NORMVIZ: A Benchmark and Framework for Grounding Multimodal Reasoning in Global Cultures

Akhila Yerukola[♡] **Fabrice Harel-Canada**[♣] **Simran Khanuja**[♡] **Abhinav Rao**[♡]

Ashima Suvarna[♣] **Nanyun Peng**[♣] **Saadia Gabriel**[♣] **Maarten Sap**[♡]

[♡]Carnegie Mellon University [♣]University of California, Los Angeles

ayerukol@andrew.cmu.edu

Abstract

AI systems are used worldwide, but they struggle to serve the needs of culturally diverse populations. Prior work on cultural understanding evaluates AI systems on text-only settings or on visual artifact recognition (e.g. foods, clothing). The ability to reason about visually observable behaviors through local social norms, which we call *visual norm understanding*, remains unexamined. We introduce NORMVIZ-BENCH, a high quality, human-validated benchmark of 3,272 contrastive image pairs (6,544 images) spanning 16 countries. Each pair varies only in the culturally relevant behavior (e.g., objects, attributes, spatial relations, and actions) that alters how each image is interpreted. Each image is labeled as conforming to, violating, or irrelevant to local social norms, and pair-level evaluation requires both images to be correctly classified, thereby preventing reliance on superficial visual shortcuts. Even the strongest VLMs, Gemini 3.0 Flash and Qwen2.5 VL 7B, succeed on only **25.3%** and **23.0%** of pairs, struggling most with identifying violating and culturally benign visual behaviors. Towards bridging this, we introduce NORMVIZ-TRAIN, a training dataset of 64*k* images paired with explanations. Though absolute performance remains low ($< 30\%$), finetuning on NORMVIZ-TRAIN improves pair accuracy relatively by up to 86% and 36% for Qwen3-VL 4B and 8B respectively, showing a path forward to teach models to connect visual perception to cultural significance. Together, NORMVIZ-BENCH and NORMVIZ-TRAIN establish visual norm understanding as a challenging and consequential frontier for multimodal AI. We make our code and data publicly available. [1]

1 Introduction

As multimodal AI systems are increasingly deployed across culturally diverse settings, the ability to identify and reason about *visually observable* social dynamics is a critical capability, which we refer to as **visual norm understanding** (Matsumoto & Hwang, 2016; Knapp, 1972; Geertz, 2017). Consider a celebratory event in Japan where a guest gifts a ceramic plate with four koi fish (Fig. 1 (a), left image). A Japanese native would recognize this as culturally offensive (in Japanese the number 4, 'shi', is homophonous with the word for 'death'; Simon, 1952); in contrast, a plate with three fish (Fig. 1 (a), right image) would be culturally benign — neither meaningful nor offensive. A crucial step towards AI systems effectively serving users across global contexts is this ability to not just recognize *what* is in an image, but also understand *how* specific visual behaviors relate to local social and cultural norms.

Yet, current AI training and evaluation has largely ignored this critical ability. Prior work on evaluating cultural understanding in AI has mostly operated in text-only settings, where cultural context is explicitly provided as input (Durmus et al., 2023; Rao et al., 2024; Kabra et al., 2023). Efforts in multimodal understanding, meanwhile, have focused on *cultural objects*

[1]Code and data available at: https://github.com/Akhila-Yerukola/NormViz

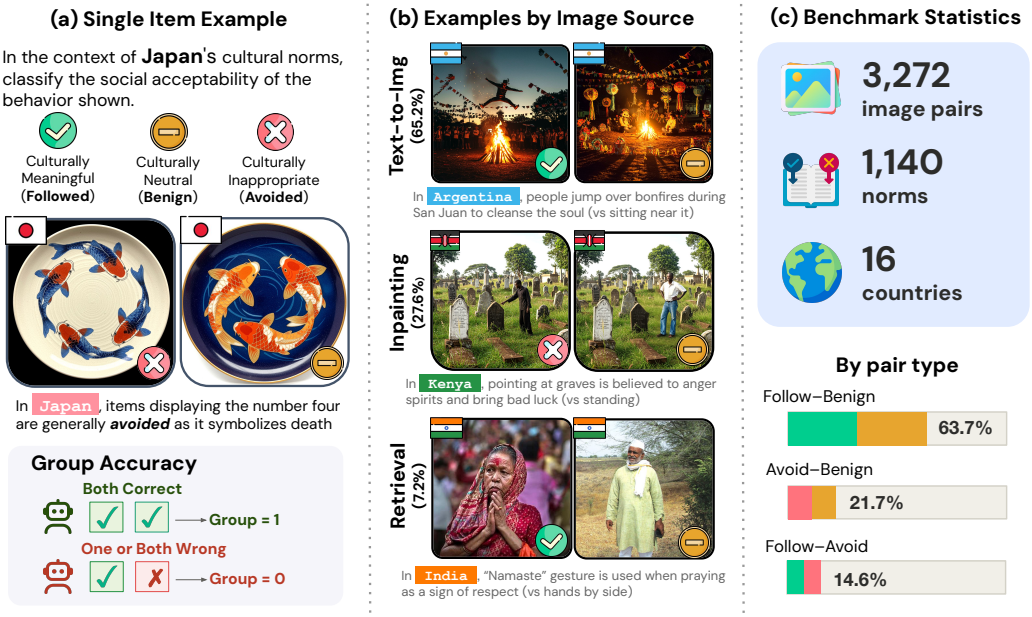


Figure 1: Overview of NORMVIZ-BENCH for visual norm understanding, consisting of 3,272 contrastive image pairs across 16 countries that differ only in the culturally relevant behavior. (a) A contrastive pair example from Japan; group accuracy requires a model to correctly classify both images in a pair. (b) Images are sourced via Text-to-Image generation, inpainting, and Retrieval. (c) Benchmark statistics and pair type distribution.

and artifacts such as foods, clothing, and landmarks (Urailertprasert et al., 2024; Nayak et al., 2024; Romero et al., 2024). *Culturally situated behaviors*, such as gifting practices, are hard to isolate in web-scale image data (Van Miltenburg et al., 2017; Slingerland et al., 2020), making it difficult to curate data that goes beyond artifact recognition to visual norm reasoning.

To fill this gap, we introduce NORMVIZ-BENCH, a high-quality, diverse, and human-validated benchmark of 3,272 contrastive image pairs (6,544 images) for evaluating visual norm understanding across 16 countries. We adopt contrastive pair evaluation since prior work show models often exploit superficial shortcuts in single-image evaluative settings (e.g., easier to detect a cup 'on' vs. 'under' a table) whereas contrastive minimal-pair designs are more robust to such biases (Thrush et al., 2022; Kamath et al., 2023). In our benchmark, each pair differs only in the culturally relevant behavior, requiring both visual perception and normative judgment. Given each image, models must determine if the depicted behavior conforms to (*Follow*), violates (*Avoid*), or is irrelevant (*Benign*) to local social norms, and must classify both images correctly. Visually detectable cultural norms are sourced from three human-validated text-based cultural understanding benchmarks (Chiu et al., 2024; Rao et al., 2024; Yin et al., 2024), transformed into images with benign counterparts through text-to-image generation, inpainting, and real image retrieval, with each pair rigorously validated by qualified annotators from the target country (Figure 1).

Using NORMVIZ-BENCH, we evaluate 21 vision-language models (VLMs) using *group accuracy*, the percentage of contrastive pairs where both images are correctly classified. We find that even the strongest open and closed-source models, Qwen2.5 VL-7B and Gemini 3.0 Flash, achieve only **23.0%** and **25.3%** respectively. Models struggle most with behaviors that violate social norms or are culturally irrelevant, and perform worst on cultures from the Global South, including India, Indonesia, and Kenya. To disentangle sources of VLM failures, we provide models with (i) image descriptions alongside images, and (ii) descriptions in place of images. Performance improves in both settings, yet the strongest models achieve below 45.3% in group accuracy. This gap suggests that cultural normative reasoning, not visual perception or vision-language misalignment, is the hardest barrier.

To improve visual norm understanding, we introduce NORMVIZ-TRAIN, a large-scale synthetic training dataset of 64*k* images paired with explanations of the depicted behavior and its cultural significance. We source cultural norms from CANDLE, a cultural knowledge

graph (Nguyen et al., 2023), and generate images using Stable Diffusion 3 (SD3). Supervised finetuning Qwen3-VL models on NORMVIZ-TRAIN yields meaningful improvements in group accuracy across model sizes: Qwen3-VL 4B improves group accuracy by 86% (14.9% to 27.7%); Qwen3-VL 8B improves by 36% (19.6% to 26.7%). With 64k training examples, both models **beat Gemini 3.0 Flash** (25.3%). Despite these significant gains, absolute group accuracy remains below 30%, establishing visual norm understanding as an open and consequential challenge that NORMVIZ-BENCH and NORMVIZ-TRAIN together help advance.

2 Related Work

AI Benchmarks for Cultural Understanding. Cultural competence in AI requires not just knowledge of diverse cultures, but also the ability to reason about cultural norms when interacting with or representing pluralistic needs (Deardorff, 2009; Bhatt & Diaz, 2024). Most AI benchmarks studying this have focused solely on text-based evaluation, assessing knowledge of cultural artifacts and facts (Chiu et al., 2024; Seth et al., 2024; Singh et al., 2025), cultural values and beliefs (Durmus et al., 2023; Arora et al., 2023; Cao et al., 2023; Ramezani & Xu, 2023), and understanding social norms (Rao et al., 2024; Shi et al., 2024).

Multimodal benchmarks, in contrast, have primarily focused on recognition of cultural artifacts such as food, clothing, and landmarks (Liu et al., 2021; Romero et al., 2024; Nayak et al., 2024; Vayani et al., 2025; Yue et al., 2024; Schneider et al., 2025; Satar et al., 2025; Kim et al., 2025a), essentially evaluating object recognition rather than compositional reasoning. Cultural norms, however, depend on compositional reasoning of social behaviors: understanding not just what is present in a scene, but who performs an action, toward whom, and in what cultural setting (Knapp, 1972; Burgoon et al., 2011). VLMs are known to struggle with this kind of reasoning in general settings, exhibiting a bag-of-words failure mode and performing near chance on tasks requiring sensitivity to relational structure, object attributes, or actions (Yuksekgonul et al., 2022; Ma et al., 2023; Li et al., 2024a; Thrush et al., 2022; Parcalabescu et al., 2022). Yet whether this weakness extends to culturally situated visual settings remains unexplored (Yerukola et al., 2025; Kim et al., 2025b; Malakouti et al., 2025). No existing work, to our knowledge, has systematically evaluated the visual norm understanding of culturally situated actions, gestures, and behaviors in VLMs.

Data Bottleneck for Training Culturally Grounded Models. A key bottleneck for cultural alignment is the difficulty of mining high-quality data at scale that encodes not just cultural facts but also norms, values, and situated behaviors (Pawar et al., 2025). In the text domain, recent work has addressed this through synthetic data such as LLM-generated cultural dialogues (Li et al., 2024c;b; Xu et al., 2025), multilingual critique data (Feng et al., 2025), and cross-cultural blended data (Yuan et al., 2024). In the visual modality, the scarcity problem is more acute: cultural norms and social practices are difficult to isolate, label, and index in web data (Slingerland et al., 2020), leaving multimodal approaches largely confined to artifact-centric image-QA pairs (Liu et al., 2025; Nyandwi et al., 2025) or real images curated for cultural safety (Qiu et al., 2025). No existing work has crossed the modality gap, using culturally grounded textual knowledge to generate visual training data covering social behaviors and norms that are systematically absent from naturally occurring image corpora.

