpenAI. 2025. Gpt-5.

Tushar Pranav, Eshan Pandey, Lyka Diane Bala Austria, Aman Chadha, Indriyati Atmosukarto, and Donny Cheng Lock Soh. 2025. Rice-vl: Evaluating vision-language models for cultural understanding across asean countries. *ArXiv*, abs/2512.01419.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22238–22254, Suzhou, China. Association for Computational Linguistics.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. GIMMICK – globally inclusive multimodal multitask cultural knowledge benchmarking. *Preprint*, arXiv:2502.13766.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark & model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20009–20033, Suzhou, China. Association for Computational Linguistics.

Darshan Singh, Arsha Nagrani, Kawshik Manikantan, Harman M. Singh, Dinesh Tewari, Tobias Weyand, Cordelia Schmid, Anelia Angelova, and Shachi Dave. 2026. Minerva-cultural: A benchmark for cultural and multilingual long video reasoning.

Atnafu Lambebo Tonja, Srija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A. Mohamed, Debela Desalegn Yadeta,

Negasi Haile Abadi, Abigail Oppong, Nnaemeka Casmir Obiefuna, Idris Abdulmumin, Naome A Etori, Eric Peter Wairagala, Kanda Patrick Tshinu, Imanigirimbabazi Emmanuel, Gabofetswe Malema, Alham Fikri Aji, David Ifeoluwa Adelani, and Thamar Solorio. 2026. Afri-mcqa: Multimodal cultural question answering for african languages. *Preprint*, arXiv:2601.05699.

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.

Nikhil Varimalla, Yunfei Xu, Arkadiy Saakyan, Meng Wang, and Smaranda Muresan. 2025. Videonorms: Benchmarking cultural awareness of video language models.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, {Henok Biadglign} Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa {Al Ghallabi}, Mihail Mihaylov, Chao Qin, {Abdelrahman M.} Shaker, Mike Zhang, {Mahardika Krisna} Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 19565–19575. Publisher Copyright: © 2025 IEEE.; 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025 ; Conference date: 11-06-2025 Through 15-06-2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. 2025. Timesuite: Improving mllms for long video understanding via grounded tuning. *Preprint*, arXiv:2410.19702.

Weihua Zheng, Zhengyuan Liu, Tanmoy Chakraborty, Weiwen Xu, Xiaoxue Gao, Bryan Chen Zhengyu Tan, Bowei Zou, Chang Liu, Yujia Hu, Xing Xie, Xiaoyuan Yi, Jing Yao, Chaojun Wang, Long Li, Rui Liu, Huiyao Liu, Koji Inoue, Ryuichi Sumida, Tatsuya Kawahara, and 16 others. 2025. Mma-asia: A multilingual and multimodal alignment framework for culturally-grounded evaluation. *Preprint*, arXiv:2510.08608.

A Appendix

A.1 Human Evaluation Design

To quantify the cultural gap between current video-language models and SEA native speakers, we conduct a human evaluation on Stage 1 (text–MCQ) under two complementary roles: a within-country **Native** ceiling and a within-region but out-of-country **Non-native** floor. Each rater receives exactly the same S1 question text and answer options as the models did, so the human and model numbers are directly comparable on the same 306 concepts.

Human Rater pool. We recruit **14 qualified raters**, two per SEA country (Cambodia, Indonesia, Malaysia, Myanmar, Philippines, Thailand, Vietnam). Every rater serves in two roles: they are one of two *Local* raters for their own country, and one of two *Non-local* raters for exactly one other country. Consequently, each of the seven countries is evaluated by four distinct raters: two Native raters and two Non-native raters.

Non-local rotation. Non-local assignments follow a fixed $+1/+2$ rotation. Ordering the countries $\mathrm{KH}=0$, $\mathrm{ID}=1$, $\mathrm{MY}=2$, $\mathrm{MM}=3$, $\mathrm{PH}=4$, $\mathrm{TH}=5$, $\mathrm{VN}=6$, person a of country i rates country $(i+1) \bmod 7$ as Non-local, and person b of country i rates country $(i+2) \bmod 7$. This guarantees that each country receives Non-local ratings from two *different* home countries; the resulting pairings are shown in Table 5.

Split-and-overlap rating structure. Within each country's 306-concept universe, both Local raters split the concepts: 20% of the concepts form an **overlap subset** rated by *both* Locals, while the remaining 80% is partitioned equally, with each Local rating 40% solo. The two Non-local raters for that country follow the same split on the same 20% overlap subset. Each non-overlapping concept therefore receives two ratings (one Native and one Non-native); each overlapping concept receives four ratings (two Native and two Non-native).

Overlap-subset selection. The 20% overlap subset is drawn by stratified random sampling within each country–category cell using $\max(1, \mathrm{round}(0.2\, n_{\mathrm{cat}}))$, with a fixed random seed. The selection is frozen before annotation begins and released as part of the benchmark for reproducibility; raters are not told which items are overlap versus solo.

Aggregation. For each concept, the *Native score* is the mean of available Local ratings, and the *Non-native score* is the mean of available Non-local ratings (binary $\{0, 1\}$ on solo items; continuous $\{0, 0.5, 1\}$ on overlap items). Per-country Native and Non-native floors are the mean of per-concept scores over that country's concepts. Model numbers reported alongside use the same per-country concept subsets, so all three quantities (Native / Non-native / Model) are computed on the same items.

Workload. Per-rater load ranges from 42 to 64 S1 MCQs (approximately 45 min per rater), producing approximately 732 ratings total ($\sim$9.3 person-hours aggregate).

Acknowledged structural asymmetry. Under any design with two Non-local raters per country, the Latin-script vs. non-Latin-script country imbalance (4 vs. 3) forces exactly one country to deviate from the "1 Latin + 1 Local Non-local" pattern. We use the *+1/+2 rotation*, as shown in Table 5, in which each country's two Local raters are also paired with the next country (+1) and the country after that (+2) in a cyclic order Under this rotation, the deviation lands on **Myanmar**, whose two Non-locals are both from the Latin-script cluster (MY and ID). Because Myanmar is itself a non-Latin-script country, both are still cross-cluster raters from its perspective; six of seven countries receive mixed (Latin + Local) Non-local perspectives, and Myanmar's Outsider floor is computed entirely from Latin-script raters. Script-cluster sub-analyses (same-cluster vs. different-cluster Non-native floor) are available on the six mixed countries.

A.2 Per-country inter-rater agreement

Table 7 reports 4-way Fleiss' κ on S1 and S2 for each of the seven countries' overlap items ($n = 6$–12 per country). Per-country values are noisy at this sample size but follow the pattern observed in the pooled analysis: S2 κ values are positive across all seven countries, while S1 κ values fluctuate around zero. This is consistent with S2's moment-grounded design reducing ambiguity relative to S1's abstract symbolic prompts. The near-chance pooled S1 κ is consistent with, rather than evidence against, the design: cultural concepts of a neighboring country are hard even for trained native raters, which is exactly the gap the Outsider Filter (§3.3) is intended to harden.

A.3 Modality Ablation Details

We re-ran the three lowest-cost models (Gemini 3.1 Pro, Qwen3-VL-32B, Qwen3.5-27B) on the modality ablations; the remaining three (GPT-5.4, InternVL3.5-14B in both configurations) were excluded due to API and compute budget.

A.4 Modality Role by Country Cluster

The *Distracting* role shares splits sharply along the writing-system clusters. For audio, the cluster gap is $+15$ points (z=3.84, p<.001 from a two-proportion z-test); for subtitles, the gap is smaller ($+8.5$ points) but points in the same direction.

The Distracting concepts in the non-Latin-script cluster are dominated by traditional music concepts: of 26 such concepts, 8 are music (47% of non-Latin music concepts vs 4% in Latin), and 10 are traditional games. The music subset features solo instruments (Saung, Sralai, Chhing, Sor Duang, Tapone, Maung, Kong Chai, Tone) whose audio plays uniformly across the clip, providing no temporal signal for the specific sub-event the question asks about. The pattern is also consistent with audio being anchored less reliably for underrepresented languages in pretraining, though we cannot verify the training-side mechanism because the models' training mixtures are not disclosed.

A.5 Human Raters

Expert raters comprise 2 Ph.D. holders, 3 Ph.D. candidates, 1 B.A. graduate, and 1 B.A. student; 5 have published cultural-research papers, and 2 have documented hands-on cultural fieldwork experience. Non-expert raters are native speakers without cultural-research publications or formal cultural-studies training; they were screened only on native fluency and confirmation that they had not previously seen CMB material.

A.6 More on Implementation Details

Joint score: zero-imputation and worked example. In the Joint formula $\text{Joint} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\text{S1}_i\} \cdot \mathbf{1}\{\text{S2}_i\} \cdot \text{IoU}_i$, IoU_i is the per-concept S3 IoU, set to zero if the model produces no S3 prediction for concept i. An equivalent form makes the components explicit: $\text{Joint} = \frac{n_{\text{S1}\cap\text{S2}}}{N} \cdot \overline{\text{IoU}}_{\text{S1}\cap\text{S2}}$, where $n_{\text{S1}\cap\text{S2}}$ is the number of concepts with both S1 and S2 correct, and $\overline{\text{IoU}}_{\text{S1}\cap\text{S2}}$ is the mean S3 IoU over those concepts. For example, with $n_{\text{S1}\cap\text{S2}} = 138$ and $\overline{\text{IoU}}_{\text{S1}\cap\text{S2}} = 0.32$, $\text{Joint} = (138/306) \cdot 0.32 = 14.4\%$.

