SEEING CULTURE WHEN WORDS FALL SHORT: UNCERTAINTY-GATED VISUAL GROUNDING OF CULTURAL CUES FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have achieved strong performance across a wide range of tasks, yet they remain vulnerable to cultural bias due to the uneven distribution of cultural knowledge in training corpora. This limitation often leads to inaccurate understanding and generation in cross-cultural contexts, particularly when textual cues alone are insufficient. To address this issue at inference time, we propose **V-CUE** (Visual-Cultural Understanding Enhancement), a visual-space-assisted framework that improves culturally grounded reasoning by incorporating generated visual cues. Specifically, V-CUE first employs an uncertainty detection mechanism to identify potentially unreliable responses. Based on this signal, a text-to-image model generates culturally relevant visual content, including symbols, clothing, and contextual elements. These visual signals are incorporated into the generation process to refine model predictions and enhance cultural grounding. Evaluated on understanding task (CulturalBench) and generation task (CARE) across LLM-text, LLM-reasoning, and LLM-VL systems, V-CUE yields consistent accuracy gains of up to +19 points, while uncertainty gating reduces token cost by 46% with negligible accuracy loss. Further analyses confirm that gains scale with visual–query alignment quality and that V-CUE remains complementary to retrieval-augmented approaches. Without requiring additional training data, V-CUE offers a practical and generalizable multimodal pathway for enhancing the cultural robustness of LLMs[1].

1 INTRODUCTION

In recent years, Large Language Models (LLMs), such as ChatGPT, have been widely adopted across the globe, attracting millions of users (Achiam et al., 2023; Chen et al., 2025). To unlock the full potential of artificial intelligence and realize the vision of "AI for All", researchers have begun to call for greater emphasis on cross-cultural relevance in the design and development of LLMs (Adilazuarda et al., 2024; Kabir et al., 2025).

Although several countries have developed localized LLMs (Chang et al., 2024), such as Qwen (Team, 2025c) and DeepSeek (DeepSeek-AI et al., 2024) from China, and HyperCLOVA (Team, 2025a) from South Korea—these models often fail to adequately consider cultural differences during pretraining and alignment tuning (Xu et al., 2025; Huang et al., 2024). This limitation further manifests during inference, leading to culturally inaccurate understanding and generation (Choenni & Shutova, 2024).

This cultural neglect can lead to user misunderstandings and even bias in cross-cultural contexts (Prabhakaran et al., 2022), thereby undermining the applicability of LLMs in diverse cultural settings. As a result, cultural bias remains a core challenge in current LLM development (Huang et al., 2024; Li et al., 2024a).

To address cross-cultural bias in LLMs during the inference, we introduce **V-CUE** (**V**isual-**C**ultural **U**nderstanding **E**nhancement), a visual-space-assisted generation framework for culture tasks (Figure 1). When an LLM detects high uncertainty or potential cultural bias, V-CUE proactively queries a

[1] Our code and data will be available once the paper is accepted.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

visual space to create images rich in cultural cues—symbols, clothing, or situational elements. These visual signals are then fed back into the original generating process.

To assess the effectiveness of the proposed approach, we conduct comparisons on two levels of culturally grounded tasks: classification and question answering (CulturalBench (Chiu et al., 2025)) and generation (CARE (Guo et al., 2025b)), against LLM-text systems, LLM-reasoning systems, and LLM-VL systems. Experimental results demonstrate that our method exhibits stronger generalizability and robustness in these multicultural grounded tasks. Analysis experiments show that V-CUE reduces cultural errors by enhancing the salience of existing cultural cues through visual signals, rather than introducing new knowledge. The main contributions of this study are as follows:

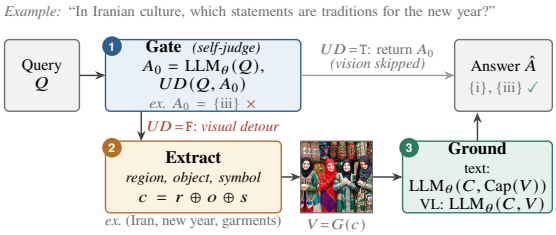

Figure 1: Overview of V-CUE. When the LLM is uncertain about cultural questions, a visual module generates culturally relevant images to refine outputs and improve cultural understanding.

- We propose V-CUE, a visual-space-assisted framework that enhances culturally sensitive tasks and mitigates cross-cultural bias in LLMs through generated visual cues.
- We systematically evaluate multiple LLMs across diverse cultural tasks, revealing the limitations of current LLMs and the effectiveness of V-CUE.
- Our analysis identifies key factors affecting the impact of visual signals on cultural tasks, offering insights for future multimodal cultural modeling.

2 Related Works

From an evaluation perspective, prior work has examined LLMs' cultural cognition by analyzing responses to social survey questions on human values and attitudes (Cao et al., 2023). Researchers have also proposed culture-specific evaluation tools, including cultural fact knowledge bases (Yin et al., 2022) and entity-based cultural benchmarks (Naous et al., 2024). In addition, question–answer datasets have been constructed to cover diverse cultural domains, such as cooking habits (Palta & Rudinger, 2023), social norms (Rao et al., 2025), and social etiquette (Chiu et al., 2025). More comprehensive benchmarks have recently emerged, including CulturalBench (Chiu et al., 2025), CultureBank (Shi et al., 2024), BLEND (Myung et al., 2024), and CARE (Guo et al., 2025b), which span multiple cultures, regions, and languages with human verification or preference judgments. Collectively, these studies show that advanced LLMs tend to exhibit Western-centric biases, highlighting the growing importance of cultural alignment beyond evaluation.

From a methodological perspective, early studies mainly relied on prompt engineering to elicit culturally situated reasoning in LLMs (Durmus et al., 2023). Subsequent work embedded cultural information directly into the context, such as through scenario descriptions or role-playing (Kwok et al., 2024; Kharchenko et al., 2024), or adopted soft prompt tuning to mitigate biases (Masoud et al., 2025). A more robust line of research fine-tunes LLMs for specific cultures using curated datasets, including regional models that continue pre-training on large-scale local corpora (Almeida et al., 2024; Nguyen et al., 2023; Team, 2025c). However, such approaches are computationally costly and limited by single-language text data. To improve efficiency, CultureLLM (Li et al., 2024a) induces consistent cultural behavior via global value survey responses, while CulturePark (Li et al., 2024b) adopts a multi-agent framework for cross-cultural communication, and CultureSPA (Xu et al., 2025) analyzes differences between culturally insensitive and sensitive settings. Overall, existing methods remain text-centric, whereas our approach introduces a visual modality for the first time to complement and correct text-induced biases, thereby enhancing cultural cognition in large models.