3 NORMVIZ-BENCH Construction

We curate NORMVIZ-BENCH, a benchmark of contrastive image pairs for evaluating visual norm understanding in VLMs. Our pipeline (Fig. 2) transforms cultural norm descriptions into image pairs through five stages: (1) norm extraction and structuring (§3.1), (2) metadata extraction and filtering for visual detectability (§3.2), (3) prompt construction (§3.3), (4) multi-source image curation (§3.4), and (5) human validation (§3.5).

Task Design Motivation. Prior work on (general) visual compositional reasoning has shown that single-image evaluation design fails to expose the visual and linguistic shortcuts that models exploit (Yuksekgonul et al., 2022; Ma et al., 2023; Li et al., 2024a; Hsieh et al., 2023). Thus, we adopt a contrastive minimal-pair evaluation setup (Gardner et al., 2020;

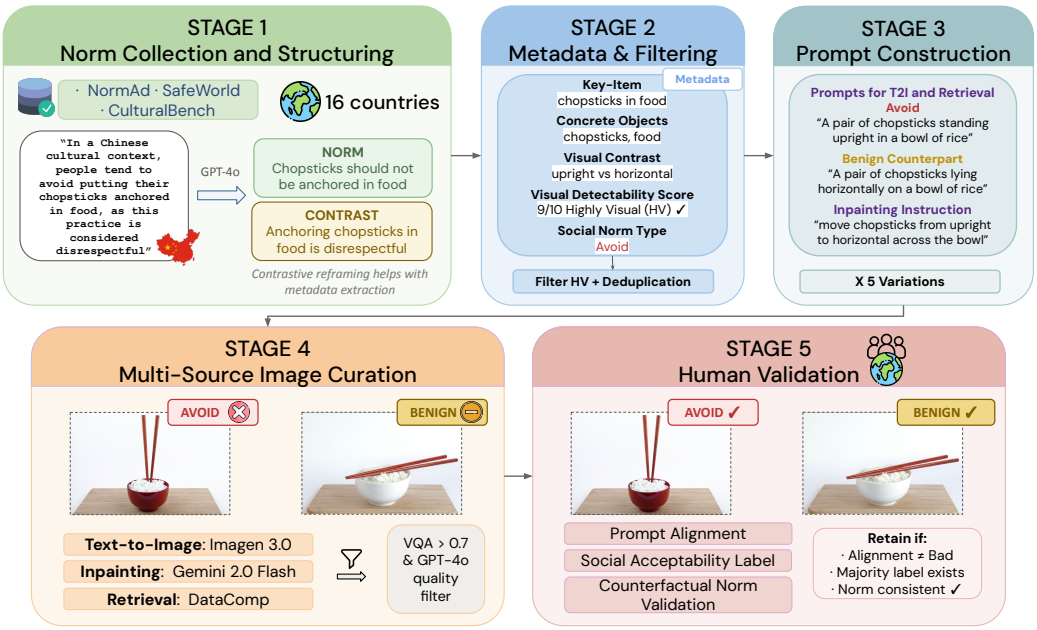


Figure 2: NORMVIZ-BENCH construction pipeline across five stages: (1) collecting cultural norms from 3 text benchmarks and transforming them into norm-contrast pairs; (2) extracting metadata and social norm type, and then filtering for visual detectability; (3) constructing prompt descriptions of the contrasting pair, along with inpainting instructions; (4) curating images via text-to-image generation, inpainting, and retrieval, with initial quality filtering; and (5) human validation of the entire set from annotators from each of the 16 countries.

Parcalabescu et al., 2022; Thrush et al., 2022), where the two images differ only in culturally relevant behavior (e.g, number of items in Figure 1 (a)), and models must correctly classify both. A model that can correctly classify only one of the images might be relying on superficial features, rather than truly understanding the cultural norm.

Task Formulation. Each image in these pairs depicts nonverbal visual behaviors (gestures, object attributes, spatial arrangements, or actions). Given an image and its country context, models must independently classify the depicted behavior as Culturally Significant (*Follow*), Culturally Inappropriate (*Avoid*), or Culturally Neutral (*Benign*). We evaluate performance through two metrics: (1) pairwise classification via **group accuracy**, where both images in a pair must be correctly classified, (2) **single-image classification F1** performance.

3.1 Stage 1: Norm Collection and Structuring

Norm Collection. We source descriptions of cultural norms from three text benchmarks: CulturalBench (Chiu et al., 2024), NormAd (Rao et al., 2024), and SafeWorld (Yin et al., 2024), selected for their broad cultural coverage and extensive human validation. 'Culture' is multifaceted, and we focus on several visual nuances covered across these benchmarks, such as social etiquette, religious practices, celebrations, and daily life. We target 16 countries prioritizing the Global South and based on data availability: Argentina, Brazil, China, Egypt, France, India, Indonesia, Italy, Japan, Kenya, Nigeria, Peru, South Korea, Turkey, United States, and Vietnam (spans 10 UN geoscheme regions). While culture transcends national borders, countries serve as a meaningful proxy (Adilazuarda et al., 2024), reflecting shared history and lived experience, and providing a practical starting point for geo-tagging data.

Norm Structuring. We transform each cultural description into a *norm–contrast* pair using GPT-4o,[2] extracting a behavioral norm (e.g., "chopsticks should not be anchored in food")

[2]We qualitatively validated GPT-4o outputs at every stage of our pipeline, and the final test set undergoes comprehensive human validation (§3.5)

and rephrasing it as a contrastive statement that preserves its meaning (e.g., "anchoring chopsticks in food is considered disrespectful"). We found that visual metadata extraction is significantly more reliable when using norm–contrast pairs rather than norms alone.

3.2 Stage 2: Metadata and Filtering

Metadata Extraction. Not all cultural norms can be meaningfully depicted in images. For example, norms about attitudes (e.g., "remain humble") or verbal conventions (e.g., "honorific titles for elders") lack concrete visual anchors. To identify visually observable norms, we use GPT-4o to extract structured metadata for each *norm–contrast* pair, spanning *key items*, *concrete items*, *visual contrast*, among others (see Appendix D.2 for more details). We additionally classify each norm as *Follow* or *Avoid* based on its social norm type.

Visual Detectability Filtering and Deduplication. Using the extracted metadata, we prompt GPT-4o to score each norm on visual detectability (0–10) based on clarity of representation and ease of distinguishability, retaining norms scoring $\geq 7/10$ as Highly Visual (HV). For example, "the *bride*'s family gifts a Japanese tea set" (Low Visual; LV) is filtered out as it's hard to visually determine which family gave the gift. We then embed the filtered *key items* and apply K-means clustering to merge semantically equivalent norms across the three source benchmarks, reducing the norm set by two-fold.

3.3 Stage 3: Prompt Construction

For each norm classified as either *Follow* or *Avoid*, we use GPT-4o to generate three prompt types conditioned on the extracted metadata: (1) ***Follow*/*Avoid* image descriptions**: detailed scene descriptions depicting the relevant social norm behavior; (2) **Benign counterfactual descriptions**: scenes retaining the key item in a culturally neutral configuration—neither meaningful nor offensive (e.g., "A person holding chopsticks while reaching for food"). We construct *Benign* counterfactuals for both the norms to *Follow* ($Benign_{Follow}$) and *Avoid* ($Benign_{Avoid}$); (3) **Inpainting instructions**: minimal editing directives to transform a *Follow/Avoid* image into its Benign counterpart (e.g., "Change the red envelope to a yellow envelope"). We generate 5 variations of prompts per category to increase scene diversity.

3.4 Stage 4: Multi-Source Image Curation

For each norm, we curate both synthetic and real images to ensure diversity in visual style. With a growing prevalence of AI-generated content, there is a need to detect cultural inconsistencies in generated images, as well as to understand how social norms manifest in real images within target cultures.

Text-to-Image Generation. We first generate images for both *Avoid/Follow* and *Benign* behaviors independently. We select `Imagen 3.0` for its high-quality and culturally competent image generation (Kannen et al., 2024). We generate 2 images per prompt variation.

Inpainting. To create challenging image pairs with minimal variation, we apply inpainting instructions only to the generated *Follow/Avoid* images (e.g., change a red envelope to yellow). We select Gemini 2.0 Flash (Comanici et al., 2025) for its strong inpainting capabilities that preserve image coherence while making precise edits to produce *Benign* counterparts.

Real Image Retrieval. Generated images may not capture how these behaviors naturally appear in the target country, so we retrieve from country specific image databases instead. We use the *Avoid, Follow, Benign* prompts to query country-specific subsets of DataComp (Gadre et al., 2023), and retrieve up to 20 candidates per prompt.

Quality Filtering. Prompt-image alignment is difficult to control at scale, particularly when models misinterpret or refuse culturally specific prompts (Yerukola et al., 2025). Before human validation (§3.5), we apply two types of automatic filtering: (1) VQA-based scoring (Lin et al., 2024), retaining images with semantic alignment above 0.7 (scale 0–1), and (2) GPT-4o-based evaluation verifying that key objects, attributes, and spatial

relationships are correctly depicted, classifying each image as good, partial, or bad. Only good images are retained.

3.5 Human Validation

To ensure the ecological validity of our benchmark, we recruited cultural in-group annotators from each of the 16 countries via Prolific (https://www.prolific.com/). Annotators were prescreened using a pre-qualification test to ensure annotation quality. We obtain 3 annotations per image pair. For each pair, annotators perform three tasks: (1) rate the **prompt-image alignment** for each image (good, partial, or bad); (2) independently **classify each image** as culturally appropriate, culturally inappropriate, or culturally benign; and (3) verify that **both images depict the same cultural norm** (*counterfactual validation*). To ensure benchmark quality, we retain only pairs where (1) neither image is rated as a bad prompt-image match, (2) annotators reach majority agreement on cultural alignment labels, and (3) norm consistency is confirmed. Since labels are assigned freely, the resulting pairs are not constrained to the intended *Follow/Avoid–Benign* structure, resulting in *Avoid–Benign*, *Follow–Benign*, and *Follow–Avoid* pairs, allowing annotators to surface and correct mislabeled or culturally ambiguous cases. We filter out pairs with the same label.

After our initial quality filtering (§3.4), we obtained 8,456 candidate image pairs for annotation. We recruited 212 annotators across 16 countries (113 male, 98 female, 1 undisclosed), with each annotator evaluating 70-150 pairs (varied due to availability). After filtering spam annotations, we retained 21,743 annotations. Of these, 38.6% passed rigorous quality checks, yielding the final benchmark of 3,272 image pairs across 16 countries. This rigorous filtering process also specifically guards against any potential AI generator-introduced biases being passed through to the benchmark. Please find details on the prescreening, annotation scheme, IRB approval and fair pay in Appendix B.