Country (target)	Cluster	Local raters	Non-local raters	Non-local clusters
Cambodia	Local	KH-a, KH-b	VN-a, TH-b	Latin + Local
Indonesia	Latin	ID-a, ID-b	KH-a, VN-b	Local + Latin
Malaysia	Latin	MY-a, MY-b	ID-a, KH-b	Latin + Local
Myanmar	Local	MM-a, MM-b	MY-a, ID-b	**Latin + Latin**
Philippines	Latin	PH-a, PH-b	MM-a, MY-b	Local + Latin
Thailand	Local	TH-a, TH-b	PH-a, MM-b	Latin + Local
Vietnam	Latin	VN-a, VN-b	TH-a, PH-b	Local + Latin

Table 5: Annotator assignment under the $+1/+2$ rotation. Each of the seven countries is evaluated by two Local raters (own-country) and two Non-local raters drawn from two different other countries. Six of seven countries receive Non-local ratings from both script clusters (Latin-script: Indonesia(ID)/Malaysia(MY)/Philippines(PH)/Vietnam(VN); non-Latin-script: Cambodia(KH)/Myanmar(MM)/Thailand(TH)); Myanmar is the structural exception discussed in the text.

#	Annotator	Local (60% of own)	Non-local country	Non-local (60%)	Total
1	KH-a	17 (of 28)	Indonesia	29	46
2	KH-b	17	Malaysia	26	43
3	ID-a	29 (of 48)	Malaysia	26	55
4	ID-b	29	Myanmar	22	51
5	MY-a	26 (of 44)	Myanmar	22	48
6	MY-b	26	Philippines	31	57
7	MM-a	22 (of 37)	Philippines	31	53
8	MM-b	22	Thailand	25	47
9	PH-a	31 (of 52)	Thailand	25	56
10	PH-b	31	Vietnam	33	64
11	TH-a	25 (of 42)	Vietnam	33	58
12	TH-b	25	Cambodia	17	42
13	VN-a	33 (of 55)	Cambodia	17	50
14	VN-b	33	Indonesia	29	62

Table 6: Per-annotator S1 workload under the split-and-overlap scheme. Each annotator rates 60% of their own country's concepts (40% solo + 20% overlap) plus 60% of one Non-local country under the same split. Per-annotator range: 42–64 MCQs ($\sim$32–48 min); total ratings collected: $\sim$732.

Country	S_1 κ	S_1 n	S_2 κ	S_2 n
Cambodia	$+0.005$	6	$+0.120$	6
Indonesia	$+0.079$	11	$+0.175$	11
Malaysia	$+0.056$	9	$+0.193$	9
Myanmar	-0.172	8	$+0.099$	8
Philippines	-0.202	10	$+0.225$	10
Thailand	-0.052	9	$+0.152$	9
Vietnam	$+0.106$	12	$+0.024$	12
Pooled (Fleiss, all countries)	$+0.011$	65	$+0.166$	65

Table 7: Per-country 4-way Fleiss' κ on S1 and S2 across the 20% overlap items. Pooled row uses all 65 overlap items together. S2 κ is positive across all seven countries; S1 κ fluctuates around zero, consistent with S2's visual grounding reducing rater disagreement relative to S1's abstract symbolic prompts.

The main S1/S2/S3 experiments all use a single query form (Latin-common), so the split is a post hoc property of the countries, not an experimentally manipulated condition.

Ablation	n pairs	$\overline{\Delta}$ mIoU points	p
No-audio	631	-0.02	.55
No-subtitles	631	-0.65	.10
Visual-only	631	-0.81	.045

Table 8: Aggregate mIoU shift per modality ablation, pooled across the three ablation-instrumented models (Gemini 3.1 Pro, Qwen3-VL-32B, Qwen3.5-27B). p-values from paired tests against the baseline. Per-pair role classifications derived from these shifts appear in Figure 3.

Cluster	*Distr.* audio (%)	*Distr.* subs (%)
Latin-script (ID/MY/PH/VN)	14.0	11.9
non-Latin-script (KH/MM/TH)	29.1	20.4

Table 9: Distracting role share (%) by writing-system cluster, pooled across Gemini 3.1 Pro, Qwen3-VL-32B, and Qwen3.5-27B over the same 631 S3 pairs as Figure 3. Non-Latin-script cluster shows a $+15$ pp audio gap ($p < .001$) and a $+8.5$ pp subtitle gap.

Implementation Details. In all evaluated models, each frame is resized to 224×224 before being passed to the model. S1/S2 prompts use the four-way MCQ template shown in Figure 1; S3 prompts ask the model to return the predicted $[\hat{t}_{\text{start}}, \hat{t}_{\text{end}}]$ span in seconds.

API and compute budget. Closed-source inference was conducted through the Gemini and OpenAI APIs. The total API cost for the full experiment grid (3 stages $\times$ 3 modes on S1/S2 plus three modality ablations on S3) was approximately USD 700, split between Gemini and GPT. Open-

source models were run on a single multi-GPU node.

A.7 Model-specific intervention priorities

Beyond diagnosis, CMB suggests model-specific intervention priorities. Adding cultural knowledge is the priority for the smallest models, where naming itself is the limit. Improving visual recognition is the priority for the two Qwen models, which name well but do not carry the name to a video moment. Temporal localization is a persistent bottleneck across all the models.

A.8 Conceptual Knowledge vs. Procedural Traditions

CMB stages probe two complementary kinds of cultural knowledge. **Conceptual Knowledge** (S1, S2) is knowing *what a concept symbolizes*: the meaning behind the name and its visual signature. **Procedural Traditions** (S3) is knowing *how the tradition unfolds in time*: which sub-events occur, in what order, and how they bind a recognizable moment. We use this distinction to motivate the textual-only LLM audits: R1's S1 pass guards Conceptual Knowledge items against text-level shortcuts, and its S3 pass guards Procedural Traditions items against under-specified sub-event descriptions. Visual judgment for both kinds remains exclusively human.

A.9 Annotator recruitment and screening

Procedural rules. Every rater completes their Non-local batch before their Local batch to avoid own-country priming. Question text is identical to the model prompt. Raters cannot skip items or revisit previous answers after submission. No external aids: raters answered each S1 item from memory and regional priors only; web search, LLM assistants, and consultation with peers were prohibited. The rating interface required raters to confirm this rule at the start of each session.

Local Annotators. Recruited via the authors' academic contacts at universities in each of the seven Southeast Asia countries. Inclusion criteria: i) born and raised in the country, ii) native fluency in the country's primary language, iii) general cultural exposure documented through a short intake survey on festivals, family traditions, and regional dialects, iv) willingness to reach a unanimous consensus with one or two peers from the same country. Two Local Annotators per country, three for the Philippines and Myanmar, to cover their wider linguistic diversity.

Cultural Annotators. A Cultural Annotator is a Local Annotator who additionally meets one of the following criteria: i) authored or co-authored a publication in cultural studies, anthropology, ethnomusicology, or adjacent fields covering the country's traditions; ii) sustained co-curricular involvement (cultural-society leadership, festival organization, or museum work) verified by two reference letters. Two Cultural Annotators are co-authors (Vietnam, Malaysia); the remaining five were recruited through the same academic networks and screened on a three-concept calibration task before formal annotation began.

Reviewers. Non-local researchers, all co-authors, were selected based on documented Southeast Asian cultural exposure, either through prior peer-reviewed publications on SEA culture or at least 6 months of fieldwork in the region. Reviewers are explicitly not permitted to be a Local Annotator for the country they are reviewing.

14-rater human-study pool. Recruited via the authors' extended network and a public call distributed through partner cultural-studies departments. Each rater was screened on i) native fluency, ii) self-reported expert vs non-expert status (verified by follow-up), and iii) confirmation that they had not previously seen CMB material.

A.10 Released record schema

Each released CMB record corresponds to one of the 306 concepts and carries the following fields:

`concept_id` Stable identifier, format `<country>-<category>-<NN>`.

`country` ISO-3166 alpha-2 code (one of KH, ID, MM, MY, PH, TH, VN).

`category` One of {Celebration, Dance, Game, Music, Wedding}.

`concept_name` Concept name in native script with a Latin transliteration where applicable.

`description` English-language symbolic description used as the stem of the S1 and S2 questions.

`s1_options` List of four candidate name strings.