3 Methodology

To address cultural bias in culturally dependent tasks, we introduce a visual-space-assisted generation mechanism. When an LLM exhibits uncertainty in cultural understanding, external visual cues

are invoked to refine and correct the generation process. Specifically, we design a visual-guided framework that models interactions between language and visual spaces. As illustrated in Figure 1, the framework consists of three components: uncertainty detection (UD), visual generation (VG), and visual-based generation reconstruction (VR).

Uncertainty Detection. The first component of the V-CUE framework introduces a self-judge mechanism to estimate response reliability. Crucially, the judge is not an external evaluator or a separate model; it is performed by the *same backbone LLM* that generates the initial response. Following the LLM-as-Judge paradigm (Dong et al., 2023; Zheng et al., 2023; Gu et al., 2024), UD operates in two steps. Given an input query Q, the backbone first produces a preliminary answer:

$$A_0 = \text{LLM}(Q) \tag{1}$$

The (Q, A_0) pair is then fed back to the same model with a judge prompt P_{ud} to obtain a binary reliability label:

$$UD(Q) = \text{LLM}\big(P_{\text{ud}},\ Q,\ A_0\big) \in \{\texttt{True},\ \texttt{False}\} \tag{2}$$

If $UD(Q) = \texttt{False}$, the system activates the *Visual Generation* module for auxiliary cultural reasoning (Figure 1); otherwise, when $UD(Q) = \texttt{True}$, A_0 is returned directly, reducing inference cost and latency. This self-judge design avoids the overhead of an additional judge model while leveraging the backbone's own cultural knowledge to identify its uncertain outputs. Prompt templates and decision criteria for UD are provided in the supplementary material.

Visual Generation. Upon detecting high uncertainty ($UD(Q) = \texttt{False}$), the system invokes an external visual space to support culturally grounded reasoning (Figure 2, VG row). Which is realized via a text-to-image diffusion model that provides culturally informed visual cues.

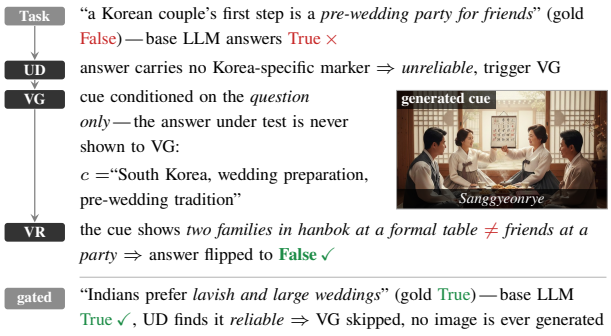

Figure 2: **A worked example of the V-CUE pipeline.** Top: UD flags the answer, VG builds the cue from the question alone, and VR turns the cue–answer mismatch into a corrected label. Bottom: a reliable answer skips VG entirely.

The process begins with a culture-aware extraction step. Given the query Q, a structured parsing prompt P_{ext} instructs the backbone LLM to extract three culturally salient fields:

$$(r, o, s) = \text{LLM}\big(P_{\text{ext}},\ Q\big) \tag{3}$$

where r, o, and s denote the cultural `region`, `object`, and `symbol`, respectively. To prevent label leakage, P_{ext} accesses only the question text without answer options or gold labels. The three fields are concatenated into a culture-aware condition unit:

$$c = r \oplus o \oplus s \tag{4}$$

where $\oplus$ denotes string concatenation.

The condition unit c is then passed to a text-to-image model G to synthesize a culturally relevant image:

$$V = G(c) \tag{5}$$

Prompt templates and extraction rules are detailed in the supplementary material.

Visual-based Generation Reconstruction. We introduce a visual-assisted reconstruction mechanism that integrates an image generation module during inference to provide auxiliary cultural cues, which interact with the LLM to facilitate self-correction of its outputs.

The overall process is formulated as:

$$\text{Answer} = \mathcal{F}(C, V) \tag{6}$$

where C denotes the original textual context, V is a culturally relevant image generated from C, and $\mathcal{F}$ is the response generation function, whose execution depends on the underlying model architecture.

Specifically, $\mathcal{F}$ follows two distinct pathways based on the LLM's multimodal capability.

For text-only LLMs, we introduce an image-to-text mediation strategy. The generated image V is first converted into a textual description $\text{Cap}(V)$, which is concatenated with the original context C and fed into the LLM:

$$\mathcal{F}(C, V) = \text{LLM}_{\text{text}}(C, \text{Cap}(V)) \tag{7}$$

where $\text{Cap}(V)$ encodes culturally informative visual cues. This pathway embeds visual knowledge into the textual reasoning process, enabling culturally informed generation without native visual perception.

For multimodal LLMs, the model directly processes both textual and visual inputs without intermediate conversion:

$$\mathcal{F}(C, V) = \text{LLM}_{\text{VL}}(C, V) \tag{8}$$

where LLM_{VL} jointly models text and images. This direct cross-modal fusion aligns linguistic and visual representations, making it more effective for reasoning under complex cultural contexts. Prompt generation templates are detailed in the supplementary material.

Overall pipeline. Figure 2 instantiates both branches of V-CUE on a single item.

4 Evaluation Setting

4.1 Dataset

We evaluate our approach on two benchmarks: CulturalBench for cultural understanding, and CARE for cultural generation.

The **CulturalBench** (Chiu et al., 2025)[2] is selected to evaluate cultural understanding of LLMs across a broad and diverse set of regions. The benchmark focuses on assessing cultural breadth and diversity by covering *45* global regions. It comprises 4.91k True-or-False and 1.23k Multiple-Choice human-written cultural knowledge QA pairs, each verified by five independent reviewers, and is designed to systematically challenge LLMs on cross-regional and cross-cultural reasoning.

The **CARE-eval** (Guo et al., 2025b)[3] benchmark is adopted to evaluate LLM performance on culture-specific generation tasks, comprising 150 questions for each of the Chinese, Japanese, and Arab cultures, with 30 samples randomly drawn from each of five cultural categories.

4.2 Systems Setting

To provide a thorough assessment of the V-CUE approach, comparisons are conducted at three distinct capability levels: LLM-text, LLM-reasoning, and LLM-VL, each reflecting a progressively higher degree of reasoning and modality integration.