3.6 NORMVIZ-BENCH Characteristics

Our final benchmark comprises **3,272 contrastive image pairs** (**6,544 images**; Table 3) spanning **16 countries** from three image sources: *text-to-image generation* (65.2%), *inpainting* (27.6%), and *natural image retrieval* (7.2%), with three pair types: *Follow–Benign* (63.7%), *Avoid–Benign* (21.7%), and *Follow–Avoid* (14.6%). Notably, the retrieval subset of 7.2% reflects how rarely these norms manifest in naturally occurring real images, and their retrieval is non-trivial. NORMVIZ-BENCH covers a semantically rich range of cultural nuances—including public conduct, greetings and gestures, festivals, symbolic objects and practices, gift-giving and dining customs—and requires diverse visual reasoning over contrasting object states, actions, presence/absence, and gesture differences.

The per-country distribution in NORMVIZ-BENCH is uneven (Figure 8; Vietnam: 410 pairs, China: 344, vs. United States: 92), partly due to the country coverage decisions inherited from our seed benchmarks (see §3.1), but also reflects a well-established distinction between high-context and low-context cultures in cross-cultural communication theory (Hall, 1976; 2000). High-context cultures (e.g., Japan, China) embed meaning in nonverbal cues, body language, and situational context, which are precisely the visually observable norms our benchmark targets, whereas low-context cultures (e.g., the United States) rely more on explicit verbal communication and thus exhibit fewer visually distinctive, detectable nonverbal norms. Please refer to Appendix C and D for full details on benchmark construction, prompts used, and benchmark characteristics.

4 Benchmarking: Setup, Results, and Analysis

Setup and Metrics. We evaluate 21 VLMs covering proprietary and open-weight systems. We prompt models with each image and ask them to classify its social acceptability within a specified cultural context: *"In the context of {country}'s cultural and social norms, classify the social acceptability of the behavior shown in this image."*. Models choose from three options: *(A) Culturally Meaningful/Significant, (B) Culturally Inappropriate/Offensive, or (C) Culturally*

Benign/Neutral. We set temperature to 0.0 for all experiments. We report two metrics: (1) **Group accuracy** for pairwise classification, and (2) **macro F1** for single-image classification.

Condition	*Closed source*			*Open source*									Avg
	Gemini 3 flash	GPT 5.2	GPT 4o	Qwen 2.5 7B	LLaVA OV 7B	Qwen 3 8B	Qwen 2.5 3B	GLM-4V 9B	Phi-3V 4.2B	Aya Vis. 8B	Qwen 3 4B	DS-vl2 1B	
Group Accuracy (%)													
NORMVIZ-BENCH	**25.3**	22.7	16.3	**23.0**	19.8	19.6	18.3	16.7	16.1	15.4	14.9	13.8	**18.5**
oracle + Image Descriptions	32.7	37.8	23.5	30.8	27.3	28.1	27.5	23.1	29.9	26.7	23.0	19.7	**27.5**
oracle + Image Desc. (w/o img)	40.8	45.3	37.2	32.7	34.3	40.2	29.7	31.0	35.8	28.7	30.7	2.7	**32.4**
Macro F1													
NORMVIZ-BENCH	**55.8**	45.6	41.5	44.2	44.3	43.6	**44.6**	42.5	40.2	37.3	37.7	40.6	**43.2**
oracle + Image Descriptions	63.6	60.8	50.5	48.9	48.5	50.0	52.2	47.4	49.9	44.3	43.7	45.4	**50.4**
oracle + Image Desc. (w/o img)	68.0	66.5	62.8	53.2	53.6	60.2	56.0	53.9	56.4	48.5	52.9	26.8	**54.9**

Table 1: Performance across the top 12 evaluated models, ordered by group accuracy within closed- and open- source groups. **Best** and second-best model are highlighted. **Group accuracy for all models is critically low**. We run two oracle ablations: (1) + *Image Descriptions* - image with text description; (2) + *Image Desc. (w/o img)* - replace input image entirely with text description. We find that performance generally improves in both settings, indicating cultural visual perception and vision-language alignment are bottlenecks. However, the best model under these settings achieves 45.3% group accuracy (GPT 5.2), suggesting that cultural normative reasoning is the biggest barrier. Full results for all 21 models in App. E.

Main results. Current VLMs struggle significantly with visually grounded cultural reasoning. Across all models, **group accuracy is strikingly low**, with the best-performing models Gemini 3.0 flash, GPT 5.2, Qwen2.5 VL 7B, and LLaVA OneVision-7B achieving only 25.3–19.8% (see Table 1; full results in Table 4). **Macro F1 scores for single-image classification are also weak**, ranging from 55.8 to 44.2 across these top models. This gap reveals a fundamental limitation: since contrastive pairs differ only in a culturally relevant visual detail, getting both images right is necessary for assessing visual norm understanding. Relatively stronger F1 but strikingly low group accuracy suggests models rely on shallow cultural associations rather than the subtle visual distinctions that separate these categories. Performance further varies substantially across regions: **Sub-Saharan African, Southern Asian, and Southeast Asian regions show the lowest group accuracy**, while Western European and South American countries (particularly Argentina and Brazil) perform comparatively better (Appendix E.3). Finally, while larger models tend to perform slightly better, **scaling effects are inconsistent across model families**. Refer to Appendix E for more details.

Three Barriers: Visual Perception, Cultural Vision-Language Alignment, and Cultural Normative Reasoning To probe whether failures stem from visual perception, vision-language misalignment, or cultural reasoning, we conduct two oracle ablations using NORMVIZ-BENCH (rows 2, 3 in Table 1). First, we provide models with ground-truth image descriptions alongside the image (+ *Image Descriptions*).[3] Second, we remove the image entirely, providing only the image description (+ *Image Desc. w/o img*). We note that these ablations do not perfectly isolate each factor and are best interpreted as generous upper bounds rather than clean separations. Nonetheless, the consistent significant performance improvements are indicative of several barriers.

The first ablation yields consistent improvements (avg. +9.0 group accuracy, +7.3 F1; Wilcoxon signed-rank across models, $p < 0.001$ for both metrics), indicating that visual perception is a significant bottleneck with VLMs struggling to correctly interpret what is depicted in culturally-laden images. Yet absolute performance achieved remains critically low (avg. 27.5% group accuracy, 50.4 F1), suggesting that perception alone does not explain the failures. Strikingly, removing the image entirely improves performance further (avg. 32.4% group accuracy, 54.9 F1; Wilcoxon signed-rank, $p < 0.05$ for both metrics). This points to a vision-language misalignment between encoding culturally-specific image content

[3]Image descriptions are the exact prompts used for image generation and retrieval (e.g., "A pair of chopsticks standing upright in a bowl of rice")

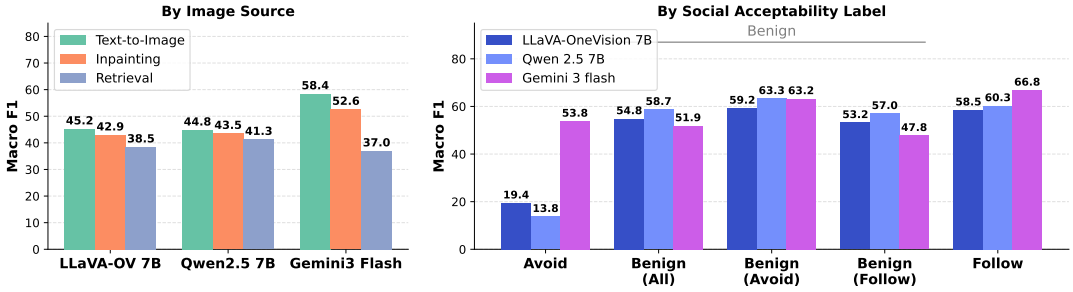


Figure 3: Single-image F1 performance of the top three models: (a) across data sources and (b) across social acceptability labels. Naturally retrieved images are the most challenging, followed by inpainted images, with independently generated images being the easiest. Detecting offensive content is the most difficult task, while identifying the benign counterparts of offensive behavior is easier than detecting offensive content.

and its textual description. Modern VLMs are trained to align visual content with general linguistic descriptions (Yu et al., 2022; Tschannen et al., 2025; Bai et al., 2025), but perhaps *cultural* alignment is sparse in that training signal, causing the two signals to conflict rather than reinforce each other—turning the image into a source of noise rather than information. Yet even under this most favorable condition, absolute performance remains far from perfect. This points to a second, deeper barrier: models lack the cultural knowledge to make normative judgments about the behaviors described, regardless of how those behaviors are conveyed. Visual perception and cultural VL alignment are meaningful bottlenecks, but the ability to reason about the cultural appropriateness remains the harder barrier to bridge.

Performance by Data Source. Figure 3 shows the best closed source model (Gemini 3 Flash) and top 2 open source models (Qwen 2.5 VL 7b, LLaVa-OneVision 7b). We find that retrieval-based images are the hardest (F1 drops 2–21 points), followed by inpainted images, while independently generated images are easiest (Figure 3 (a)). Natural retrieval-based images likely contain more visual variability and noise than synthetic alternatives. For inpainting, the minimal differences between image pairs likely make it harder to detect changes in cultural behaviors, whereas independently generated pairs tend to exhibit more pronounced differences, making it easier to identify them.

Performance by Social Acceptability Labels. As seen in Figure 3 (b), models are considerably better at detecting culturally significant behavior (*Follow*, 58–67 F1) than culturally offensive behavior (*Avoid*, 14–54 F1), with Qwen 2.5 7B exhibiting the most extreme disparity (58.7 vs. 13.8). For *Benign* images, models more reliably identify the benign counterpart of an *Avoid* norm than that of a *Follow* norm (e.g., Qwen2.5 7B: 63.3 vs. 57.0), suggesting that, to models, the *visual contrast of benign situations with offensive behavior is more salient* than the contrast with culturally significant behavior.

5 Improving Cultural Visual Reasoning

Our ablations reveal three sources of failure in VLMs: difficulty perceiving cultural content, vision-language misalignment, and limited cultural knowledge and normative reasoning capabilities. We hypothesize these gaps stem partly from training data limitations: current VLMs rarely encounter culturally grounded examples that connect visual content to cultural significance (Van Miltenburg et al., 2017). However, mining high-quality data at scale for improving cultural competence remains an open and hard challenge (Slingerland et al., 2020). To test whether targeted finetuning on synthetically generated data can help fill these gaps, we construct NORMVIZ-TRAIN and finetune Qwen3-VL models.

5.1 NORMVIZ-TRAIN Construction

We source descriptions of cultural norms from CANDLE (Nguyen et al., 2023), a large-scale broad-coverage cultural knowledge graph extracted from C4 web crawls (Raffel et al., 2020). We use GPT-4o to filter out norms appearing in NORMVIZ-BENCH's evaluation set

to prevent data leakage. Then we apply the same pipeline as §3, limiting to text-to-image generation to prioritize data scale over source diversity. We use Stable Diffusion 3 (SD3) (Esser et al., 2024), a strong open-source T2I model, to avoid closed-source API costs and forgo human validation. Hence, the resulting training set thus consists of cultural norms explicitly disjoint from NORMVIZ-BENCH, and further differs from it in image source, using SD3 rather than Imagen 3.0 (and Gemini 2.0 Flash for inpainting) used in NORMVIZ-BENCH. Although the resulting training dataset may contain inconsistencies, noisy training data remains beneficial at scale (Karpukhin et al., 2019; Nguyen et al., 2024; Zhai et al., 2023; Goyal et al., 2024), and no such large-scale culturally-grounded training resource otherwise exists. To further mitigate quality concerns, we aggressively filter image-text pairs using VQA scoring (as in §3.4). Please refer to Appendix F.1 for additional details on decontamination and data-quality efforts.