`s1_correct_index` Integer in $\{0, 1, 2, 3\}$.

s1_distractor_origin Per-distractor flag in {same-country, different-country}.

video_A_url, video_A_s2_moment YouTube URL and (start_s, end_s) timestamps for the S2 source clip.

s2_options Four (video_url, start_s, end_s) triples, one correct and three semantic-similarity distractors.

s2_correct_index Integer in $\{0, 1, 2, 3\}$.

video_B_url, video_B_s3_span YouTube URL and (start_s, end_s) ground-truth span for S3, with video_B $\neq$ video_A.

s3_cue Textual description of the target sub-event used as the S3 prompt.

writing_cluster Derived flag in {Latin, Local}.

audio_role Derived label in {Complementary, Redundant, Distracting} computed from the mean Δ mIoU of the three modality-instrumented models (Gemini 3.1 Pro, Qwen3-VL-32B, Qwen3.5-27B); see §4.4.

annotator_role_id Anonymised Cultural-Annotator identifier (one of seven).

review_outcome One of {passed, rewritten, discarded-and-replaced} from the Outsider Filter.

A.11 Specialized temporal-localization models on S3

To check whether models designed specifically for video temporal grounding outperform the general-purpose VLMs on S3, we evaluate TimeSuite (Zeng et al., 2025), TRACE (Guo et al., 2024), and MUSEG (Luo et al., 2025) on the same 306 S3 items. All three models read only the S3 textual cue plus the corresponding source video B; no S1 or S2 information is provided. Table 10 reports mIoU per category and overall.

	TimeSuite mIoU (%)	**TRACE** mIoU (%)	**MUSEG** mIoU (%)
Celebration	9.5	8.8	22.4
Dance	3.9	2.0	15.5
Game	11.2	6.0	21.6
Music	10.6	18.6	26.4
Wedding	10.5	13.5	13.8
Overall	**9.2**	**5.9**	**19.3**

Table 10: S3 temporal-localization mIoU of specialized temporal-grounding models on CMB, broken down by category. All three models receive the concept name and sub-event cue as text input and produce a free-form $[\hat{t}_{\text{start}}, \hat{t}_{\text{end}}]$ span on video B. They all evaluate at 1 fps with 224x224 resolution.

B Dataset Analysis

This appendix summarizes CMB's concept and question distributions, video and span durations, and the difficulty of the semantic-similarity distractors. All statistics are derived from the released CMB dataset: the per-concept analyses use n=306 items, and the S3 analyses use n=631 pairs across 318 unique source videos (some Set B source videos contribute more than one sub-event span).

Concept and question distribution. Figure 11 shows the per-country and per-category breakdowns. Each concept yields one S1 and one S2 question on Set A, so subplots (a) and (b) describe both the concept inventory and the Set A question pool. Subplots (c) and (d) show the Set B S3 question pool, which is larger because some source videos host more than one sub-event span. The distribution is dominated by Celebration (91) and Music (68); Wedding (11) is the smallest category, reflecting the relatively narrow space of distinct wedding traditions per country (typically one canonical ceremony name).

Question length and video duration. Figure 12 reports word-count statistics for the S1, S2, and S3 questions, video-duration histograms for Set A (S2 moments) and Set B (S3 source videos), and a per-category box plot of S3 sub-event span lengths. Questions are tightly clustered in length (S1: 26.4 words on average, S2: 26.4, S3: 23.5), reflecting the constraint that descriptions stay focused on a single symbolic concept. Set A moments are short (median 20 s) by design, while Set B source videos are substantially longer (median 316 s; mean 371 s, matching the $\sim$6.2 min figure, giving S3 a non-trivial localization span to search. Sub-event spans are short across all categories (median 13 s), with Game spans being the shortest because game rounds are typically self-contained and ritualized.

Distractor similarity. Figure 13 confirms that the three S1/S2 distractors selected per concept are systematically harder than a random sample from the same candidate pool. The selected distractors have a Sentence-BERT cosine similarity to the correct concept of median 0.70 (vs 0.60 for the non-selected candidate pool), so the model cannot fall

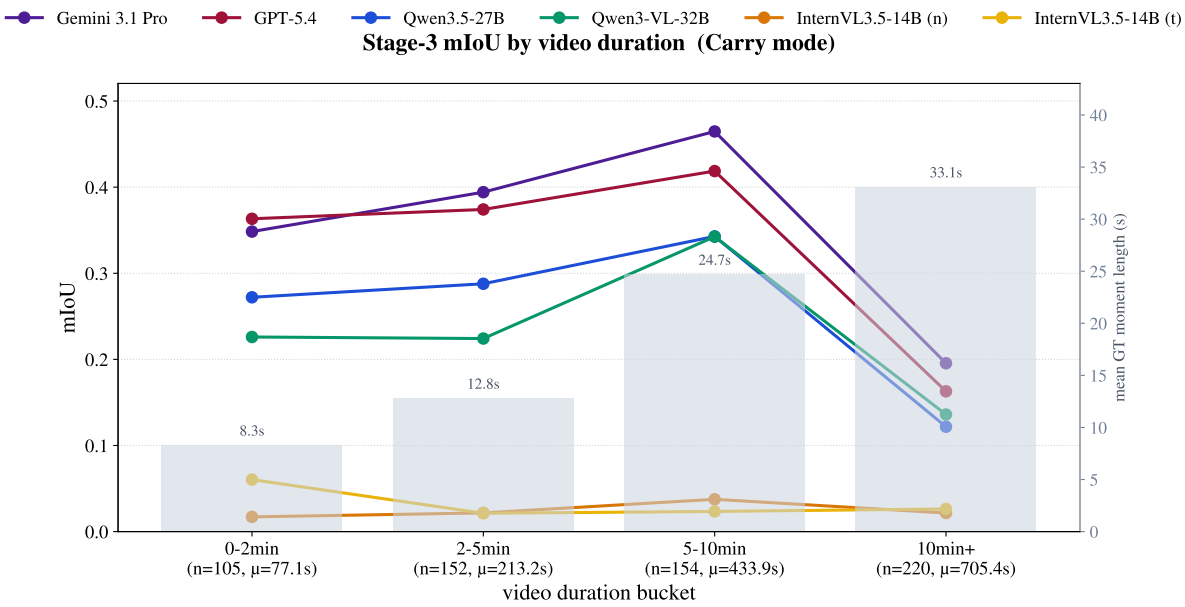

Figure 5: S3 mIoU under Carry mode, bucketed by Set B source-video duration. Below each bucket: n = number of S3 pairs in the bucket, and $\overline{\ell}$ = mean source-video duration. Markers annotate the mean ground-truth sub-event span in the bucket (right axis). All six models degrade with video length; closed-source models degrade more gracefully.

back on coarse keyword matching to discriminate the four options.

Summary statistics. Table 12 consolidates the headline numbers used elsewhere in the paper.

B.1 S1/S2 distractor sampling algorithm

Each S1 and S2 question has four options: one correct concept plus three distractors. The same three distractors are reused across S1 (text) and S2 (video moment), so the difficulty profile is identical at both stages. Distractors are drawn from a per-category candidate pool using a two-phase constrained greedy algorithm (Algorithm 1).

Inputs and constraints. For each category c with N_c concepts (Celebration: 91, Dance: 60, Game: 76, Music: 68, Wedding: 11), the algorithm consumes (i) a Sentence-BERT cosine-similarity matrix over the symbolic concept descriptions, (ii) a concept-to-country mapping, and (iii) a manually curated set of *ambiguity groups* $\mathcal{A}$: sets of concept names that refer to the same underlying ritual, dance form, game, or instrument under different country naming conventions. Across CMB we curated 17 ambiguity groups in total:

- **Celebration** (4): {Visak Bochea, Waisak, Wesak Day} (Vesak); {Lebaran, Hari Raya Aidilfitri} (Eid al-Fitr); {Makha Bucha, Meak Bochea}; {Idul Adha, Hari Raya Haji} (Eid al-Adha).
- **Dance** (2): {Tinikling, Mua Sap} (bamboo-pole dance); {Khon, Yama Zatdaw} (Ramayana dance drama).
- **Game** (8): {Congklak, Congkak} (mancala); {Engklek, Piko} (hopscotch); {Lompat Tali, Nhay Day} (jump rope); {Gasing, Danh Cu, Mark Khang} (spinning top); {Wau, Tha Dieu} (kite flying); {Guli, Kelereng} (marbles); {Capteh, Da Cau, Sipa} (shuttlecock); {Sepak Raga, Chinlone} (rattan ball).
- **Music** (3): {Kulintang (ID), Kulintang (MY), Kulintang (PH)}; {Rebab (ID), Rebab (MY)}; {Suling (ID), Suling (MY)}.
- **Wedding** (0): no ambiguity groups required.

Two concepts in the same group are never placed in the same MCQ.

Each key concept has three distractor slots. Slot 0 holds a same-country distractor (forcing fine-grained intra-country discrimination); slots 1 and 2 hold distractors from two *different* other countries, so no two distractors of the same item come from the same country. Each concept may be used as a distractor at most $N_{\max} = \lfloor 0.1 \cdot N_c \rfloor$ times across the category, capping per-concept re-use at $\sim$10%.