We evaluate representative baselines from three categories: **LLM-text** (DeepSeek-V3 (DeepSeek-AI et al., 2024), Qwen3 (Team, 2025c), ChatGPT-3.5 (Ouyang et al., 2022), CulFiT (Feng et al., 2025), CultureLLM (Li et al., 2024a), CulturePark (Li et al., 2024b)), **LLM-reasoning** (Qwen3 (Team, 2025c), DeepSeek-R1 (Guo et al., 2025a), GPT-o3-mini (OpenAI, 2024)), and **LLM-VL** (Qwen2.5-VL-72B (Team, 2025b), GPT-4o (Achiam et al., 2023)).

For the image generation model and the image captioning model in V-CUE, we used stable-diffusion-2-1 (Rombach et al., 2022) [4] and Qwen2.5-vl-72B (Team, 2025b), respectively.

Component selection guideline. V-CUE consists of three pluggable components: *UD*, a backbone-internal self-judge requiring no extra model; *visual generator*, where Stable Diffusion 2.1 (1.3B) is used as a cost-efficient default; and *visual mediator*, where larger captioners improve performance (7B–72B, Appendix J).

Implementation details. Experiments are conducted on a single NVIDIA A100 (80GB). We use Stable Diffusion 2.1 with guidance scale 7.5, 30 denoising steps, and 512×512 resolution. Statistical significance is evaluated with five trials on CulturalBench and three on CARE using paired one-sided

[2] https://huggingface.co/datasets/kellycyy/CulturalBench

[3] https://huggingface.co/datasets/geyang627/CARE-eval

[4] https://huggingface.co/stabilityai/stable-diffusion-2-1

Student's t-tests (`scipy.stats.ttest_rel`); unless stated otherwise, improvements over the corresponding base models are significant at $p < 0.01$.

5 Main Experimental Results

5.1 Results on the CulturalBench

Table 1: Evaluation results on CulturalBench (Multiple-Choice and True/False, accuracy %). ↑: gain of V-CUE over the base model; *: results from original papers; –: not reported. Gray cells: V-CUE is not applied, since CulFiT has no released checkpoints and CultureLLM/CulturePark bake cultural knowledge into the Qwen3-32B weights at training time; they thus serve as training-time reference points for our inference-time gains on Qwen3-32B rather than as backbones to be further augmented. All subsets were evaluated five times.

Model	Multiple-Choice (ACC)			True or False (ACC)		
	Base	+V-CUE	↑	Base	+V-CUE	↑
LLM-text Cultural Methods						
CulFiT (Qwen 2.5-7B)*	–			72.9		
CulFiT (Llama3.1-8B)*	–			73.1		
CultureLLM (Qwen3-32B)	79.4			74.1		
CulturePark (Qwen3-32B)	77.4			75.5		
LLM-text systems						
Qwen3-235B-A22B	83.1	85.6	2.5	82.5	89.0	6.5
Qwen3-32B	79.2	82.9	3.7	74.1	81.5	7.4
DeepSeek-V3	81.7	84.8	3.1	83.3	89.9	6.6
GPT-3.5-turbo	67.8	75.8	8.0	63.1	70.9	7.8
LLM-reasoning systems						
Qwen3-235B-A22B	73.9	**88.8**	14.9	85.9	**93.3**	7.4
Qwen3-32B	64.2	83.3	19.1	83.0	89.4	6.4
DeepSeek-R1	87.3	91.0	3.7	85.2	90.7	5.5
o3-mini	90.5	92.7	2.2	89.0	92.1	3.1
LLM-VL systems						
Qwen2.5-vl-72B	84.9	87.5	2.6	86.8	90.0	3.2
GPT-4o	87.0	88.7	1.7	89.6	91.8	2.2

We evaluate V-CUE on CulturalBench. The experimental results are shown in Table 1. We observe accuracy gains across models and tasks, demonstrating strong generalizability and effectiveness.

LLM-text systems. V-CUE delivers stable improvements across mainstream models. Qwen3-235B-A22B, Qwen3-32B, and DeepSeek-V3 improve by +3.1 (Multiple-Choice) and +6.8 (True or False) on average, while GPT-3.5-turbo shows the largest gains (nearly +8 points on both tasks).

LLM-reasoning systems. V-CUE provides strong compensatory effects despite inherent cultural bias. Qwen3-235B-A22B and Qwen3-32B gain over +14.9 and +19.1 points on Multiple-Choice and over +6 points on True or False tasks; DeepSeek-R1 and o3-mini achieve consistent +2–5 point gains.

LLM-VL systems. Even with multimodal pretraining, V-CUE yields additional gains. Qwen2.5-vl-72B improves by +2.6 (Multiple-Choice) and +3.2 (True or False), indicating that explicit visual–cultural cues further enhance cultural understanding.

5.2 Results on the CARE Benchmark

We evaluated V-CUE on the CARE-eval benchmark following the original protocol. Using the same LM-as-a-judge setup as CARE (Guo et al., 2025b), GPT-4o[5] scores model outputs on generality, cultural relevance, and literary quality at the language level. CARE-eval demonstrates high reliability, with a Pearson correlation of 0.93 against human judgments under fixed prompts. The experimental results are shown in Figure 3.

[5]The version of GPT-4o is gpt-4o-2024-05-13.

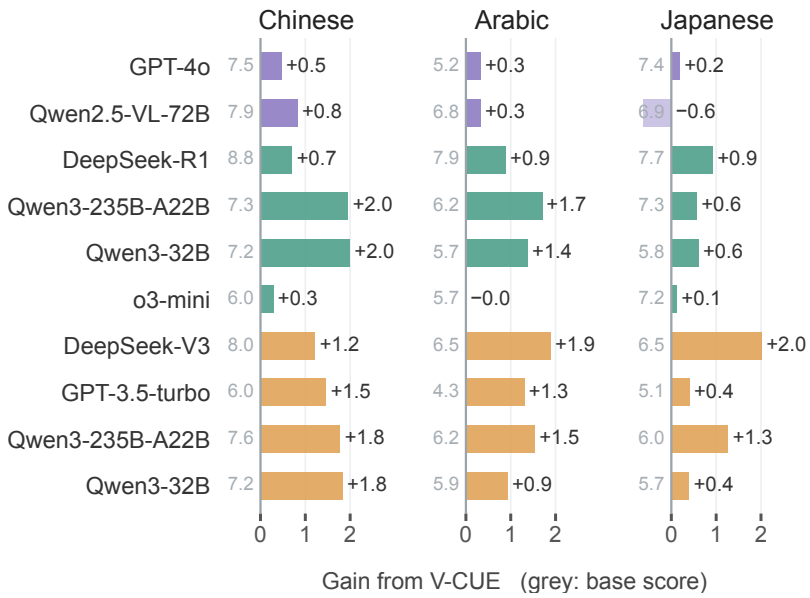

Figure 3: Results compare base LLMs and their V-CUE–enhanced versions on CARE; rows denote before/after V-CUE. Scores are averaged over three runs. Average scores (1–10) on Chinese, Arabic, and Japanese cultural questions in local languages.