5.2 Supervised Finetuning (SFT)

Training data is formatted as VQA instances mirroring the evaluation setup in §4. Each instance consists of an image, the question[4] with answer options (representing *Follow*, *Avoid*, or *Benign*), and a structured explanation generated by GPT-4o. The explanation targets two barriers identified by our ablations (§4), consisting of: (1) a visual grounding component i.e., *what* is depicted and (2) a reasoning component explaining *how* it is culturally appropriate. Models are trained to produce this explanation followed by the social acceptability label, encouraging verbalization of visual content before reasoning over cultural appropriateness.

Experimental Setup. We finetune on individual image–text pairs (not image–image pairs), allowing flexible resampling across social acceptability labels. We experiment with three settings: *balanced* with equal label distribution; *follow-heavy* which oversamples *Follow* norms; and *avoid-heavy* which oversamples *Avoid* norms. To ensure balanced representation across all cultural norms, we select the top 20 images *per norm* ranked by VQA score, yielding 64,288 (64k) training examples. All Qwen3-VL models are finetuned for 3 epochs with a learning rate of 2e-7. We initially experiment with Qwen2.5 VL 7B, and switch to Qwen3 VL upon its release. We see moderate gains on Qwen2.5 VL 7B with NORMVIZ-TRAIN. See Appendix F.2 and F.3 for further details.

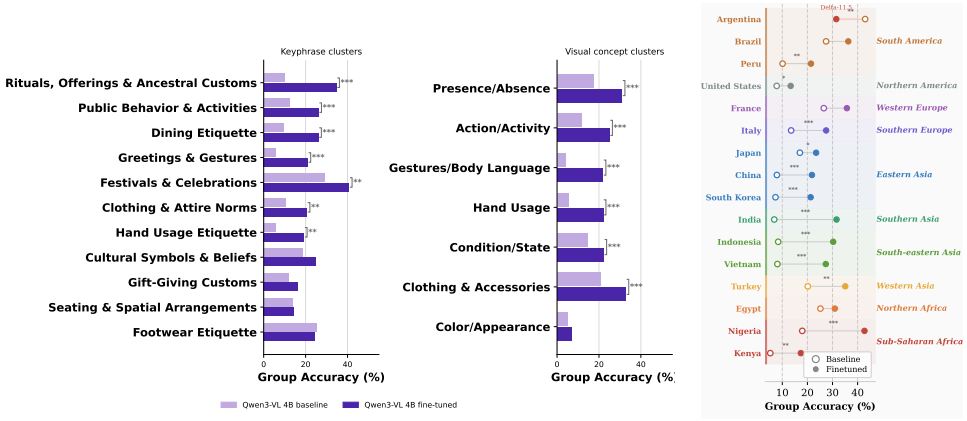
5.3 Results

	Overall		By Source (F1)			By Label (Macro F1)				
Model	**Group Acc**	**Macro F1**	**T2I**	**Inp.**	**Ret.**	**Follow**	**Avoid**	**Benign**	**B_{-Avoid}**	**$B_{-Follow}$**
Qwen3-VL 4B										
Baseline	14.9	37.7	37.3	40.5	28.2	59.7	12.1	41.3	43.6	40.4
Finetuned	**27.7**	**51.0**	**52.5**	**48.7**	**36.4**	**60.5**	**29.8**	**62.6**	**66.4**	**61.3**
Qwen3-VL 8B										
Baseline	19.6	43.6	43.5	45.9	32.6	61.1	19.3	50.5	56.6	48.1
Finetuned	**26.7**	**51.5**	**53.5**	**50.9**	**35.9**	**62.5**	**35.0**	**57.1**	**63.1**	**54.9**

Table 2: Finetuning results on Qwen3-VL 4B and 8B (*follow-heavy 64K*), by image source and label type. Inp.=Inpainting; Ret.=Retrieval; B_{-Avoid}=$Benign_{-Avoid}$; $B_{-Follow}$=$Benign_{-Follow}$

Finetuning on NORMVIZ-TRAIN consistently improves performance across all model sizes and label distributions (Table 2). *Follow-heavy* with 64k training data leads to the largest gains: group accuracy improves by 86% and 36% for Qwen3-VL 4B and 8B, respectively. Gains are consistent across image sources (e.g., for 4B: generated +41%, inpainting +20%, retrieval +29%) and social acceptability labels, with particularly strong improvements in detecting offensive content (+145% for 4B; +82% for 8B). The finetuned 4B model also beats both the finetuned 8B (27.7 vs. 26.7) and Gemini 3.0 Flash (27.7 vs. 25.3), suggesting that task-specific finetuning on culturally grounded data can be particularly beneficial for smaller models. Further details of all finetuning experiments are in Appendix F.3.

[4] "In the context of {country}'s cultural and social norms, classify the social acceptability of the behavior shown in this image."



(a) (b)

Figure 4: Improvements from finetuning Qwen3 VL 4B on NORMVIZ-TRAIN **(a)** Group accuracy gains, broken down by cultural norm domains and visual contrast types (left and right, respectively) in NORMVIZ-BENCH; asterisks (*) indicate significance level. **(b)** Country-wise group accuracy gains organized by UN geoscheme (shown on right); gains are largest in Southern Asia, South-Eastern Asia, and Sub-Saharan Africa.

Improvements across countries and semantic cultural nuances. Finetuning yields consistent gains across most countries (Figure 4b of Qwen3 4B finetuned vs baseline; significance via paired bootstrap test, BH-FDR corrected). The largest group accuracy improvements are seen in culturally underrepresented regions such as Southern Asia, South-eastern Asia, and Sub-Saharan Africa (15-25 points), followed by modest gains in European countries (8-11 points); Argentina is the only country to regress (-11.5 points). Beyond improvements across countries, fine-tuning also yields significant gains along visual and semantic cultural dimensions. Among visual contrast types, gains are strongest and statistically significant (McNemar's test, BH-FDR corrected, $p < 0.001$) for Presence/Absence, Action/Activity, Gestures/Body Language, and Clothing/Accessories (Figure 4a). Among norm domains, the largest gains appear in *Rituals, Offerings & Ancestral Customs*, *Public Behavior & Activities*, and *Dining Etiquette* ($p < 0.001$), with *Greetings & Gestures* also significant ($p < 0.01$) (Figure 4a). Full country-level and semantic category breakdowns are provided in Appendix F.5.

Discussion. We show that finetuning on synthetically generated images and structured explanations of NORMVIZ-TRAIN is a significant first step, though we demonstrate this only for the Qwen model family, where it yields gains across all image sources and labels. However absolute group accuracy remains less than 30%. Further improvement likely requires richer training signals such as spatial supervision (e.g., bounding boxes), contrastive objectives that better align vision and language for cultural content, or new architectures better suited to perceiving and reasoning about cultural behaviors. Retrieval at inference time from varied sources, such as external text or image knowledge bases or richer modalities like video, combined with data transformation methods (e.g., converting between modalities), is a promising path forward. Together, NORMVIZ-BENCH and NORMVIZ-TRAIN pave the way for systematic progress on cultural visual norm reasoning.

6 Conclusion

We introduce NORMVIZ-BENCH, a human-validated benchmark for evaluating visual norm understanding comprising 3,272 image pairs across 16 countries. VLMs perform poorly on this task, struggling most with offensive behaviors and non-Western cultures, due to three compounding barriers: cultural visual perception, vision-language misalignment, and normative reasoning. We introduce NORMVIZ-TRAIN, a 64k synthetic training dataset; finetuning on it leads to significant gains on Qwen3-VL, beating Gemini 3.0 Flash. We highlight visual norm understanding as an open challenge for equitable, globally competent AI.

Acknowledgments

We would like to thank Shaily Bhatt, Karina Halevy, Jocelyn Shen, Vijay Viswanathan, and Fernando Diaz for their valuable feedback on this work. Akhila Yerukola is supported by the 2025-2026 K&L Gates Presidential Fellowship.

Ethics Statement

Our work advances the culturally grounded visual norm understanding in vision-language models, which bears immense potential to build inclusive and context-aware systems to prevent harm and ensure equitable applicability. At the same time, it is important to consider both ethical factors and societal impact:

Existence of Multiple Cultural Proxies Cultural norms are not monolithic; multiple variations often exist within a single country, region, or social group. Defining culture in the context of AI systems is challenging, with prior work categorizing approaches by cultural proxies, linguistic interactions, and measurement strategies (Adilazuarda et al., 2024). Our work uses country boundaries as a demographic proxy for culture. While this approach is a practical starting point, broader evaluation through diverse proxies is needed to capture the full cultural landscape (Bhatt & Diaz, 2024).

Risks in Annotation Recent work has shown that exposure of potentially offensive content can be harmful to the annotators (Roberts, 2016). To mitigate these risks, we obtained informed consent before participation and offered fair compensation at $12/hour. Only essential demographic information was collected, and our annotation study is also supervised by an Institutional Review Board (IRB).

Biases from LLM-generated pipeline Our pipeline relies on LLMs and generative text-to-image models to filter cultural norms, generate contrastive prompt pairs, and synthesize benchmark and training data. Consequently, the dataset may inherit systemic biases from these underlying architectures, such as Western-centric or stereotypical defaults. To mitigate this, we conduct rigorous multi-stage validation by in-country annotators across 16 countries.

Harm Prevention and Intended Use While documenting culturally offensive behaviors carries inherent risks, such as the potential for misuse or the misrepresentation of cultural practices, we are committed to minimizing them. We believe the benefits of improving AI systems' cultural awareness, inclusivity, and safety outweigh the potential harms (Larimore et al., 2021; Ipsos, 2016). The research is intended to contribute to the development of AI systems that are less likely to inadvertently cause cultural offense or misinterpretations. We explicitly do not endorse the use of the data for harmful purposes, including generating offensive content, exploiting cultural differences for malicious intent, or developing biased and discriminatory AI technologies.

References

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, 2024. URL https://arxiv.org/abs/2403.15412.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pp. 114–130, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shaily Bhatt and Fernando Diaz. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16055–16074, 2024.

Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. Nonverbal signals. 2011.

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. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pp. 53–67, 2023.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Darla K Deardorff. *The SAGE handbook of intercultural competence*. Sage Publications, 2009.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633, 2024.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22413–22430, 2025.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Clifford Geertz. *The interpretation of cultures*. Basic books, 2017.

Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22702–22711, 2024.

Edward T Hall. *Beyond culture*. Anchor, 1976.

Edward T Hall. Context and meaning. *Intercultural communication: A reader*, 9:34–43, 2000.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

MORI Ipsos. Attitudes to potentiall offensive language and gestures on tv and radio. 2016.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863*, 2024.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 42–47, 2019.