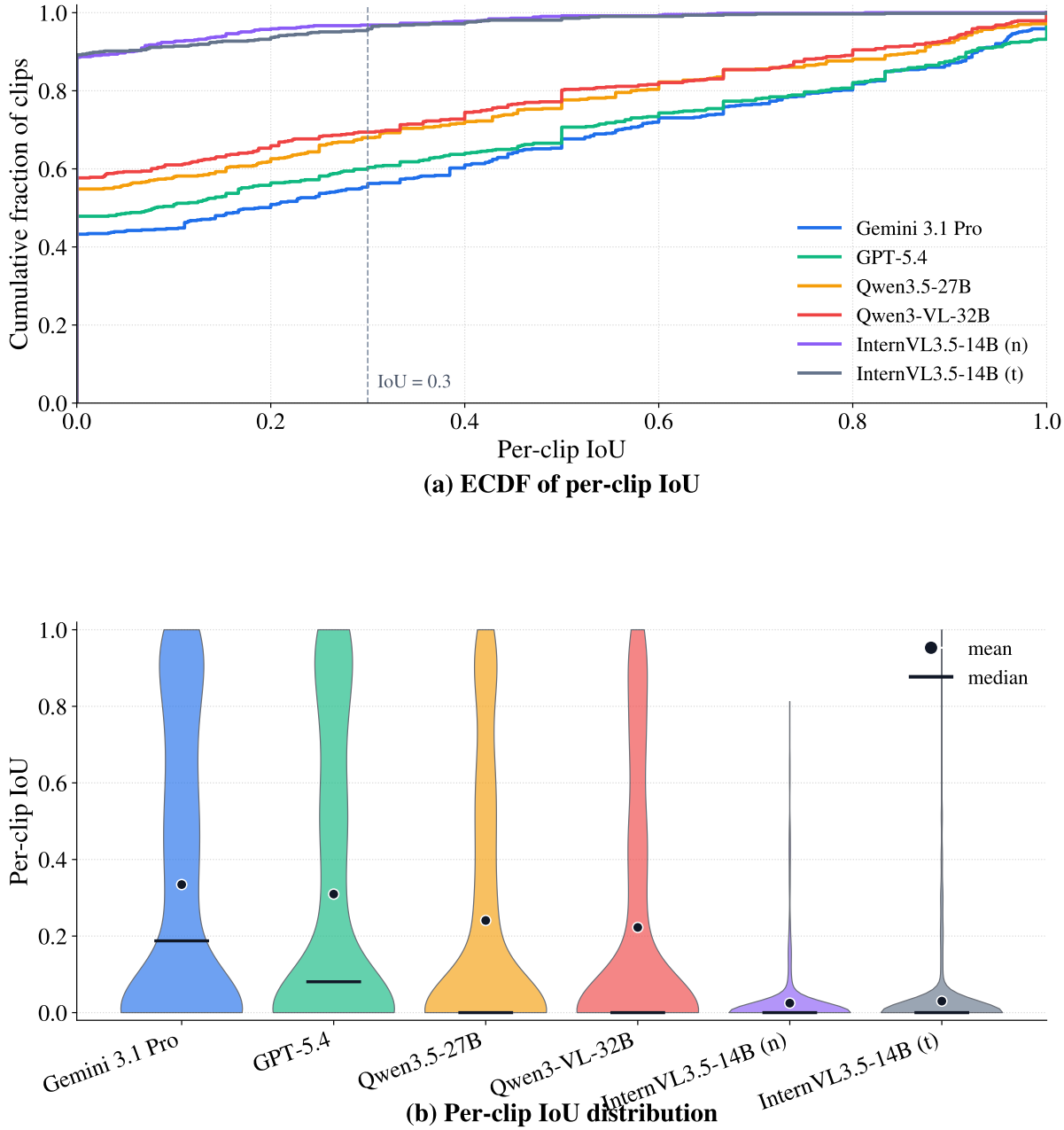

Figure 6: Per-clip S3 IoU distribution under Carry mode, across all 306 concepts. (a) ECDF of per-clip IoU; the dashed vertical line at IoU=0.3 reads off the IoU@0.3 hit rate per model. (b) Violin plot of the same distributions, with a strip of mean (filled dot) and median (horizontal bar) markers. A single mIoU number hides a long-tailed distribution with a heavy mass near zero for every model.

Algorithm. The assignment is computed in two phases. **Phase 1 (coverage guarantee):** concepts are sorted by *placability*—the number of valid open slots they fit into—and each unused concept is placed in the highest-similarity key with a compatible open slot, guaranteeing every concept appears as a distractor at least once. **Phase 2 (similarity-first greedy fill):** for each remaining open slot, the most-constrained key is processed first, and the slot is filled with the highest-similarity candidate that satisfies the validity predicate (self, repeat, $N_{\max}$, ambiguity, country). The fill terminates when all slots are populated or no further valid placement exists.

Validation. Across all five categories the algorithm produced no unfilled slots, no unused concepts, no over-cap usage, no country-constraint violations, and no ambiguity violations. The median Sentence-BERT cosine similarity between a correct concept and its three sampled distractors

Country	Category	Concepts
Cambodia	Celebration (5)	Bon Chrat Preah Neangkoal, Choul Chnam Thmey, Meak Bochea, Pchum Ben, Visak Bochea
	Dance (7)	Robam Bes Kravanh, Robam Chuon Por, Robam Kngaok, Robam Phlet, Robam Phloy Suoy, Robam Tep Apsara, Robam Yike
	Game (6)	Bos Angkunh, Chol Chhoung, Krob Moin, Leak Kanseng, Ok Chaktrong, Sey
	Music (9)	Chhing, Khaen, Khloy, Kong Thom, Kong Touch, Roneat Ek, Skor Thom, Sneng, Sralai
	Wedding (1)	Kar Khmer
Indonesia	Celebration (14)	Bulan Puasa, Festival Gamelan Yogyakarta, Festival Rawa Pening, Festival Tabot, Galungan, Idul Adha, Kuningan, Lebaran, Nyepi, Pasola, Perang Topat, Pesta Kesenian Bali, Sekaten, Waisak
	Dance (8)	Tari Barong, Tari Cendrawasih, Tari Gambyong, Tari Kecak, Tari Legong, Tari Merak, Tari Reog Ponorogo, Tari Saman
	Game (8)	Bola Gebok, Congklak, Egrang, Engklek, Galah Asin, Gasing, Kelereng, Lompat Tali
	Music (13)	Angklung, Gambang, Gong, Kecapi, Kempul, Keroncong, Kolintang, Kulintang, Rebab, Sape', Suling, Talempong, Tanjidor
	Wedding (5)	Pernikahan Adat Bali, Pernikahan Adat Batak, Pernikahan Adat Betawi, Pernikahan Adat Jawa, Pernikahan Adat Sunda
Malaysia	Celebration (16)	Bon Odori, Chinese New Year, Deepavali, Hari Gawai, Hari Merdeka, Hari Raya Aidilfitri, Hari Raya Haji, Hungry Ghost Festival, King's Birthday, Mid-Autumn Festival, Pesta Kaamatan, Ponggal, Qing Ming, Thaipusam, Thimithi, Wesak Day
	Dance (11)	Mak Yong, Tarian Asli, Tarian Bhangra, Tarian Inang, Tarian Joget, Tarian Kuda Kepang, Tarian Ngajat, Tarian Silat, Tarian Sumazau, Tarian Ulek Mayang, Tarian Zapin
	Game (10)	Baling Selipar, Batu Seremban, Capteh, Congkak, Galah Panjang, Guli, Lari Dalam Guni, Sepak Raga, Silat, Wau
	Music (6)	Gambus, Kompang, Kulintang, Rebab, Sape, Suling
	Wedding (1)	Perkahwinan Melayu
Myanmar	Celebration (13)	Hle Pwe, Kason Festival, Naga Hnit Thit Kuu Pwe Daw, Nat Pwe, Sanda Muni Pwe, Sein Taw Ya Paya Pwe, Shwedagon Pagoda Pwe, Taungbyon Nat Pwe, Tazaungdaing Mee Pone Pyan Pwe, Thadingyut Pwe, Thei Pone Cedi Pwe, Thingyan Festival, Tipitaka Chanting Ceremony
	Dance (10)	Kinnari Kinnara A Ka, Kyaukse Cin A Ka, Shan Daung A Ka, Si Mi Gwet A Ka, Ta Bin Daing A Ka, Thangyat A Ka, Yama Zatdaw A Ka, Yein A Ka, Zat Pwe A Ka, Zawgyi A Ka
	Game (7)	Chaw Taing Tet, Chinlone, Hlei Pyaing Pwe, Htoke Si Toe, Lethwei, Lun Swe Pwe, Paik Kyaw Chin
	Music (6)	Hne, Maung, Palwei, Pattala, Saung, Si Neh Wa
	Wedding (1)	Mingala Saung
Philippines	Celebration (25)	Ati-Atihan, Banga Festival, Buwan ng Wika, Carabao Festival, Dinagyang Festival, Kaamulan, Kadayawan, Kalilangan, Kasanggayahan, Kawayan Festival, MassKara, Obando Fertility Rites, Pahiyas Festival, Panagat Festival, Panagbenga Festival, Parada ng Lechon, Pintados-Kasadyaan, Pista ng San Juan, Sanduguan, Santacruzan, Sillag Festival, Sinakulo, Sinulog Festival, Suman Festival, Ylang-Ylang Festival
	Dance (7)	Binasuan, La Jota Moncadena, Pandanggo sa Ilaw, Sagayan, Singkil, Subli, Tinikling
	Game (12)	Agawang Buko, Bati-Cobra, Dama, Jolen, Langit-Lupa, Luksong Baka, Luksong Tinik, Patintero, Piko, Pukpok Palayok, Sipa, Tumbang Preso
	Music (7)	Banduria, Gabbang, Kudyapi, Kulintang, Luntang, Tamburin, Tumpong
	Wedding (1)	Kasal
Thailand	Celebration (4)	Constitution Day, End of Buddhist Lent, Makha Bucha, Songkran
	Dance (9)	Fon, Khon, Lakhon Nai, Manorah, Rabam Chao Na, Ram Krabi Krabong, Ram Muay, Ram Si Nuan, Ram Thoet Thoeng
	Game (17)	Chon Kho Khon, Dueng Nang, Ka Dot Chueak, Ka Fak Kai, Ka Toeng Ka Toi, Kaeng Kwien, Kaeng Ruea Bok, Kee Ma Song Muang, Kee Tu Klang Na, Ling Ching Lak, Mark Kep, Mark Khang, Morn Son Pa, Tee Gai, Tee Jub, Ti Luk Lo, Wing Piao
	Music (11)	Angkalung, Jakay, Khlui, Kong Chai, Kong Wong Lek, Pee, Pin Pia, Ranat, Sor Duang, Ta-pone, Tone
	Wedding (1)	Ngarn Taeng Ngarn
Vietnam	Celebration (14)	Gio To Hung Vuong, Hoi Dua Bo Bay Nui, Hoi Dua Voi Buon Don, Hoi Lim, Le Hoi Dan Toc Co Tu, Le Hoi Den Long Hoi An, Le Hoi Long Tong, Le Hoi Thap Ba Ponagar, Le Hoi Tien Cong, Le Phat Dan, Le Tich Dien, Tet Doan Ngo, Tet Nguyen Dan, Tet Trung Thu
	Dance (8)	Mua Cong Chieng, Mua Gay Senh Tien, Mua Khen, Mua Lan, Mua Non, Mua Phach Senh Tien, Mua Quat, Mua Sap
	Game (16)	Banh Chuyen Banh Dua, Bau Cua, Bit Mat Bat De, Bit Mat Dap Nieu, Da Cau, Danh Cu, Di Ca Kheo, Keo Co, Nem Con, Nhay Bao Bo, Nhay Day, Nhay Lo Co, O An Quan, Oan Tu Xi, Rong Ran Len May, Tha Dieu
	Music (16)	Cong Chieng, Dan Bau, Dan Da, Dan Day, Dan Nhi, Dan T'rung, Dan Tranh, Dan Ty Ba, Ken La, Khen, Phach Ca Tru, Phach Hat Van, Sao, Senh Tien, Song Loan, Trong
	Wedding (1)	Dam Cuoi