For different model types, V-CUE consistently improves performance across LLM-text, LLM-reasoning, and LLM-VL systems, demonstrating its effectiveness on generative cultural tasks. For different cultures, CARE-eval benchmarks cover three distinct cultural settings—Chinese, Arabic, and Japanese. V-CUE yields consistent gains across all three, with the most significant improvements observed on Chinese-based LLMs. For different languages, CARE-eval uses prompts in local languages rather than English. V-CUE enhances performance across all language settings, highlighting its robustness in multilingual scenarios.

6 Analysis Experiment

To ensure consistency across different system architectures, we conducted our analytical experiments using the Qwen series, covering LLM-text, reasoning, and VL systems, unless stated otherwise. Building on the overall gains in Section 5, we next analyze why V-CUE works and under which conditions the gains are most pronounced. We ask five questions in turn: whether the gains depend on the particular image generator (§6.1), whether the image modality and its cultural content are what matters (§6.2), how much each component contributes (§6.3), whether uncertainty gating preserves the gains at lower cost (§6.4), and whether visual cues offer anything beyond textual augmentation (§6.5).

Additional analyses of visual quality, cultural recovery, and image-to-text mediation are provided in Appendices H–J.

6.1 Effect of VG on V-CUE

Appendix H shows that V-CUE benefits from higher-quality visuals, but every result so far comes from a single generator, so the gains may reflect properties of Stable Diffusion rather than of V-CUE itself. We test this by swapping the visual generation module (VG) for Gemini-3 Pro-Image (Nano Banana Pro)[6] while holding the backbones (Qwen3-235B-A22B for LLM-text and LLM-reasoning, Qwen2.5-VL-72B for LLM-VL) and all downstream settings fixed, and we measure the swap at both ends of the pipeline: upstream image–query alignment (CLIPScore (Hessel et al., 2021) on CARE-test) and downstream accuracy (CulturalBench multiple-choice).

[6]The version of Gemini-3 Pro-Image is gemini-3-pro-image-preview.

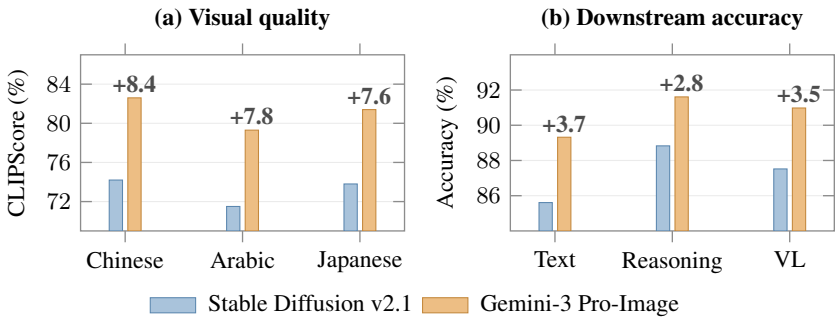

Figure 4: Replacing the visual generation module of V-CUE. (a) CLIPScore image–query alignment on CARE-test; (b) V-CUE accuracy on CulturalBench multiple-choice for LLM-text, LLM-reasoning, and LLM-VL backbones. Labels give the gain of Gemini-3 Pro-Image over Stable Diffusion v2.1; y-axes are truncated.

Figure 4 reports both measurements. Gemini-3 Pro-Image aligns its images with the cultural query far more closely than Stable Diffusion v2.1, raising mean CLIPScore from 73.2 to 81.1, and this upstream gap propagates to every downstream setting, improving accuracy by $+2.8$ to $+3.7$ points without reordering LLM-text, LLM-reasoning, and LLM-VL. V-CUE is therefore robust to the choice of generator and scales with visual fidelity, yet the downstream gain is clearly sub-linear in alignment quality, indicating that a moderately faithful generator already captures most of the achievable benefit.

6.2 Comparing Visual and Textual Modalities

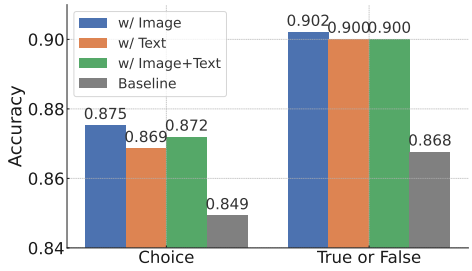

Figure 5: Performance of Qwen2.5-vl-72B on CulturalBench with different input modalities. w/ Image denotes generated image input (V), w/ Text uses its caption $Cap(V)$, and w/ Image+Text combines both.

Table 2: V-CUE performance across visual input types on CulturalBench (True or False) with Qwen2.5-VL-72B.

Image Type	Accuracy	Correction Rate
Cultural Visuals	**90.04%**	**23.90%**
Neutral Visuals	88.95%	18.94%
Unrelated Visuals	85.24%	1.40%

To compare the roles of visual and textual modalities in cultural understanding, we conduct controlled experiments on two CulturalBench tasks (Multiple-Choice and True or False) using Qwen2.5-VL-72B. We evaluate three input settings: images (w/ Image), image-generated captions (w/ Text), and their combination (w/ Image+Text).

As shown in Figure 5, images achieve the best performance on both tasks (w/ Image), with 87.5% accuracy on Multiple-Choice and over 90.0% on True or False. Caption-based inputs provide only limited gains (w/ Text). These results indicate that preserving native visual signals is more effective for cultural understanding in LLM.

Does the visual *content* matter? Having established that the image modality is the effective channel, we further ask whether the gain depends on what the image depicts. We evaluate V-CUE on LLM-VL (Qwen2.5-VL-72B) under three conditions with all other inputs fixed: culturally relevant images (Cultural Visuals), visually neutral images (Neutral Visuals), and unrelated images (Unrelated Visuals). Neutral and unrelated visuals are obtained by prompting DeepSeek-V3 to remove culture-specific elements from the extracted cues or replace them with random ones, followed by Stable Diffusion synthesis (Appendix C). Besides accuracy, we report the correction rate, i.e., the share of originally incorrect answers that become correct after V-CUE (Appendix I).