Eunsu Kim, Junyeong Park, Na Min An, Junseong Kim, Hitesh Laxmichand Patel, Jiho Jin, Julia Kruk, Amit Agarwal, Srikant Panda, Fenal Ashokbhai Ilasariya, et al. World in a frame: Understanding culture mixing as a new challenge for vision-language models. *arXiv preprint arXiv:2511.22787*, 2025a.

Youngmin Kim, Jiwan Chung, Jisoo Kim, Sunghyun Lee, Sangkyu Lee, Junhyeok Kim, Cheoljong Yang, and Youngjae Yu. Speaking beyond language: A large-scale multimodal dataset for learning nonverbal cues from video-grounded dialogues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2265, 2025b.

Mark L. Knapp. Nonverbal communication in human interaction. 1972. URL https://api.semanticscholar.org/CorpusID:142952146.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pp. 81–90, 2021.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024a.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*, 2024b.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216, 2024c.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10467–10485, 2021.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*, 2025.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. Culture in action: Evaluating text-to-image models through social activities. *arXiv preprint arXiv:2511.05681*, 2025.

David Matsumoto and Hyisung C Hwang. The cultural bases of nonverbal communication. 2016.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5769–5790, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.329. URL https://aclanthology.org/2024.emnlp-main.329/.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang W Koh, and Ranjay Krishna. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459, 2024.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pp. 1907–1917, 2023.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. Grounding multilingual multimodal llms with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24198–24242, 2025.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8253–8280, 2022.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3): 907–1004, 2025.

Haoyi Qiu, Kung-Hsiang Huang, Ruichen Zheng, Jiao Sun, and Nanyun Peng. Multimodal cultural safety: Evaluation framework and alignment strategies. *arXiv preprint arXiv:2505.14972*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 428–446, 2023.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A benchmark for measuring the cultural adaptability of large language models. 2024.

Sarah T Roberts. Commercial content moderation: Digital laborers' dirty work. 2016.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

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

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9605–9668, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.500. URL https://aclanthology.org/2025.findings-acl.500/.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. Dosa: A dataset of social artifacts from different indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5323–5337, 2024.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, 2024.

Gwladys Hughes Simon. Some japanese beliefs and home remedies. *The Journal of American Folklore*, 65(257):281–293, 1952.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Edward Slingerland, Quentin D Atkinson, Carol R Ember, Oliver Sheehan, Michael Muthukrishna, Joseph Bulbulia, and Russell D Gray. Coding culture: Challenges and recommendations for comparative cultural databases. *Evolutionary Human Sciences*, 2:e29, 2020.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang (eds.), *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 173–185, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.alvr-1.15. URL https://aclanthology.org/2024.alvr-1.15/.

Emiel Van Miltenburg, Desmond Elliott, and Piek Vossen. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 21–30, 2017.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19565–19575, 2025.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6859–6877, 2025.

Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. Mind the gesture: Evaluating ai sensitivity to culturally offensive non-verbal gestures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25041–25080, 2025.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. Safeworld: Geo-diverse safety alignment. *Advances in Neural Information Processing Systems*, 37: 128734–128768, 2024.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions*, 2022, 2022.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*, 2024.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

A NORMVIZ-BENCH Examples

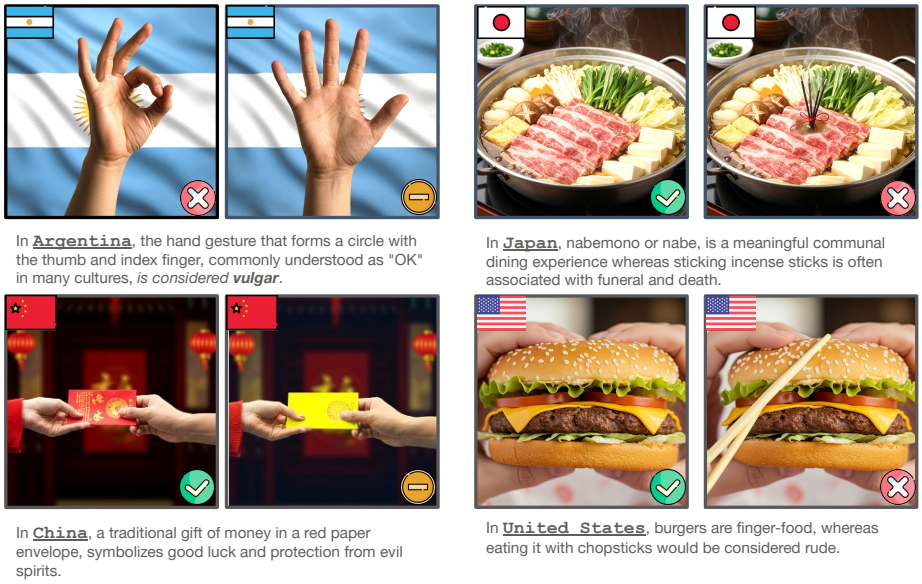
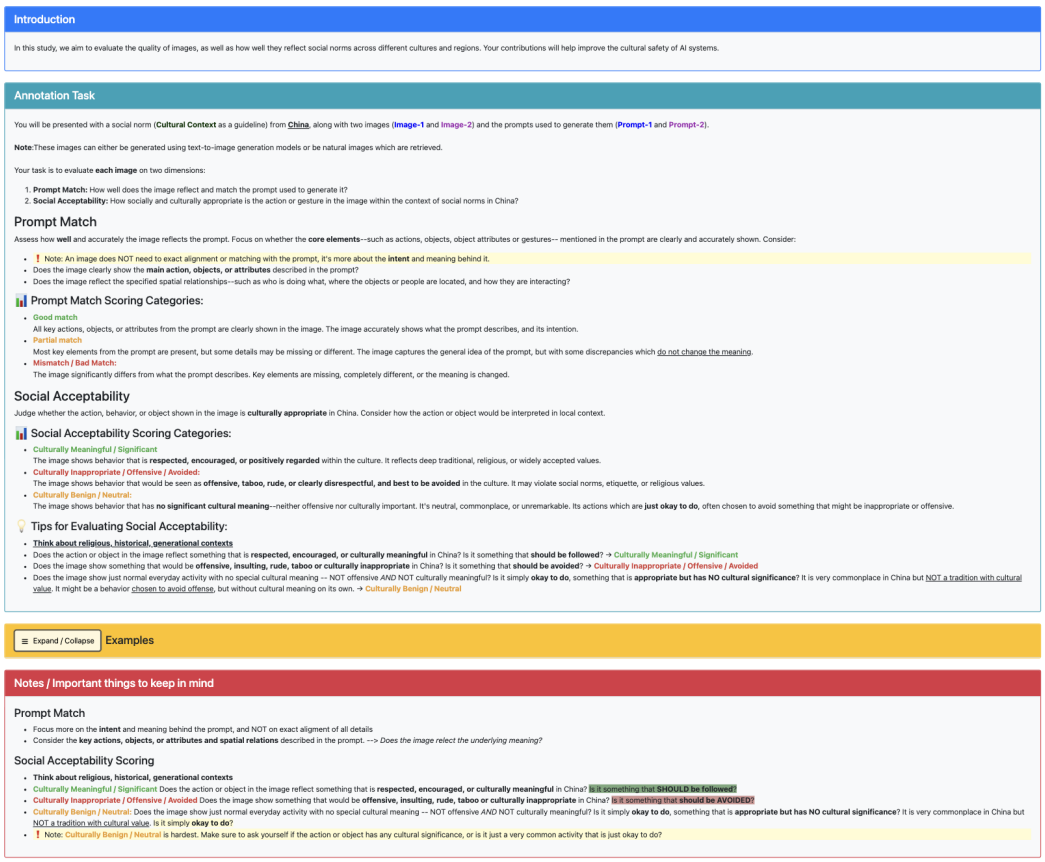


Figure 5: More examples of image pairs from NORMVIZ-BENCH

B Annotation Framework Details

We use Prolific https://www.prolific.com/ to collect annotations. We first pre-screen annotators from the specific country in question (e.g., China), approval rate: 90-100% and 20-10,000 number of previous submissions, and corresponding Ethnicity if applicable (e.g., East Asian). Next, to ensure annotation quality, we required all annotators to pass a qualification task before participating in the main annotation study. The qualification task consisted of 5 high-quality, unambiguous image pairs with known ground-truth labels. Annotators were required to achieve 80% accuracy (4 out of 5 pairs correct) to qualify. A pair was considered correct only if both images in the pair were labeled correctly. Figures 6 and 7 present the annotation instructions and the annotation framework questions. Annotators were compensated at the rate of $12/hr. Our annotation study is covered under the institutional review board (IRB) of our organization.



Introduction

In this study, we aim to evaluate the quality of images, as well as how well they reflect social norms across different cultures and regions. Your contributions will help improve the cultural safety of AI systems.

Annotation Task

You will be presented with a social norm (**Cultural Context** as a guideline) from China, along with two images (**Image-1** and Image-2) and the prompts used to generate them (**Prompt-1** and Prompt-2).

Note: These images can either be generated using text-to-image generation models or be natural images which are retrieved.

Your task is to evaluate **each image** on two dimensions:

1. **Prompt Match:** How well does the image reflect and match the prompt used to generate it?
2. **Social Acceptability:** How socially and culturally appropriate is the action or gesture in the image within the context of social norms in China?

Prompt Match

Assess how **well** and accurately the image reflects the prompt. Focus on whether the **core elements**--such as actions, objects, object attributes or gestures-- mentioned in the prompt are clearly and accurately shown. Consider:

- ❗ Note: An image does NOT need to exact alignment or matching with the prompt, it's more about the **intent** and meaning behind it.
- Does the image clearly show the **main action, objects, or attributes** described in the prompt?
- Does the image reflect the specified spatial relationships--such as who is doing what, where the objects or people are located, and how they are interacting?

📊 Prompt Match Scoring Categories:

- Good match
  All key actions, objects, or attributes from the prompt are clearly shown in the image. The image accurately shows what the prompt describes, and its intention.
- Partial match
  Most key elements from the prompt are present, but some details may be missing or different. The image captures the general idea of the prompt, but with some discrepancies which do not change the meaning.
- Mismatch / Bad Match:
  The image significantly differs from what the prompt describes. Key elements are missing, completely different, or the meaning is changed.

Social Acceptability

Judge whether the action, behavior, or object shown in the image is **culturally appropriate** in China. Consider how the action or object would be interpreted in local context.

📊 Social Acceptability Scoring Categories:

- Culturally Meaningful / Significant
  The image shows behavior that is **respected, encouraged, or positively regarded** within the culture. It reflects deep traditional, religious, or widely accepted values.
- Culturally Inappropriate / Offensive / Avoided:
  The image shows behavior that would be seen as **offensive, taboo, rude, or clearly disrespectful, and best to be avoided** in the culture. It may violate social norms, etiquette, or religious values.
- Culturally Benign / Neutral:
  The image shows behavior that has **no significant cultural meaning**--neither offensive nor culturally important. It's neutral, commonplace, or unremarkable. Its actions which are **just okay to do**, often chosen to avoid something that might be inappropriate or offensive.