Table 11: The 306 CMB cultural concepts (Common Latin script), grouped by country and category. Counts in parentheses. Official Latin and native-script (Local) variants are released in the JSON dataset; see Appendix A.11.

Statistic	Value
Concepts	306
S1 questions (= concepts)	306
S2 questions (= concepts)	306
S3 questions (Set B, with repeats)	631
Unique Set A S2-moment videos	306
Unique Set B S3-source videos	318
S2 moment duration (Set A, n=304): median (IQR), seconds	20 (5–48)
S3 source duration (Set B, n=318): median (IQR), seconds	316 (173–540)
S3 sub-event span (n=631 pairs): median (IQR), seconds	13 (6–24)
S1 question length: mean words	26.4
S2 question length: mean words	26.4
S3 question length: mean words	23.5
Chosen distractor similarity: median	0.70
Non-chosen candidate similarity: median	0.60

Table 12: CMB dataset summary statistics. Set A counts are 1-to-1 with concepts (one S2-moment video per concept); Set B counts are larger because multi-event traditions contribute more than one S3 sub-event span.

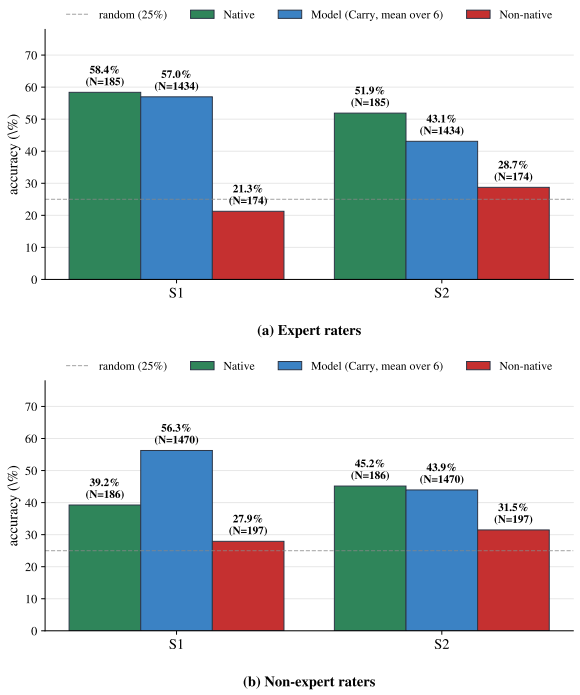

Figure 7: Human evaluation accuracy on a stratified 20% CMB subset. **(a) Expert raters** = 7 kits (ID-a, KH-b, MM-a, MY-b, PH-a, TH-b, VN-a) with prior cultural-research experience. **(b) Non-expert raters** = 7 kits (ID-b, KH-a, MM-b, MY-a, PH-b, TH-a, VN-b) from the general population. "Native" = answering one's own country; "Non-native" = answering one other country via a fixed $+1/+2$ rotation. The Model bar averages over the six Carry-mode models on the same concept set the humans saw (N = unique concepts $\times$ 6 models).

is 0.70, versus 0.60 for non-selected candidates in the same per-category pool, confirming the chosen distractors are systematically harder than a random baseline by construction.

B.2 Benchmark categorization for Table 1

This appendix explains how each entry in Table 1 was categorized, with a focus on properties that required interpretation of the original paper.

SCB (Satar et al., 2025). SCB is image-based and covers seven Southeast Asia countries with 138 cultural concepts. We mark it as multi-stage with cascade because it has two distinct ability stages: Stage 1 is a visual MCQ for cultural reasoning, and Stage 2 evaluates segmentation of the cultural artifact only when Stage 1 is correct. This is a sequential cascade with conditional dependency. Stage 2 is spatial grounding, not temporal localization, so we mark it as ✗on free-form temporal localization. SCB does not report human accuracy on the same task as models (described as future work in the paper).

VideoNorms (Varimalla et al., 2025). VideoNorms covers American and Chinese cultural norms through TV show clips. Each clip is fixed at 15 seconds (0.25 min). The benchmark has three parallel tasks (binary adherence/violation classification, evidence extraction, and norm generation), but these are scored independently rather than sequentially. Evidence quality (Task 2) is computed only on correctly-labeled samples, which is a partial form of conditional analysis; we mark Conditional metrics as ✗because the headline