As shown in Table 2, Cultural Visuals achieve the best results (90.04% accuracy, 23.90% correction rate). Neutral Visuals yield moderate correction (18.94%) but lower accuracy (88.95%), while

Unrelated Visuals perform worst (1.40% correction, 85.24% accuracy), below the baseline (86.76%). This confirms that culturally aligned visuals aid correction, whereas unrelated visuals mislead the model: the benefit of V-CUE comes from the cultural cues the image carries, not from the mere presence of an image.

6.3 Ablation Study of V-CUE Components

Table 3: Ablation results of V-CUE on CulturalBench (Multiple-Choice and True or False tasks). Modules include uncertainty detection (UD), visual generation (VG), and visual-based reconstruction (VR). The last column of each task removes VR for text/reasoning systems and VG for VL systems.

Model	Multiple-Choice			True or False		
	+V-CUE	w/o UD	w/o VR/VG	+V-CUE	w/o UD	w/o VR/VG
LLM-text systems						
Qwen3-235B-A22B	85.6	84.6	83.7	89.0	87.9	84.6
Qwen3-32B	82.8	81.7	79.9	81.5	77.4	76.9
LLM-reasoning systems						
Qwen3-235B-A22B	88.8	85.5	83.3	93.2	89.3	87.6
Qwen3-32B	83.2	80.3	78.6	89.3	86.8	84.5
LLM-VL systems						
Qwen2.5-vl-72B	87.5	86.4	85.5	90.0	88.5	86.9

Table 3 reports the ablation results of V-CUE components on CulturalBench. Removing uncertainty detection (w/o UD) leads to a consistent 1–3 point ACC drop across models (e.g., 85.6 $\rightarrow$ 84.6 for Qwen3-235B-A22B; 82.8 $\rightarrow$ 81.7 for Qwen3-32B), indicating that LLMs retain limited inherent capability for cultural attribute recognition. Removing visual generation (w/o VG) forces LLM-VL models to rely solely on UD outputs without visual cues, resulting in an ACC decrease of around 3 points (90.0 $\rightarrow$ 86.9), which highlights the contribution of visual signals to cultural correction. In contrast, removing visual-based reconstruction (w/o VR) causes a substantial performance degradation for LLM-text and LLM-reasoning systems, with drops of up to 4–5 points on both Multiple-Choice and True/False tasks (e.g., 82.8 $\rightarrow$ 79.9; 83.2 $\rightarrow$ 78.6), underscoring the critical role of VR-converted textual information in enhancing cultural understanding.

6.4 Effect of UD on V-CUE

Table 4: Uncertainty detection (UD) as a gate for V-CUE. (a) Accuracy (%), trigger rate (%), tokens and latency (ms) per query on 1,000 CulturalBench multiple-choice instances (Qwen2.5-VL-72B); †Oracle-gated consults gold labels and is an upper bound rather than a deployable strategy. (b) Trigger rate (%) per culture on CARE, with the sample standard deviation across cultures.

(a) Accuracy–cost trade-off on CulturalBench-Choice

Method	Acc. $\uparrow$	Trig. $\downarrow$	Toks. $\downarrow$	Lat. $\downarrow$
Always-on	85.9	100.0	1820	1260
UD-gated (Ours)	85.6	34.6	980	720
Oracle-gated†	86.3	31.2	940	690

(b) UD trigger rate across cultures on CARE

Backbone	Chinese	Arabic	Japanese	σ
DeepSeek-V3	28.6	31.2	29.7	1.31
Qwen3-235B-A22B	29.4	30.8	30.1	0.70

V-CUE improves cultural understanding by generating and consuming an image for every query, so whether the method is practical depends on a gate that invokes enhancement only when it is actually needed. We compare three triggering strategies on 1,000 instances sampled from CulturalBench multiple-choice with Qwen2.5-VL-72B: *Always-on*, which enhances every instance; *UD-gated* (ours), which enhances only when uncertainty detection flags the base response as unreliable; and *Oracle-gated*, which enhances only the instances the base model actually answers incorrectly and therefore upper-bounds any gating rule.

Table 4(a) reports accuracy against cost. UD-gated fires on 34.6% of instances, cutting token usage by 46% and latency by 43% relative to Always-on, while the three strategies span only 0.7 accuracy points—seven instances out of 1,000, which is below the sampling noise of this evaluation ($\sigma \approx 1.1$ points). Uncertainty detection thus retains the full benefit of always enhancing at roughly half the cost, and approaches the oracle trigger rate of 31.2%.

Two further checks indicate that the gate is genuinely selective rather than merely conservative. Measured against ground-truth error labels on CulturalBench (True/False), UD reaches 0.74 precision and 0.68 recall (F1 $= 0.71$); and its trigger rate on CARE stays within 28.6–31.2% across the Chinese, Arabic, and Japanese subsets for both backbones (Table 4(b), $\sigma \leq 1.31$), so the cost savings are not obtained by systematically skipping any one culture.

6.5 V-CUE Text Augmentation Analysis

To isolate the effect of visual signals—rather than additional context or textual tokens, especially in the LLM-text setting—we compare V-CUE with two text-only baselines based on textual knowledge augmentation: MAPS (He et al., 2024) (knowledge retrieval) and Self-correct (Shinn et al., 2023) (self-reflection). Both rely solely on text and use no visual signals; prompt details are provided in the supplementary material. All methods are evaluated under LLM-text and LLM-VL settings on Multiple-Choice of CulturalBench.

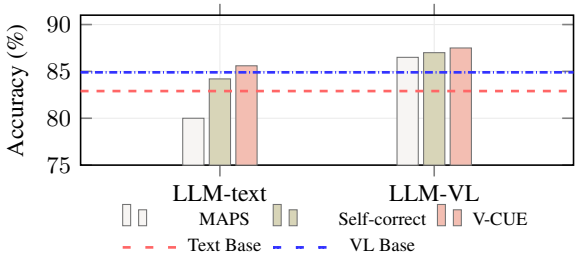

Figure 6: Comparison of text-only methods and V-CUE under LLM-text and LLM-VL settings.

Table 5: Comparison with retrieval-based augmentation (RAG) on CulturalBench (Multiple-Choice).

Method	**Accuracy (%)**
Text-only	82.9
RAG	84.1
V-CUE	85.6
RAG + V-CUE	**86.2**

As shown in Figure 6, text-only methods yield moderate gains, whereas V-CUE achieves stable and significant improvements after introducing visual signals in both settings, indicating that the gains stem from the visual cues rather than increased textual reasoning.