💡 Tips for Evaluating Social Acceptability:

- Think about religious, historical, generational contexts
- Does the action or object in the image reflect something that is **respected, encouraged, or culturally meaningful** in China? Is it something that **should be followed**? → Culturally Meaningful / Significant
- Does the image show something that would be **offensive, insulting, rude, taboo or culturally inappropriate** in China? Is it something that **should be avoided**? → **Culturally Inappropriate / Offensive / Avoided**
- Does the image show just normal everyday activity with no special cultural meaning -- NOT offensive *AND* NOT culturally meaningful? Is it simply **okay to do**, something that is **appropriate but has NO cultural significance**? It is very commonplace in China but NOT a tradition with cultural value. It might be a behavior chosen to avoid offense, but without cultural meaning on its own. → **Culturally Benign / Neutral**

≡ Expand / Collapse **Examples**

Notes / Important things to keep in mind

Prompt Match

- Focus more on the **intent** and meaning behind the prompt, and NOT on exact aligment of all details
- Consider the **key actions, objects, or attributes and spatial relations** described in the prompt. --> *Does the image relect the underlying meaning?*

Social Acceptability Scoring

- **Think about religious, historical, generational contexts**
- **Culturally Meaningful / Significant** Does the action or object in the image reflect something that is **respected, encouraged, or culturally meaningful** in China? Is it something that **SHOULD be followed?**
- **Culturally Inappropriate / Offensive / Avoided** Does the image show something that would be **offensive, insulting, rude, taboo or culturally inappropriate** in China? Is it something that **should be AVOIDED?**
- **Culturally Benign / Neutral:** Does the image show just normal everyday activity with no special cultural meaning -- NOT offensive *AND* NOT culturally meaningful? Is it simply **okay to do**, something that is **appropriate but has NO cultural significance**? It is very commonplace in China but NOT a tradition with cultural value. Is it simply **okay to do?**
- ❗ Note: **Culturally Benign / Neutral** is hardest. Make sure to ask yourself if the action or object has any cultural significance, or is it just a very common activity that is just okay to do?

Figure 6: Annotator instructions

Figure 7: Annotation Framework with Example

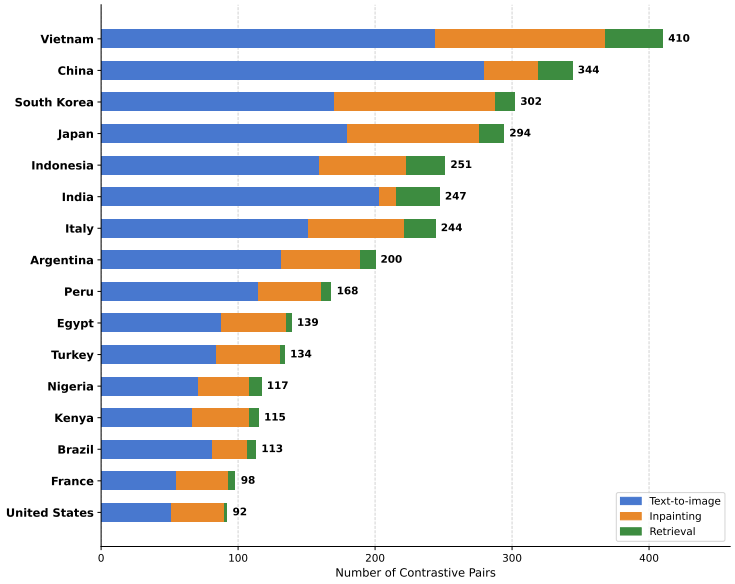


Figure 8: Country-wise distribution of the contrastive pairs in NORMVIZ-BENCH, broken down by image source type.

C NORMVIZ-BENCH Data Characteristics

Our final benchmark contains 3272 contrastive image pairs, spanning across 16 countries. It comprises of image pairs from three sources: text-to-image, inpainting and natural image retrieval. Through human validation, we have three pair types: Follow-Benign, Avoid-Benign and Follow-Avoid. Table 3 shows a summary of the data statistics of NORMVIZ-BENCH.

Metric	Count
Total Image Pairs	3272
Total Images	6544
Unique Norms	1140
Countries	16
By Image Source	
Text-to-Image	65.2%
Inpainting	27.6%
Retrieved Natural	7.2%
By Pair Type	
Follow–Benign	63.7%
Avoid–Benign	21.7%
Follow–Avoid	14.6%
By Label	
Follow	39.1%
Avoid	18.2%
Benign	42.7%

Table 3: Summary statistics of NORMVIZ-BENCH.

To characterize the semantic coverage of NORMVIZ-BENCH, we embedded keyphrases and visual contrast descriptions from the metaadata using OpenAI's `text-embedding-3-small` model and applied K-means clustering to group them into thematic categories. Number of clusters were selected by sweeping over a range of k values and evaluating elbow and silhouette scores, with GPT-4o used to assign interpretable labels to each resulting cluster. This yielded 11 norm domains (e.g., Dining Etiquette, Greetings & Gestures, Festivals & Celebrations) and 7 visual contrast types (e.g., Object State, Presence/Absence, Action/Activity). Overlapping or redundant clusters were subsequently merged by hand. Figure 9 shows the semantic coverage of NORMVIZ-BENCH across different types of cultural domains and contrasting visual differences.

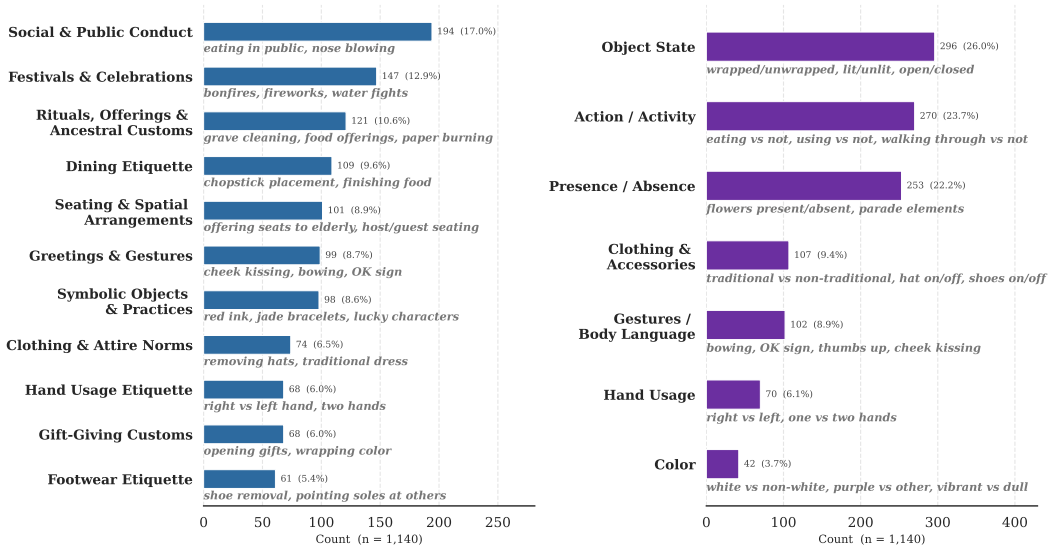


Figure 9: Semantic coverage of NORMVIZ-BENCH across norm domains and visual contrast types. Norm domains span a broad range of cultural behaviors, from everyday etiquette (e.g., Dining Etiquette, Hand Usage Etiquette, Footwear Etiquette) to ceremonial and symbolic practices (e.g., Festivals & Celebrations, Rituals, Offerings & Ancestral Customs, Symbolic Objects & Practices), with Social & Public Conduct being the most represented domain. Visual contrast types are dominated by Object State, Action/Activity, and Presence/Absence, reflecting that cultural norms are most commonly expressed through what objects are used, how actions are performed, or whether culturally significant elements appear at all.

D NORMVIZ-BENCH Construction

D.1 Stage 1: Norm Collection and Structuring

We source cultural descriptions from three benchmarks: CulturalBench, NormAd, and SafeWorld, spanning 16 countries: Argentina, Brazil, China, Egypt, France, India, Indonesia, Italy, Japan, Kenya, Nigeria, Peru, South Korea, Turkey, United States, and Vietnam. We obtain a total of 7,463 cultural norms across the 16 countries (179 from CulturalBench, 5,783 from NormAd, and 1,501 from SafeWorld). If a cultural description involves multiple behaviors (e.g., Don't give knives, or scissors as gifts"), we decompose it into separate descriptions during norm structuring (e.g., Don't give knives as gifts," "Don't give scissors as gifts") to ensure a single item focus. Figure 10 shows the prompt used for norm structuring with GPT-4o.

D.2 Stage 2: Metadata Extraction and Filtering

Metadata Extraction and Visual Detectability Details To identify if a cultural norm is visually detectable, we use GPT-4o to extract structured metadata from each norm-contrast pair. We found metadata extraction from this pair to be significantly better than using just the norms alone, since this task requires reframing the norm contrastively to determine what metadata to extract. The metadata captured includes: the key item of focus, concrete objects, visual contrast between adherence and violation, contextual setting (e.g., festival, funeral), spatial relationships, temporal independence, visual ambiguity, and any abstract concepts involved.

Using these fields, we score each norm on visual detectability on a scale of 0–10, where 0–3 is Low Visibility (LV), 4–7 is Moderately Visual (MV), and 8–10 is Highly Visual (HV). The score is computed across three sub-dimensions: clarity of visual representation of the violation (0–4), ease of distinguishing violation from adherence (0–3), and amount of additional context needed (0–3). In addition to visual detectability, we classify the social acceptability of each norm into one of three categories: norms to *Avoid* (behaviors that trigger discomfort or social rejection) and norms to *Follow* (positive or appropriate behaviors). We filter out the norms that score less than $7/10$ retaining only the highly visual (HV). Figure 11 shows the full prompt used for metadata extraction and visual detectability.

Deduplication with K-means We deduplicate norms across the three benchmarks using `BAAI/bge-m3` (Chen et al., 2024) embedding model, selected for its strong performance on multi-granularity inputs (e.g., 'knife' vs. 'cleaver') . For each norm, we embed the cultural norm description and key-item from the metadata, to capture both broader context and the focal behavior. We cluster these embeddings using K-means with $K = N/2$. We manually look through the norms and estimate a reduction in half; further we do not perfect deduplication. Within each cluster, we use GPT-4o to consolidate semantically equivalent norms into a single representative norm-transformation pair, merging their metadata accordingly.