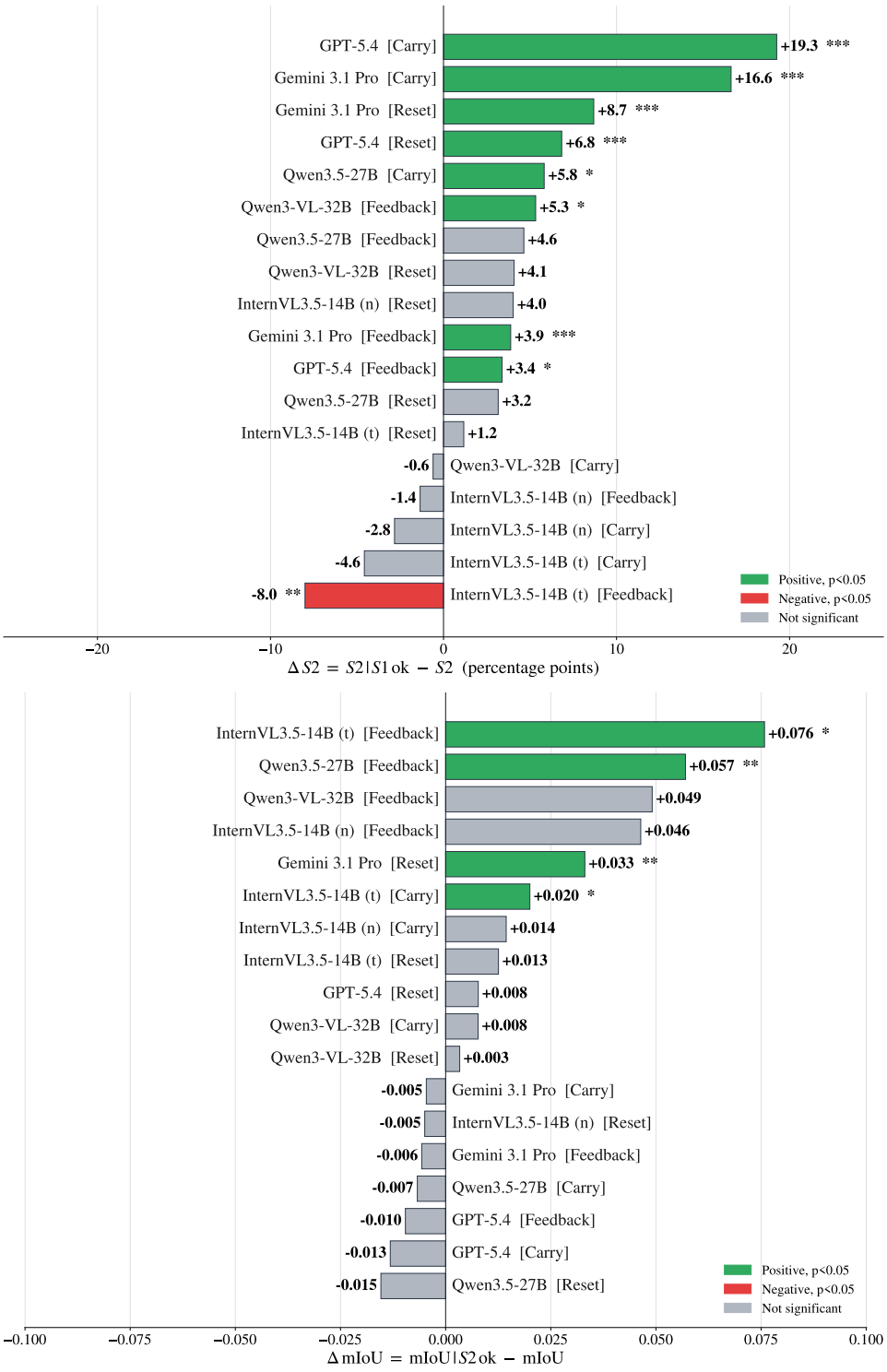

Figure 8: Per-(model, mode) cascade deltas. Top: ΔS2 = c-S2 − S2 (percentage points), measuring how much correct naming lifts visual recognition. Bottom: ΔmIoU = c-mIoU − mIoU, measuring how much correct visual recognition lifts temporal localization. Bars are sorted by effect size within each panel. Stars: $*\,p < 0.05$, $**\ p < 0.01$, $***\,p < 0.001$ against the unconditional baseline.

metric is not a stage cascade. Humans validated only a subsample of LLM-judged outputs; no human accuracy on the full task is reported.

GIMMICK (Schneider et al., 2025). GIMMICK is a multitask cultural benchmark spanning UNESCO ICH. The benchmark includes both image (CIVQA) and video (CVVQA) subsets; we compare against the CVVQA video subset, which

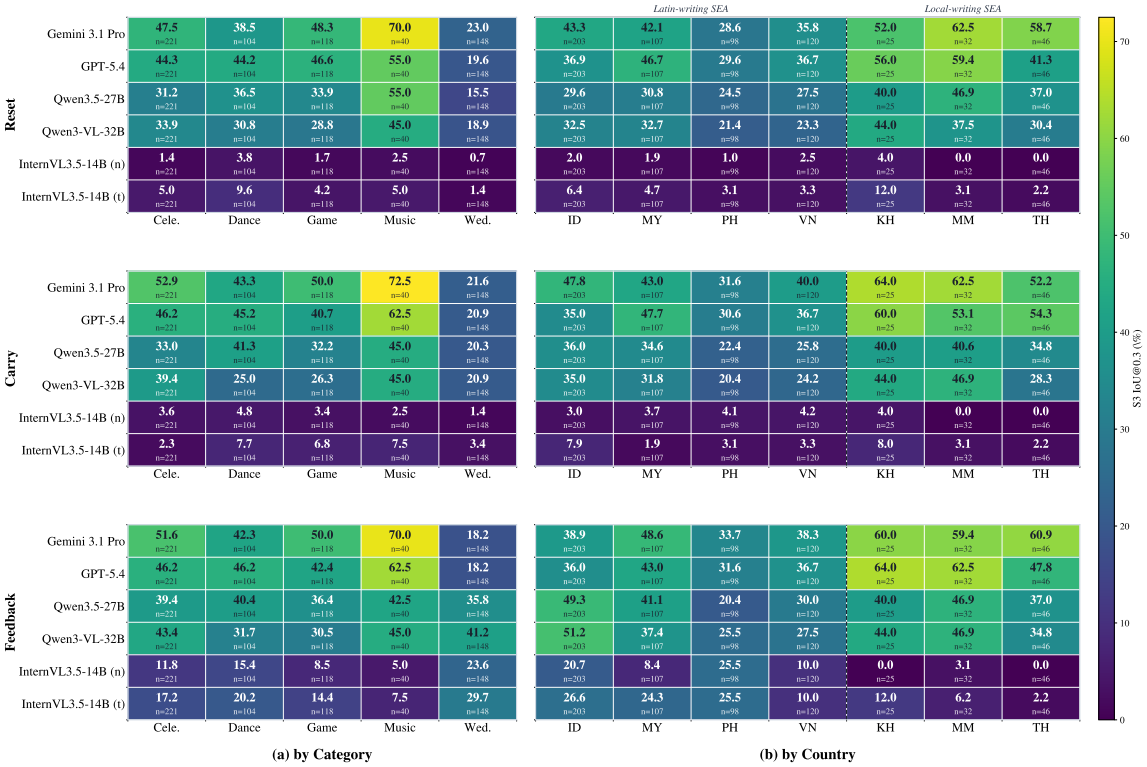

Figure 9: S3 IoU@0.3 (percentage of pairs with IoU $\geq$ 0.3) for all six models under all three modes, split by category (left) and by country writing-system cluster (right). Cells report the percentage; row n is the number of pairs aggregated.

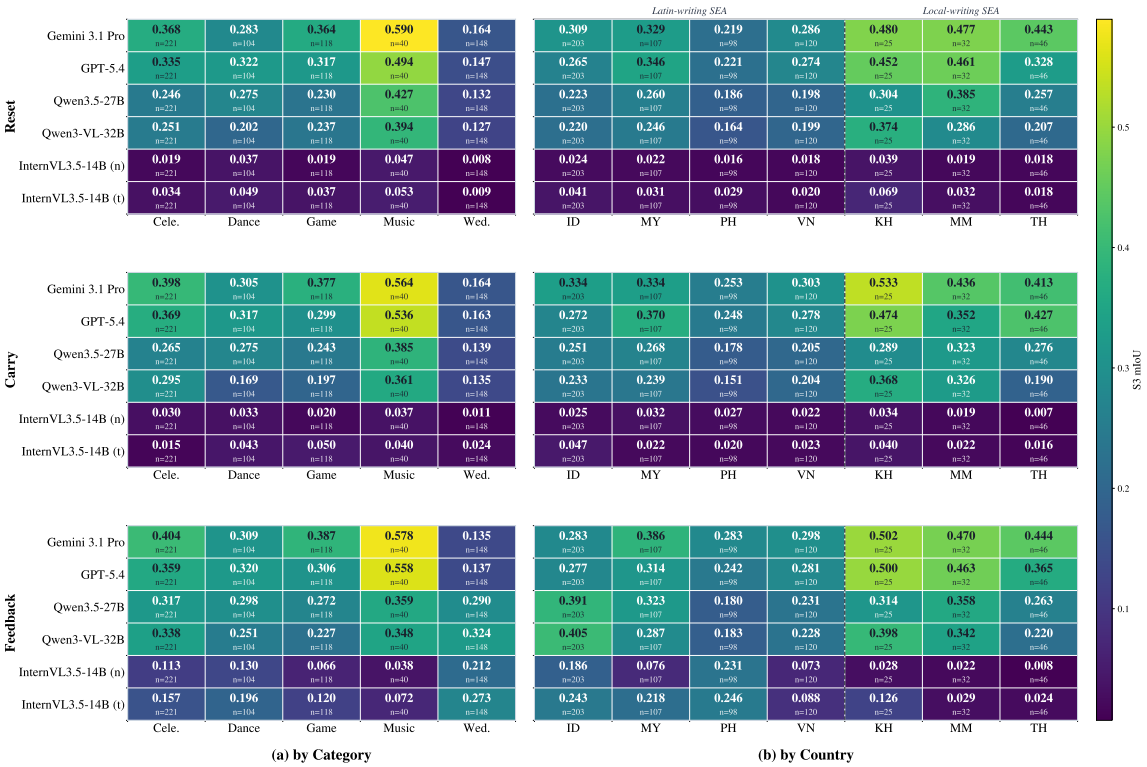

Figure 10: S3 mIoU (mean Intersection-over-Union) for all six models under all three modes, split by category (left) and by country writing-system cluster (right). Cells report mIoU on $[0, 1]$; row n is the number of pairs aggregated.