Complementarity with retrieval augmentation. A natural question is whether the generated visual cues could simply be replaced by retrieved text. Table 5 compares a text-only base, a retrieval-augmented variant (RAG) that retrieves the top-k (k=3) Wikipedia passages with BM25 and prepends them to the query, V-CUE, and their combination. V-CUE alone (85.6) outperforms RAG (84.1), and stacking the two yields the best accuracy (86.2): retrieval supplies factual context, whereas the image makes culture-specific cues already implicit in the question salient, so the two sources are complementary rather than redundant. Given that our main benchmark includes automatic scoring, we further validate whether these automatic judgments align with human expert ratings. The consistency analysis between CARE automatic evaluation and human expert judgments is reported in Appendix G.

7 Conclusion

We presented V-CUE, an inference-time framework that lets an LLM *see* culture when words fall short. A backbone-internal uncertainty detector decides when cultural grounding is needed, a text-to-image model renders the culture-specific cues implicit in the question, and the image is fed back to correct the answer. Across CulturalBench and CARE, V-CUE consistently improves LLM-text, LLM-reasoning, and LLM-VL systems without additional training, and uncertainty gating retains these gains at roughly half the inference cost. The improvement stems from the visual channel and the cultural content of the image, scales with visual–query alignment, and is complementary to retrieval. Future work will pursue finer cultural-symbol grounding to curb visual stereotypes and extend V-CUE to interactive multicultural settings.

AI USE STATEMENT

Generative AI tools were used to assist with initial drafting, translation, language editing, and consistency checks. The authors take full responsibility for the final text, claims, and related artifacts.

REFERENCES

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

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in llms: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 15763–15784. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.882. URL https://doi.org/10.18653/v1/2024.emnlp-main.882.

Thales Sales Almeida, Hugo Queiroz Abonizio, Rodrigo Nogueira, and Ramon Pires. Sabiá-2: A new generation of portuguese large language models. *CoRR*, abs/2403.09887, 2024. doi: 10.48550/ARXIV.2403.09887. URL https://doi.org/10.48550/arXiv.2403.09887.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *CoRR*, abs/2303.17466, 2023. doi: 10.48550/ARXIV.2303.17466. URL https://doi.org/10.48550/arXiv.2303.17466.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. Do localization methods actually localize memorized data in llms? A tale of two benchmarks. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 3190–3211. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.176. URL https://doi.org/10.18653/v1/2024.naacl-long.176.

Andong Chen, Yuchen Song, Kehai Chen, Xuefeng Bai, Muyun Yang, Liqiang Nie, Jie Liu, Tiejun Zhao, and Min Zhang. Make imagination clearer! stable diffusion-based visual imagination for multimodal machine translation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26567–26583, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1289. URL https://aclanthology.org/2025.acl-long.1289/.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1247. URL https://aclanthology.org/2025.acl-long.1247/.

Rochelle Choenni and Ekaterina Shutova. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *CoRR*, abs/2408.16482, 2024. doi: 10.48550/ARXIV.2408.16482. URL https://doi.org/10.48550/arXiv.2408.16482.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding,

Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. URL https://arxiv.org/abs/2304.06767.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. *CoRR*, abs/2306.16388, 2023. doi: 10.48550/ARXIV.2306.16388. URL https://doi.org/10.48550/arXiv.2306.16388.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis. *arXiv preprint arXiv:2505.19484*, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594, 2024. doi: 10.48550/ARXIV.2411.15594. URL https://doi.org/10.48550/arXiv.2411.15594.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. CARE: Multilingual human preference learning for cultural awareness. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32866–32895, Suzhou, China, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1669. URL https://aclanthology.org/2025.emnlp-main.1669/.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246, 03 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00642. URL https://doi.org/10.1162/tacl_a_00642.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. Acegpt, localizing large language models in arabic. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 8139–8163. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.450. URL https://doi.org/10.18653/v1/2024.naacl-long.450.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. Break the checkbox: Challenging closed-style evaluations of cultural alignment in llms. *CoRR*, abs/2502.08045, 2025. doi: 10.48550/ARXIV.2502.08045. URL https://doi.org/10.48550/arXiv.2502.08045.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 24174–24184. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02282. URL https://doi.org/10.1109/CVPR52733.2024.02282.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of LLM responses based on hofstede cultural dimensions. *CoRR*, abs/2406.14805, 2024. doi: 10.48550/ARXIV.2406.14805. URL https://doi.org/10.48550/arXiv.2406.14805.

Louis Kwok, Michal Bravansky, and Lewis D. Griffin. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *CoRR*, abs/2408.06929, 2024. doi: 10.48550/ARXIV.2408.06929. URL https://doi.org/10.48550/arXiv.2408.06929.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/9a16935bf54c4af233e25d998b7f4a2c-Abstract-Conference.html.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/77f089cd16dbc36ddd1caeb18446fbdd-Abstract-Conference.html.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 8474–8503. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.567/.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8eb88844dafefa92a26aaec9f3acad93-Abstract-Datasets_and_Benchmarks_Track.html.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 16366–16393. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.862. URL https://doi.org/10.18653/v1/2024.acl-long.862.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. Seallms - large language models for southeast asia. *CoRR*, abs/2312.00738, 2023. doi: 10.48550/ARXIV.2312.00738. URL https://doi.org/10.48550/arXiv.2312.00738.

OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Shramay Palta and Rachel Rudinger. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 9952–9962. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.631. URL https://doi.org/10.18653/v1/2023.findings-acl.631.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. *CoRR*, abs/2211.13069, 2022. doi: 10.48550/ARXIV.2211.13069. URL https://doi.org/10.48550/arXiv.2211.13069.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A framework for measuring the cultural adaptability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2373–2403, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério de Paula, and Diyi Yang. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.288. URL https://aclanthology.org/2024.findings-emnlp.288/.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

NAVER Cloud HyperCLOVA X Team. Hyperclova X THINK technical report. *CoRR*, abs/2506.22403, 2025a. doi: 10.48550/ARXIV.2506.22403. URL https://doi.org/10.48550/arXiv.2506.22403.

Qwen Team. Qwen2.5-vl, January 2025b. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Qwen Team. Qwen3 technical report, 2025c. URL https://arxiv.org/abs/2505.09388.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. Self-pluralising culture alignment for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6859–6877, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.350. URL https://aclanthology.org/2025.naacl-long.350/.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 2039–2055. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.132. URL https://doi.org/10.18653/v1/2022.emnlp-main.132.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023. doi: 10.48550/ARXIV.2306.05685. URL https://doi.org/10.48550/arXiv.2306.05685.