D.3 Stage 3: Prompt Construction

For each norm classified as either *Follow* or *Avoid*, we construct three kinds of prompts: 1) image descriptions to depict the exact social behvaior. We use these for Text-to-image generation and Retrieval; 2) construct benign counterparts, i.e., transform the norm into a culturally neutral situation 3) generate inpainting instructions to use with Gemini 2.0 Flash. We construct all these prompts using GPT-4o. For each norm, we construct 5 prompt variations per each category. Figure 12 shows the prompt used for the norms to avoid (similar version for norms to follow)

Norm Structuring

```
Task: You will be given a cultural description for {COUNTRY}. It may contain
↪  multiple norms. Your task is to extract these multiple norms (and their
↪  corresponding double negation violation transformations) into individual JSON
↪  dictionaries. Final output will be a list of JSONs.

Guidelines:
A cultural description may cover MULTIPLE norms involving various actions,
↪  objects, event usages (e.g., usages in different festivals) or contrastive
↪  scenario usages (e.g., look for phrases like "rather than", "instead",
↪  "alternatively" etc). If an action or object is appropriate in one setting,
↪  and inappropriate in a different setting, they should be separate norms. In
↪  all these settings, these should be separate JSONs.

Your goal is to:
1. First extract DIRECT distinct norm/s containing sufficient information of the
↪  objects, actions, and background scene information (when and where it occurs).
↪  All norms should be distinct and direct, with no unnecessary details.
For example, consider the cultural description "Taboo items for gifts are itemA,
↪  itemB, itemC etc and these are inappropriate as gifts." Here, each of the
↪  taboo item -- itemA, itemB, itemC -- should be SEPARATE JSON objects with
↪  fully sufficient information, e.g., "itemA are a taboo item and are
↪  inappropriate as gifts." is one norm, "itemB are a taboo item and are
↪  inappropriate as gifts." is the second norm, and so on.
2. Next for each of the norms, construct violation transformations. Make sure the
↪  original intended meaning of the norm is maintained. Consider the following
↪  distilled, extracted Norm: "The bride's family often gifts a tea set to the
↪  couple as a keepsake." The transformation would be: "Not gifting a tea set by
↪  the bride's family is considered not appropriate."
There are a few ways to construct this transformation: negate the core action or
↪  state, reverse or modify the specified action, change the context or setting,
↪  etc. IMPORTANT: Make sure the original intended meaning of the norm is
↪  maintained in the norm transformation.

IMPORTANT: A cultural description about {COUNTRY} which covers multiple aspects
↪  should be separate distinct norm-transformation pairs encoded as separate
↪  JSON objects. Norm - transformation JSON objects should be direct, separate
↪  and disjoint, with no overlapping information. Avoid norm - transformation
↪  pairs which are too generic. DO NOT include information that is not present in
↪  the cultural description.

Output: Provide each norm-transformation in the following structured JSON format:
{"Norm": "<text>", "Transformation": "<text>"}
Output the final output as a list of JSONs: [{"Norm": "<text>", "Transformation":
↪  "<text>"}, ..]
Only generate the final result without any additional descriptions or
↪  explanations.
```

Figure 10: We transform each cultural description into a norm-contrast pair using GPT-4o. Here a 'norm' is a singular behavior whereas 'contrast' is just a contrastive reframing of the same norm, preserving its meaning.

Metadata Extraction and Visual Detectability Classification

```
Task: You will be given a pair of sentences representing a NORM from {COUNTRY}.  Your task is to extract specific
↪  information about the norm, and classify whether it can be visually detected/depicted in an image, along with if the
↪  social offensiveness category of the norm.

Notes: The pair of sentences for a NORM are represented by "Norm" and "Transformation" -- these are rewrites of the same
↪  information. "Transformation" is the double negation interpretation of the "Norm".
Use this Norm pair to extract the information about the norm.

Your goal is to:
1. Extract meta-data about the norm in the following structured format.
a. Key item: This is the key focus point of the norm pair
b. Concrete objects: This is the list of the concrete, easily identifiable objects in the norm pair
c. Visual Contrast: Given a pair of images with one with following the norm, and the other violating the norm, is there a
↪  clear visual difference in identifying the norm?
d. Contextual: Specify the context when this norm is occurring, e.g., festival, funeral (if specified in the norm pair)
e. Spatial Relationships: Given a pair of images with one following the norm, and the other violating the norm, do they
↪  involve any specific, visually apparent arrangements?
f.  Temporal Independence: Can the following or violation of the norm be captured in a single moment, or does it require a
↪  sequence?
g. Visual Ambiguity: Given a pair of images with one following the norm, and the other violating the norm, how likely is it
↪  that the violation image could be misinterpreted or confused with adherence?
h. Abstract Concepts: Does the norm (following or violation) primarily involve intangible ideas or relationships?

2. Classify the Visual Detectability of the norm in an IMAGE
There are three types of classifications for Visual Detectability:
A. Highly Visual (HV): Norms that are easily represented and detected in a single image.
B. Moderately Visual (MV): Norms that can be represented visually but may require some additional context.
C. Low Visibility (LV): Norms that are challenging to represent or detect visually without significant context or
↪  explanation.

Develop a scoring system to quantify visual detectability:
0-3: Low Visibility (LV)
4-7: Moderately Visual (MV)
8-10: Highly Visual (HV)

Score based on:
- Clarity of visual representation of norm (0-4 points) 0: norm is not visually detectable 1: norm is barely visible or
↪  requires extensive context 2: norm is somewhat visible but could be easily missed 3: norm is clearly visible with some
↪  context 4: norm is immediately and obviously visible
- Ease of distinguishing not-adherence from adherence (0-3 points) 0: No visual difference between adherence and
↪  not-adherence 1: Little to no visual difference between adherence and not-adherence 2: Some visible difference, but
↪  could be subtle 3: Clear and obvious visual difference
- Amount of additional context needed (0-3 points) 0: Cannot understand the norm visually, regardless of context 1:
↪  Requires extensive additional context to understand the norm changes 2: Needs some context, but the core norm change is
↪  visually apparent 3: Norm and changes to it are clear with minimal or no additional context needed

Note: Quantity is often tricky and less visually detectable, if there are more than say 5-10 items essential to a norm, its
↪  hard to visually detect that (e.g., 50 students in a classroom).

3. Classify the type Social Norms. There are also three types of Social Norms:
A. Norms to Avoid (Avoid)
B. Norms to Follow (Follow)

Given each pair of norm, classify the type of Norm:
- Is the Norm something to avoid as it triggers discomfort or social rejection (Avoid)? Or is positive, appropriate and the
↪  right thing to do (Follow)?


Overall, for each cultural norm:
a. Identify the key elements of the norm .
b. Evaluate the violation scenario against only the relevant classification criteria.
c. Assign a visual detectability category (HV, MV, or LV) based on the evaluation.
d. Classify the social norm type (Avoid, Follow)


Examples:
<added 4 few-shot examples>

Output Guidelines: Provide the out in the following structured JSON format:
{"meta-data": {"Key-Item": "<text>", "Concrete Objects": "<text>", "Visual Contrast": "<text>", "Contextual": "<text>",
↪  "Spatial Relationships": "<text>", "Temporal Independence": "<text>", "Visual Ambiguity": "<text>", "Abstract
↪  Concepts": "<text>"}, "Visual Detectability": {"Clarity": <num>, "Contrast": <num>, "Additional Context": <num>,
↪  "Final Score Total": <num>, "Final Label": "HV/MV/LV"}, "Social Norm Type": "Avoid/Follow" }
Only generate the final result without any additional descriptions or explanations.
```

Figure 11: Prompt used to extract structured metadata and classify visual detectability and social norm type for each norm-contrast pair using GPT-4o.

Prompt Construction for T2I, Inpainting, and Retrieval

```
Task: You will be given a cultural norm focused on a core action from {COUNTRY}. Your task is to generate five text-to-image
↪  (T2I) generation prompts that clearly illustrate the norm being violated. You also have to generate corresponding
↪  counterfactual BENIGN alternatives of the T2I prompts within the context of {COUNTRY}, where the focus is still on the
↪  key item but the norm is neither being followed nor is it being violated, i.e, such an action has no meaning within
↪  {COUNTRY}.  Final output will be a list of five T2I prompt violating the norm--counterfactual benign pairs.

Workflow:
A. T2I prompt and counterfactual benign prompt pairs:
T2I-prompt -> T2I model -> Image-1
T2I-benign-prompt -> T2I model -> Image-2
B. Masking and In-painting with the T2I prompt:
T2I-prompt -> T2I model -> Image-1
Masking-and-In-painting-instruction + Image-1 -> In-painting model -> Image-2 (T2I-benign-prompt equivalent)

Input:
A. Norm: The specific norm to focus on.
B. Meta-data of Norm: {"Key-Item": "<text>", "Concrete Objects": "<text>", "Visual Contrast": "<text>", "Contextual":
↪  "<text>", "Spatial Relationships": "<text>", "Temporal Independence": "<text>", "Visual Ambiguity": "<text>",
↪  "Abstract Concepts": "<text>"}
C. Key-focus: The core idea of the norm.

Key steps in task:
1. Generate T2I prompts illustrating "Violating" the norm
2. Generate corresponding simple culturally relevant Counterfactual T2I prompts to illustrate "Benign" alternates
3. Instead of the Counterfactual T2I prompt, generate a simple masking and in-paining instruction to convert the image
↪  generated from the T2I prompt to the image that would be generated from the Counterfactual Benign T2I prompt.

Your goal is to:
1. Generate T2I prompts illustrating "Violating" the norms:
- Keep the focus on the key-items: concrete objects or actions in the norm. The goal is to generate T2I prompt instructions
↪  that can be used to generate images to visually represent violating the norm using the key objects or actions.  Such
↪  images will be representative of culturally inappropriate and disrespectful social behavior in {COUNTRY}.
- If the norm includes behavior to avoid, then represent those objects and actions in the T2I prompts.
- Make use of the meta-data to keep the focus on the key-items, objects, and actions. Include any contextual scene
↪  information necessary. The goal is to create realistic depictions of these norms.
- Make sure the T2I prompts are simple, direct, and focused.
IMPORTANT: The T2I prompts illustrating the "Violating" should exclusively depict the key-item and its directly related
↪  components as described in the norm. Do not introduce any additional objects, actions, or customs not explicitly
↪  mentioned in the norm description, even if they are generally associated with the broader cultural context.
IMPORTANT: DO NOT include any mention of children, kids, babies or any faces of humans in the T2I prompts.

2. Generate corresponding Counterfactual T2I prompts to illustrate culturally "Benign" alternates:
- Benign alternatives refer to manipulating the object or action attributes, such as color, item, position, etc, such that
↪  the image is culturally neutral or irrelevant with regards to the key-item. Make sure these benign alternatives are NOT
↪  culturally meaningful and NOT culturally inappropriate/discouraged. The alternative action or object should be neutral
↪  and insignificant.
- IMPORTANT: The counterfactual prompts should still be situated in the setting of the {COUNTRY} culture.
- IMPORTANT: Focus on the visually detectable alternatives, such as color, item, position, etc. CAN BE REPEATED with slight
↪  variations (e.g., yellow flowers vs white flowers vs blue flowers) Changing the contents of an item (e.g., money in bag
↪  or envelope) cannot be detected. Often slightly increasing or decreasing quantity (like number of items) cannot be
↪  detected. Easily identifiable and detectable alternatives need to be prioritized.

3. Generate masking and in-painting instructions:
- Generate instructions that can be used to convert the image generated from the T2I prompt to the image that would be
↪  generated from the Counterfactual T2I prompt.
- Make sure the instructions are simple, clear, direct, and focused on the key-items, objects, and actions.
- Sample instructions: "change the color of red item to blue", "move the item from left to right", "replace the item1 from
↪  the image with item2", "change the position of the item from upright to flat" etc. Try to focus on only one change in
↪  the instruction.
- Make sure the instructions correspond to the counterfactual prompts generated in step 2.
- Make sure the instructions have the same information as the key focus objects or actions, e.g., do not change "gifting
↪  with one hand" to "holding with one hand". Maintaining the key details as in step1 will make the in-painted
↪  instructions specific and easy to execute.
- MAKE SURE THE INSTRUCTIONS ARE SIMPLE AND EASY

Guidelines:
Content:
- T2I prompts illustrating "Violating" the norms should be simple, direct, and focused on the visual aspects of the norm
↪  and key items, objects, or actions.
- T2I prompts illustrating "Benign" alternatives should be neutral and culturally IRRELEVANT. They should NOT be culturally
↪  meaningful OR inappropriate. EASY TO IDENTIFY objects should NOT BE meaninful variations (e.g., do not replace red
↪  envelopes with teacups for China, do not replace right hand with both hands for India, etc). Always check if the benign
↪  alternative is culturally neutral and irrelevant.
- In-painting instructions should be clear, direct, and easy to execute. Focus on only 1 change in the instruction.
- IMPORTANT: The T2I prompts illustrating the "Violating" should not introduce any additional objects, actions, or customs
↪  not explicitly mentioned in the norm description, even if they are generally associated with the broader cultural
↪  context.

Output:
- Provide each violating prompt - benign prompt - benign inpainting instruction triplet in the following structured JSON
↪  format:
{"T2I-Prompt": "<text>", "T2I-Benign-Prompt": "<text>", "T2I-Benign-Inpainting-Instruction": "<text>"}
- Output the final output as a list of five JSONs: [{"T2I-Prompt": "<text>", "T2I-Benign-Prompt": "<text>",
↪  "T2I-Benign-Inpainting-Instruction": "<text>"},...]
- Ensure the prompts do not mention "based on description" or "based on norm or meta-data".
- Only generate the final result without any additional descriptions or explanations.
```