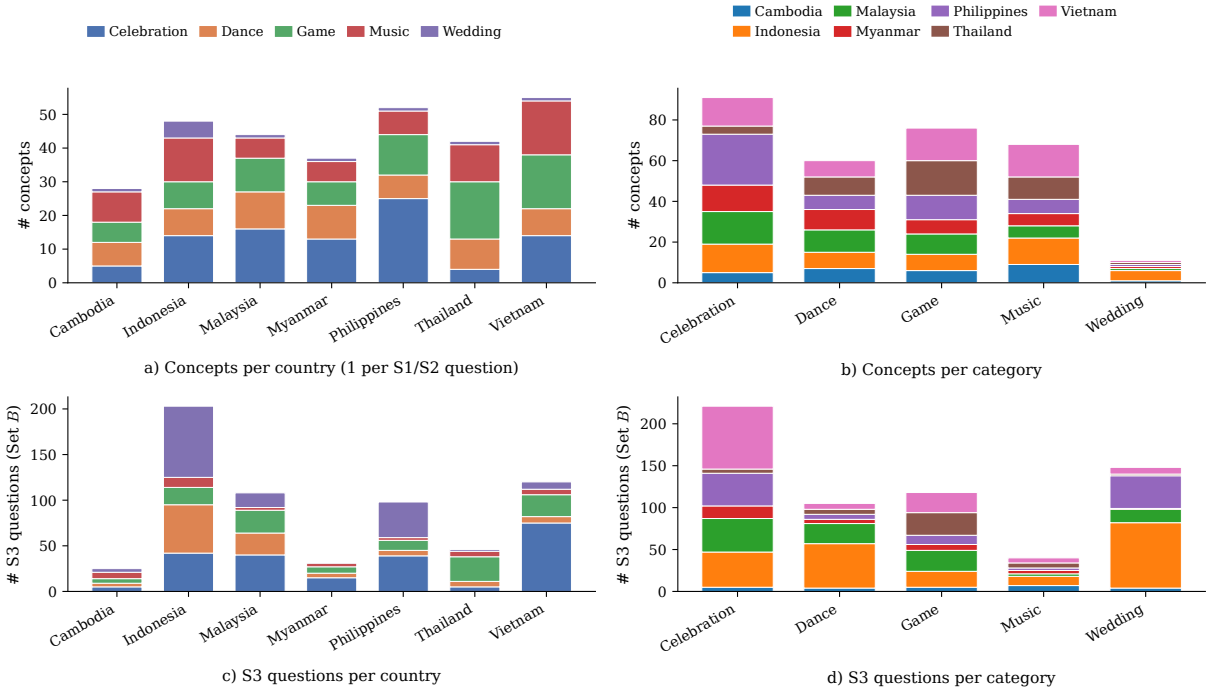

Figure 11: CMB concept and S3 question distribution. (a, b) Concepts per country and per category (n=306). Each concept corresponds to one S1 and one S2 question on Set A. (c, d) S3 questions per country and per category on Set B (n=631 across 318 source videos).

has 1,809 samples with 10-second clips (0.17 min) drawn from 139 countries. The six tasks are evaluated independently, so we mark Multi-stage cascade as ✗. GIMMICK explicitly evaluates only in English; SEA's focus is partial because the Asia-Pacific region is one of six macro-regions, but it is not the focus.

ViMUL-Bench (Shafique et al., 2025). ViMUL-Bench spans 14 languages across 14 countries. The 14 languages include Arabic, Bengali, Chinese, English, French, German, Hindi, Japanese, Russian, Sinhala, Spanish, Swedish, Tamil, and Urdu; none are Southeast Asian languages, so we mark SEA focus as ✗. Question formats (MCQ, short open-ended, long open-ended) are parallel, not cascade stages, so Multi-stage cascade is ✗. Average video duration was computed from the released dataset as 175 seconds (2.92 min). Humans only validate the LLM-judge on a sub-sample, so independent evaluation is marked partial.

MINERVA-Cultural (Singh et al., 2026). MINERVA-Cultural is a long-form (12.62 min average) multilingual benchmark across 18 locales. Of these 18 locales, 2 are Southeast Asian (id-ID for Indonesia, th-TH for Thailand), so the SEA focus is partial. The benchmark format is single-stage open-ended QA judged by an LLM on a 0/1/2 scale; we therefore mark Multi-stage cascade as ✗. MINERVA introduces Iterative Error Isolation (IEI), which traverses an evidence DAG with corrective hints across three iterations. We mark Conditional metrics as ✓because IEI reports cross-step success-given-success. MINERVA's independent human evaluators do not see ground-truth answers but are permitted to use web search for grounding unfamiliar cultural entities; we therefore mark Human eval as ✓(without the *closed* annotation), reflecting independent evaluation under an open-book protocol.

AVMeme Exam (Jiang et al., 2026). AVMeme Exam evaluates audio-visual cultural memes. The paper describes its coverage as "East and South Asia, Middle East, Europe, and North America" with contributors from the U.S., China, Japan, India, and the Middle East; no Southeast Asian country is named in the coverage list, contributor list, or reported results languages, so SEA focus is ✗. Twenty independent human participants evaluated on the same task with an explicit closed-book protocol: "without web search, collaboration, or external assistance," so we mark Human eval as ✓(*closed*). Video duration is capped at 30 seconds (0.5 min); no average is reported.

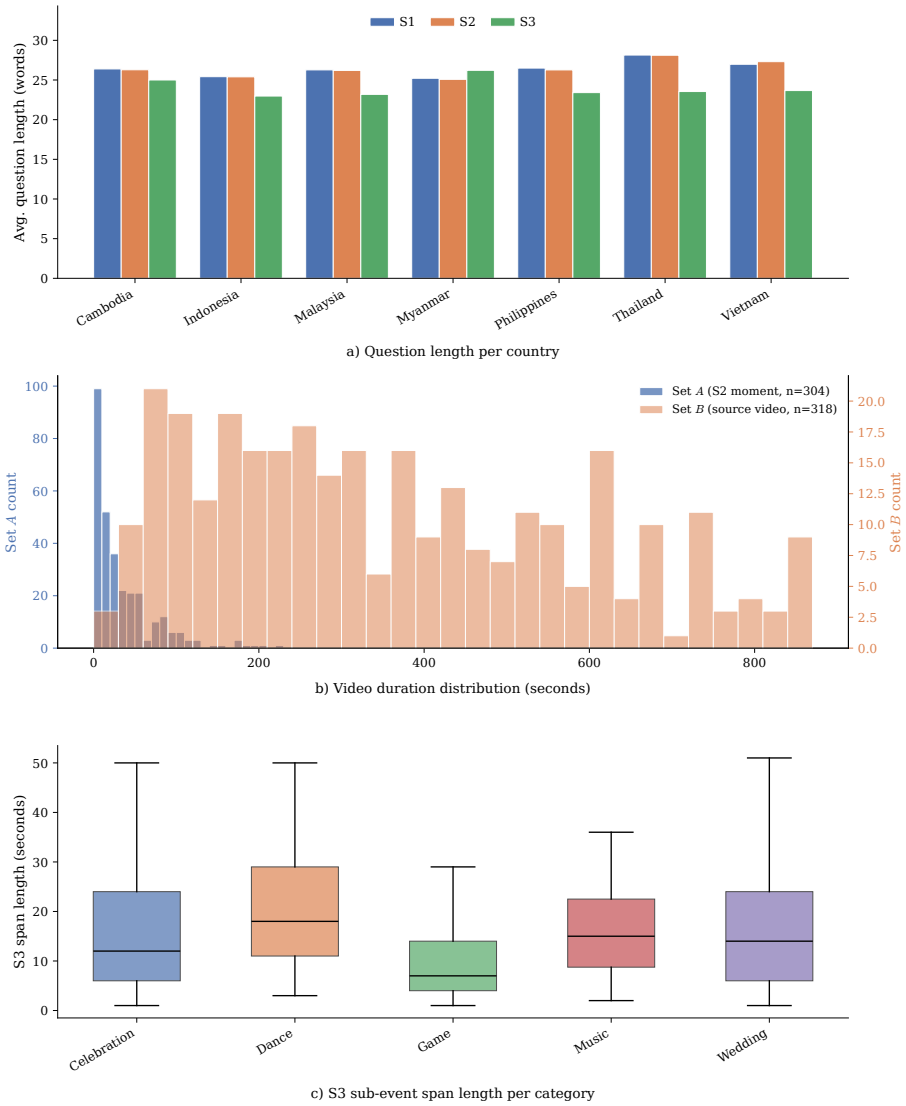

Figure 12: Question length and video duration distributions. (a) Average S1, S2, and S3 question length (words) per country. (b) Video-duration histogram: Set A S2 moments are short (median 20 s), Set B S3 source videos are long (median 385 s). (c) S3 sub-event span lengths per category, median 13 s overall.

ChineseVideoBench (Nie et al., 2025). ChineseVideoBench is monocultural (Chinese-only), with 1,625 videos averaging about 1 minute each. It has eight main task categories, including a "Temporal Localization" category, but inspection of the question format shows it is an MCQ over orderings rather than free-form span prediction; we therefore mark Free-form temporal localization as ✗. The benchmark explicitly removes audio tracks to force visual reasoning. The reported human baseline of 94.8% comes from the third subgroup of the same nine-annotator pool that built the benchmark, so we mark Human eval as ✗— the evaluators are not independent of the annotators.