A Prompt Template for Uncertainty Detection

The Uncertainty Detection (UD) agent is implemented as an LLM-based judge that evaluates the reliability of a generated response under cultural contexts. The agent takes only the user query and the model-generated answer as input.

Uncertainty Detection Prompt

System Prompt:

You are an expert evaluator for cultural understanding tasks. Your job is to assess whether the given answer is culturally reliable and confident.

User Prompt:

[Query]: {User question text}
[Answer]: {Model-generated answer}

Please evaluate the answer according to the following criteria:

1. Cultural correctness (Is the answer aligned with the cultural background implied by the query?)
2. Cultural specificity (Does the answer rely on concrete cultural knowledge rather than generic statements?)
3. Confidence and consistency (Is the answer internally consistent and decisive, without hedging or uncertainty?)

Return only one label: `True` or `False`. Do not explain your decision.

B Prompt for Cultural Cue Extraction

To construct the cultural condition unit (c), we adopt an LLM-based structured parsing prompt to extract three types of cultural elements: `[region]`, `[object]`, and `[symbol]` from the input question text. To avoid label leakage, the parsing model is strictly restricted to *question-only* input, without access to answer options or gold labels.

We use a fixed instruction prompt to enforce structured extraction rather than task solving. The prompt is shown below:

Cultural Cue Extraction Prompt

System Prompt:

You are a cultural information parser. Your task is to extract culturally relevant elements from the input question according to a predefined schema.

Rules:

- Only use information explicitly contained in the question text.
- Do not infer or predict the correct answer.
- If an element is not mentioned or cannot be inferred, output `None`.

Output Format (JSON only):

```
{
  "region": "...",
  "object": "...",
  "symbol": "..."
}
```

User Prompt:

[Question]: <*question text*>

The extracted elements are concatenated to form the conditioning text (c = `[region]` + `[object]` + `[symbol]`), which is used solely for visual generation conditioning and does not encode answer-related information.

C NEUTRAL VISUALS CONSTRUCTION

To obtain culturally neutral images, we explicitly remove culture-specific elements from the extracted visual cues (region, object, and symbol) using DeepSeek-V3. The goal is not paraphrasing but de-culturalization, i.e., preserving generic visual content while eliminating identifiable cultural markers. The neutralized text is then fed into Stable Diffusion to generate images. This ensures that Neutral Visuals retain comparable visual complexity without conveying cultural signals.

Cultural Cue Removal Prompt (DeepSeek-V3)

System Prompt:

You are a cultural cue rewriter. Your task is to neutralize culturally specific visual elements by removing identifiable cultural meaning while preserving generic visual content.

Rules:

- Rewrite each element to remove culture-specific meaning.
- Replace culturally identifiable terms with culturally neutral, generic alternatives.
- Do not add new information or attributes.
- If an element is already culturally neutral, keep it unchanged.

Output Format (JSON only):

```
{
  "region": "...",
  "object": "...",
  "symbol": "..."
}
```

User Prompt:

[Input Elements]: <*region, object, symbol*>

For the object dimension, a culturally specific term such as *Arabic headscarf* is neutralized to the generic form *headscarf*. For symbols, culturally identifiable icons like the *crescent and star* are replaced with a culturally neutral description, namely a *geometric symbol*. For regions, explicit geographic or cultural references such as the *Middle East* are generalized to a non-identifiable category, such as an *urban area*. The resulting neutralized elements are concatenated and used solely for image generation.

D PROMPT DETAILS FOR MAPS AND SELF-CORRECT

The MAPS framework mimics professional human translation by analyzing the source text through three aspects of knowledge: keywords, topics, and relevant demonstrations.

MAPS (Multi-Aspect Prompting and Selection)

Knowledge Mining Prompts:

- **Keywords:** "Extract the keywords in the following sentence."
- **Topics:** "Use a few words to describe the cultural topics of the following input sentence."
- **Demonstrations:** "Write an sentence related to but different from the input cultural query and answer."

The extracted knowledge is integrated into a final prompt for candidate generation, followed by a selection mechanism based on LLM-as-Judge to filter noisy or unhelpful knowledge.

The Self-Correct framework reinforces language agents through verbal reinforcement learning. It converts environment feedback (binary or scalar) into a textual summary called a "semantic gradient," which is stored in an episodic memory buffer.

Self-Correct

System Architecture:

- **Actor (M_a):** Specifically prompted to generate text and actions conditioned on state observations.
- **Evaluator (M_e):** Scores the outputs produced by the Actor to reflect performance within the task context.
- **Self-Reflection (M_{sr}):** Generates verbal reinforcement cues to assist the Actor in identifying and correcting prior mistakes.

E Generation Settings

All experiments are conducted on a single NVIDIA A100 (80GB) GPU. For image generation, we use Stable Diffusion v2.1 with fixed and shared parameters across all experiments. Specifically, the guidance scale is set to 7.5, the number of denoising steps is 30, and the image resolution is 512×512. To ensure reproducibility and control variance, we use a single fixed random seed per sample and do not perform multiple sampling or reranking. These settings are kept constant across all evaluations to avoid performance and efficiency deviations caused by variations in model hyperparameters.

F Human Evaluation Details

F.1 Annotator Background and Selection

To ensure the professionalism and consistency of our evaluation, we recruited three annotators with language backgrounds in Chinese, Arabic, and Japanese, as detailed below:

- **Annotator 1 (Chinese)**: A native Chinese speaker with strong proficiency in Chinese and extensive experience in academic reading and writing. The annotator has a solid understanding of common cultural references, social norms, and conventional expressions in Chinese contexts, and can identify culture-related issues in generation outputs.
- **Annotator 2 (Arabic)**: Proficient in Arabic (Modern Standard Arabic) with formal training in linguistics or area studies. The annotator understands cultural norms in Arabic contexts, including religion, etiquette, forms of address, and taboos, and can judge whether outputs exhibit cultural mismatch, pragmatic inappropriateness, or potentially offensive expressions.
- **Annotator 3 (Japanese)**: Proficient in Japanese with strong reading and writing competence. The annotator is familiar with Japanese pragmatics (e.g., honorifics, indirectness, and social distance) and its cultural context, and can evaluate the appropriateness of politeness strategies, tone, and cultural references in model outputs.

The diverse backgrounds of the annotators help analyze the model's performance from multiple perspectives, reducing individual biases.

G Consistency Between Automatic and Human Evaluation on CARE

We assess the consistency between CARE automatic evaluation (LLM-as-Judge) and human judgments. Samples are drawn from the CARE test set (Chinese, Japanese, Arabic). Three expert annotators with formal training in linguistics or cultural studies independently score each response on Generality, Cultural Relevance, and Response Quality (1–10). The final score represents the average of these three dimensions. All annotators are fluent in the target languages and have prior experience evaluating culture-sensitive language generation tasks. Human scores are averaged across annotators. Agreement is measured using Pearson r, Spearman ρ, and Krippendorff's α.

As shown in Table 6, CARE shows strong agreement with human judgments across all dimensions, with the highest consistency on Cultural Relevance. This confirms the reliability of CARE as a proxy for human evaluation.

Table 6: Agreement between CARE automatic and human evaluation.

Dim.	r	ρ	α
Generality	0.91	0.89	0.82
Cultural Rel.	0.94	0.92	0.86
Literary Qual.	0.88	0.87	0.79
Avg.	0.91	0.89	0.82

H Effect of Visual Quality on V-CUE Performance

Table 7: Accuracy comparison of Qwen models using V-CUE under high- vs. low-quality visual inputs (Stable Diffusion w/ vs. w/o EMA). V-CUE improves performance across all systems, but lower-quality visuals reduce the gain.

Model	**Multiple-Choice**		**True or False**	
	High	**Low**	**High**	**Low**
LLM-text systems (w/ V-CUE)				
Qwen3-235B-A22B	85.61	84.20	89.01	87.12
Qwen3-32B	82.87	81.23	81.54	79.67
LLM-reasoning systems (w/ V-CUE)				
Qwen3-235B-A22B	88.83	85.64	93.27	91.22
Qwen3-32B	83.26	80.78	89.38	85.99
LLM-VL systems (w/ V-CUE)				
Qwen2.5-vl-72B	87.52	86.40	90.04	88.90

We evaluate the impact of visual quality on V-CUE using three Qwen-series models across LLM-text, LLM-reasoning, and LLM-VL settings. V-CUE is applied with Stable Diffusion images generated either with Exponential Moving Average (EMA) or without EMA. EMA (Karras et al., 2024) produces more stable, high-fidelity images at inference.

As shown in Table 7, V-CUE consistently improves accuracy, but the gains are highly sensitive to visual quality. EMA-based visuals yield substantially larger improvements on both Multiple-Choice and True/False tasks, particularly for LLM-reasoning systems. In contrast, lower-quality visuals lead to reduced gains, especially for smaller text-based LLMs with weaker grounding. These results indicate that high-quality visual cues are crucial for maximizing the cultural reasoning benefits of V-CUE.

I Cultural Recovery Rate

To assess the impact of V-CUE on correcting culturally mistaken responses, we conducted an error-revision experiment across LLM-text, LLM-reasoning, and LLM-VL systems on CulturalBench. For each model, we computed the correction rate. The formula is:

$$\text{Correction Rate} = \frac{\#(\text{Incorrect} \rightarrow \text{Correct})}{\#(\text{Original Incorrect})}. \tag{9}$$

Results from the Figure 7 show that all models show noticeable improvements in correction rate after applying V-CUE, confirming its overall effectiveness in cultural error revision. LLM-reasoning systems achieve significantly higher correction rates (e.g., 52.79% on multiple-choice), indicating stronger receptiveness to external cues. This suggests that reasoning-oriented models are better able to leverage visual-cultural signals for self-correction.

J Generalization Across Image-to-Text Mediation Mechanisms

For the LLM-text setting in V-CUE, we use an Image-to-Text Mediation Mechanism (Cap(V) in Eq. (7)) to convert images into textual representations. We evaluate its generalization via (i) scalability

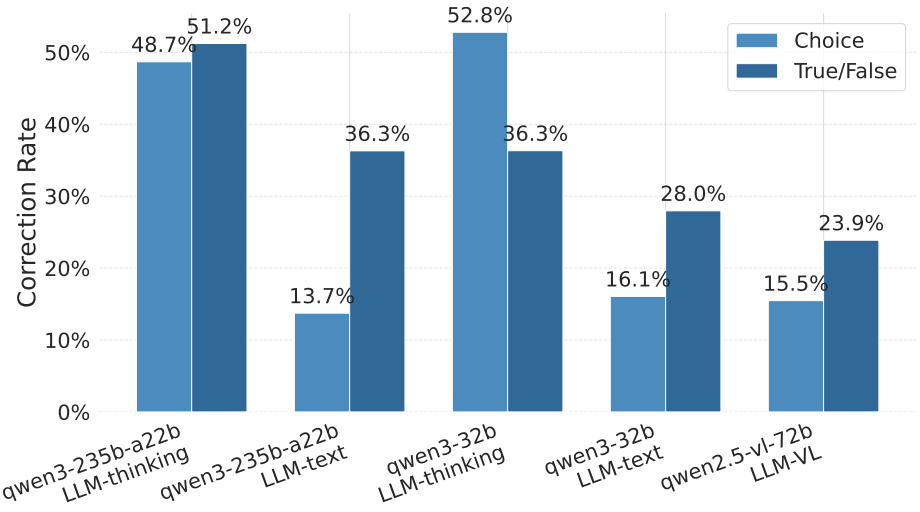

Figure 7: V-CUE significantly improves correction rates on CulturalBench, especially for LLM-reasoning systems and True or False tasks across Qwen models.

Table 8: Performance comparison on CulturalBench. Bold denotes the best result. The baseline directly uses DeepSeek-V3, whereas V-CUE incorporates advanced vision–language models as mediators.

Method	**Visual Mediator**	**Accuracy (%)**	
		Multiple-Choice	**True or False**
DeepSeek-V3		81.66	83.31
V-CUE (Ours)	*w/* Qwen2.5-VL-7B	81.92	82.79
	w/ Qwen2.5-VL-14B	82.71	84.98
	w/ Qwen2.5-VL-32B	83.68	86.31
	w/ Qwen2.5-VL-72B	**84.84**	89.87
	w/ GPT-4o	84.12	89.21
	w/ Gemini-2.5 Pro	84.66	**90.11**

across Qwen2.5-VL models from 7B to 72B, and (ii) cross-model transfer by comparing Qwen2.5-VL-72B with GPT-4o and Gemini-2.5 Pro. To avoid family-specific bias, DeepSeek-V3 is fixed as the LLM-text backbone. All experiments are conducted on CulturalBench.

As shown in the Table 8, results show that V-CUE scales consistently with model size and achieves clear gains over the DeepSeek-V3 baseline across different visual mediators, indicating robustness to the choice of Image-to-Text mediation mechanism.