VideoVista-CulturalLingo (Chen et al., 2025). VideoVista-CulturalLingo covers three cultural regions (Chinese, North American, European) across 1,389 videos, with no Southeast Asian focus, so we mark SEA focus as ✗. The benchmark is a single-stage MCQ design without conditional or cascade scoring, so Multi-stage cascade and Conditional metrics are both ✗. Average video duration reported as 4.05 min. The released paper does not describe an independent human-eval study on the same task, so Human eval is ✗.

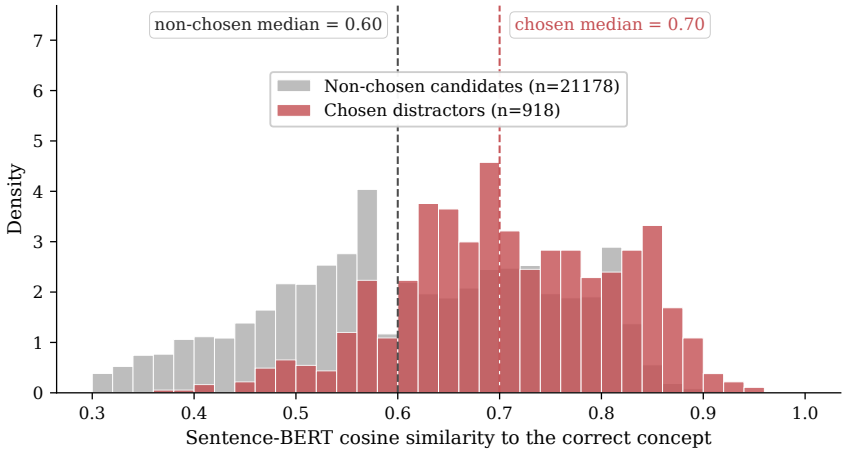

Figure 13: Sentence-BERT cosine similarity between the correct concept and (red) the three chosen S1/S2 distractors versus (gray) the remaining non-chosen candidates in the same per-category pool. Chosen distractors are systematically harder, with a median similarity of 0.70 vs 0.60 for non-chosen candidates.

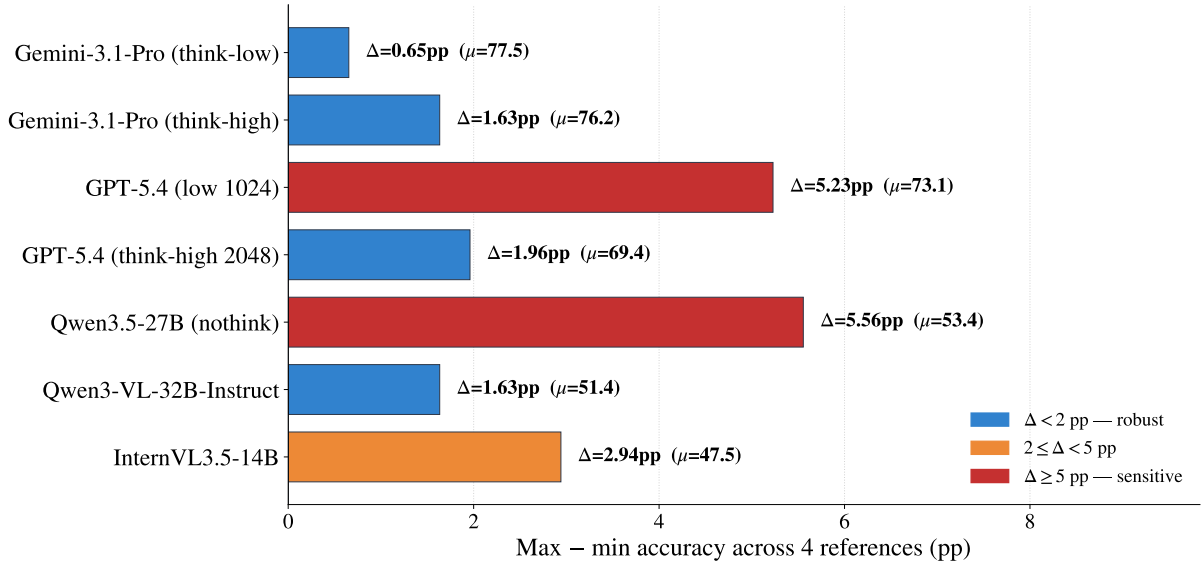

Figure 14: S1 reference variance across model configurations. For each model, S1 is rerun four times using the four name-reference variants per concept (common vs. official × Latin transliteration vs. local script); the bar reports the maximum minus the minimum S1 accuracy across the four runs (in percentage points), and the value in parentheses is the mean S1 accuracy (μ) across them. Two configurations are reference-sensitive at the ≥ 5 pp threshold (GPT-5.4 low-1024, Qwen3.5-27B nothink); the remaining five stay within 3 pp, indicating that S1 scores are robust to the choice of name reference.

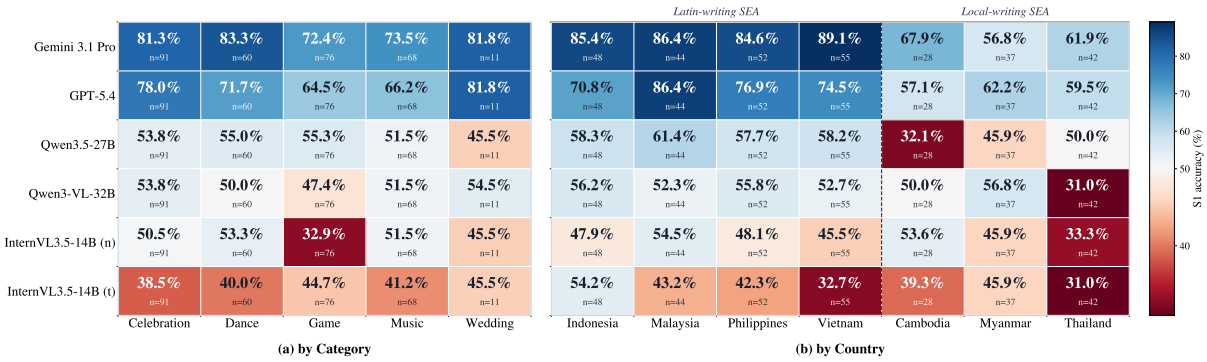

Figure 15: S1 by Category and Country.

Algorithm 1 Per-category distractor sampling (BSWA).

Require: Concepts C ($|C|{=}N$); similarity $\mathrm{sim}(\cdot,\cdot)$; country map $\mathrm{ctry}(\cdot)$; ambiguity groups $\mathcal{A}$
Ensure: Distractor triple $\mathrm{opt}[k] = (s_0, s_1, s_2)$ for each $k \in C$

$N_{\max} \leftarrow \lfloor 0.1 \cdot N \rfloor$
$\mathrm{opt}[k] \leftarrow (\perp, \perp, \perp),\ \mathrm{usage}[c] \leftarrow 0$ for all $k, c \in C$

Phase 1 (coverage guarantee).
Sort C ascending by placability $p(c) = |\{(k,s) : \textsc{Valid}(k,s,c)\}|$
for all $c \in C$ in sorted order **with** $\mathrm{usage}[c] = 0$ **do**
    $(k^*, s^*) \leftarrow \arg\max_{k,s} \mathrm{sim}(k,c)$ s.t. $\textsc{Valid}(k,s,c)$
    $\mathrm{opt}[k^*][s^*] \leftarrow c$; $\mathrm{usage}[c] \leftarrow \mathrm{usage}[c] + 1$
end for

Phase 2 (similarity-first greedy fill).
while any slot of any key is empty **do**
    Sort keys with empty slots ascending by remaining valid-candidate capacity
    for all such key k (most-constrained first) **do**
        for all empty slot s of k **do**
            $c^* \leftarrow \arg\max_c \mathrm{sim}(k,c)$ s.t. $\textsc{Valid}(k,s,c)$
            if c^* exists **then**
                $\mathrm{opt}[k][s] \leftarrow c^*$; $\mathrm{usage}[c^*] \leftarrow \mathrm{usage}[c^*] + 1$; **break**
            end if
        end for
    end for
end while

function $\textsc{Valid}(k,s,c)$
    if $c = k$ **or** $c \in \mathrm{opt}[k]$ **or** $\mathrm{usage}[c] \geq N_{\max}$ **then**
        return FALSE
    end if
    if $\exists\, a \in \{k\} \cup \mathrm{opt}[k] \setminus \{\perp\}$ s.t. $\{a, c\} \subseteq G$ for some $G \in \mathcal{A}$ **then**
        return FALSE ▷ ambiguity block
    end if
    if $s = 0$ **then**
        return $\mathrm{ctry}(c) = \mathrm{ctry}(k)$
    else
        return $\mathrm{ctry}(c) \neq \mathrm{ctry}(k)$ **and** $\mathrm{ctry}(c) \neq \mathrm{ctry}(\mathrm{opt}[k][\bar{s}])$
    end if
end function