Figure 12: GPT-4o Prompt used to construct prompts for norms to *avoid*. A similar variant is used for norms to *follow*.

D.4 Stage 4: Multi-Source Image Generation

Text-to-Image Generation We use Imagen 3.0 (`imagen-3.0-generate-002`) for T2I image generation. We generate 2 images per prompt variation (Avoid/Follow and the corresponding Benign)

Inpainting We use Gemini 2.0 Flash (`gemini-2.0-flash-exp-image-generation`) for applying inpainting instruction to generated Avoid/Follow images. This is used to generate inpainted Benign counterparts.

Real Image Retrieval We use DataComp to retrieve images using the Avoid/Follow and Benign prompts used for T2I generation. We query country-specific subsets of DataComp. These subsets are created by parsing image URLs and categorizing them based on the country-code top-level domain they contain (e.g., ".in" are assigned to the India subset). We embed all prompts and images using `CLIP ViT-bigG-14` (Radford et al., 2021) and retrieve up to 20 candidates per prompt based on cosine similarity. We pair the retrieved *Follow/Avoid* and *Benign* images, removing any near-duplicates.

Quality Filtering We apply two stages of quality filtering on the retrieved and generated images. First, each image and the prompt used for generation/retrieval are embedded using `clip-flant5-xxl` and then we compute VQAScore (Lin et al., 2024). We retain only images that exceed a similarity threshold of 0.7 (out of 1), which is determined after qualitatively evaluating the images and associated scores. Second, we apply a GPT-4o-based filter that evaluates alignment between each image and the same prompt used for generation/retrieval. GPT-4o is prompted to consider the objects, spatial relations, and actions mentioned, and classifies the image-prompt pair as a Good Match, Partial Match, or Bad Match; we retain only Good Match images. The retained images are sent for full human validation. Figure 13 contains the prompt used for quality filtering with GPT-4o.

GPT-4o Quality Filter

```
How well does the image reflect the prompt "{caption}".

Note that it doesnt have to be exact alignment with the prompt, its more about the
↪  intent behind it.

Guidelines:
1. Pay attention to the objects, object attributes and spatial relations of the
↪  objects mentioned in the prompt
2. Pay attention to the actions specified in the prompt, along with the actors.

Score whether it is:
1. Good Match
2. Partial Match
3. Bad Match
Output option number and option name.
```

Figure 13: Prompt used for quality filtering with GPT-4o

	Overall		By Source (macro F1)			By Label (macro F1)				
								Benign		
Model	Group Acc.	Macro F1	Generated	Inpainting	Retrieval	Follow	Avoid	All	Avoid	Follow
PaliGemma 2 3B	0.0	18.8	18.6	18.5	20.8	56.3	0.0	0.0	0.0	0.0
LLaVA 1.5 7B	1.5	22.3	23.3	20.8	17.9	34.7	32.4	0.0	0.0	0.0
PaliGemma 3B	2.0	19.6	19.3	18.5	21.9	0.0	31.9	27.8	25.5	28.4
BLIP-2 2.7B	3.8	26.3	25.7	25.5	33.1	7.1	13.9	57.9	**63.5**	56.3
Molmo 2 8B	5.0	31.5	32.0	32.1	24.3	56.4	31.9	6.1	10.3	4.6
LLaVA-NeXT 1.6 7B	10.8	38.5	40.8	37.1	22.8	42.6	**33.9**	39.1	36.2	39.9
Gemma 3 4B	11.7	34.6	34.7	35.1	31.0	58.7	10.5	34.7	39.6	32.9
InternVL 8B	12.9	37.7	37.9	39.3	27.1	59.4	16.4	37.4	43.9	35.0
InternVL 2B	13.4	36.2	36.3	36.8	31.9	58.9	10.2	39.4	44.0	37.7
DeepSeek-VL2 1B	13.8	40.6	41.6	39.6	34.2	53.8	27.7	40.2	40.3	40.1
Qwen 3 4B	14.9	37.7	37.3	40.5	28.2	59.7	12.1	41.3	43.6	40.4
Aya Vision 8B	15.4	37.3	37.9	37.5	28.1	59.5	7.3	45.1	52.0	42.5
Phi-3V 4.2B	16.1	40.2	40.8	39.1	38.9	57.1	14.7	48.8	53.2	47.3
GLM-4V 9B	16.7	42.5	43.4	41.8	35.3	59.2	23.8	44.6	47.3	43.7
Qwen 2.5 3B	18.3	**44.6**	**47.3**	41.9	30.9	53.7	27.7	52.2	57.3	50.4
Qwen 3 8B	19.6	43.6	43.5	**45.9**	32.6	**61.1**	19.3	50.5	56.6	48.1
LLaVA-OneVision 7B	19.8	44.3	45.2	42.9	38.5	58.5	19.4	54.8	59.2	53.2
Qwen 2.5 7B	**23.0**	44.2	44.8	43.5	**41.3**	60.3	13.8	**58.7**	63.3	**57.0**
GPT-4o	16.3	41.5	42.4	41.2	30.7	61.8	20.5	42.3	49.8	39.2
GPT-5.2	22.7	45.6	47.0	42.6	**40.5**	58.8	17.9	**60.0**	**63.9**	**58.6**
Gemini 3 flash	**25.3**	**55.8**	**58.4**	**52.6**	37.0	**66.8**	**53.8**	51.9	63.2	47.8

Table 4: Performance of 21 VLMs on NORMVIZ-BENCH. **Bold** indicates best within category (open-source or closed-source), underline indicates second-best within category. Models within 0.1 points are highlighted together. Green background = open-source, orange = closed-source.

E Benchmarking Results

E.1 Main Results

Table 4 shows the performance of all 21 VLMs on NORMVIZ-BENCH, broken down by image source and by social acceptability label. Table 5 and Table 6 show the performance of all 21 VLM models, with the two oracle ablation experiments.

E.2 Effect of Model Size.

Model size shows weak positive Pearson correlations with performance (F1: $r = 0.29$; group accuracy: $r = 0.26$), but scaling effects vary substantially across families. Qwen 3 exhibits the most robust scaling, with both F1 and group accuracy improving consistently from 4B to 8B (+5.9 and +4.7 percentage points, respectively). Other families show inconsistent patterns: Qwen 2.5 7B improves in group accuracy (23.0% vs. 18.3%) but not F1 (44.2% vs. 44.6%) over Qwen 2.5 3B; and InternVL 8B shows minimal or negative change relative to the InternVL 2B variant on both metrics (37.7% vs. 36.2% F1; 12.9% vs. 13.4% group accuracy).

E.3 Per Country Performance

Countries in the Global South and non-Western regions prove consistently harder for most models. Kenya and India show the highest fractions of underperforming models (17 out of 21), followed closely by Indonesia (16 out of 21) and a four-way tie among China, Peru, the United States, and Vietnam (15 out of 21), suggesting that VLMs struggle disproportionately with visual norms from these regions relative to their own overall performance. In contrast, Brazil, Argentina, Egypt, and France see far fewer models underperforming (3–6 out of 21), indicating more consistent recognition of visual norms from these countries. This asymmetry points to a systematic gap in how well current VLMs generalize across cultures, likely reflecting imbalances in the cultural composition of their training data. Figure 14 shows the performance of 21 VLMs across 16 countries.

	Condition	Group Acc.	Macro F1	Generated	Inpainting	Retrieval	Follow	Avoid	Benign$_{\text{-All}}$	Benign$_{\text{-Avoid}}$	Benign$_{\text{-Follow}}$
Closed source											
	Gemini 3 flash	**25.3**	**55.8**	**58.4**	**52.6**	37.0	**66.8**	**53.8**	51.9	63.2	47.8
oracle	+ Image Descriptions	32.7	63.6	66.2	60.2	51.1	72.1	68.6	57.1	69.8	52.4
oracle	+ Image Desc. (w/o img)	40.8	68.0	70.4	66.0	51.3	74.0	70.4	61.0	73.3	56.5
	GPT 5.2	22.7	45.6	47.0	42.6	**40.5**	58.8	17.9	**60.0**	**63.9**	**58.6**
oracle	+ Image Descriptions	37.8	60.8	62.9	57.0	56.2	70.0	46.8	65.6	69.2	64.3
oracle	+ Image Desc. (w/o img)	45.3	66.5	69.1	61.2	62.9	73.8	56.5	69.3	72.5	68.1
	GPT-4o	16.3	41.5	42.4	41.2	30.7	61.8	20.5	42.3	49.8	39.2
oracle	+ Image Descriptions	23.5	50.5	51.9	48.8	43.4	65.6	36.4	49.4	58.1	45.8
oracle	+ Image Desc. (w/o img)	37.2	62.8	65.3	58.8	52.3	71.3	56.3	60.7	69.1	57.5
Open source											
	Qwen 2.5 7B	**23.0**	44.2	44.8	43.5	**41.3**	60.3	13.8	**58.7**	**63.3**	**57.0**
oracle	+ Image Descriptions	30.8	48.9	49.2	48.6	46.8	67.3	17.9	61.5	63.1	60.9
oracle	+ Image Desc. (w/o img)	32.7	53.2	55.1	49.3	50.9	68.5	29.9	61.1	62.1	60.7
	LLaVA-OneVision 7B	19.8	44.3	45.2	42.9	38.5	58.5	19.4	54.8	59.2	53.2
oracle	+ Image Descriptions	27.3	48.5	49.4	47.0	44.0	65.4	21.1	59.0	62.8	57.4
oracle	+ Image Desc. (w/o img)	34.3	53.6	55.7	50.1	46.5	68.9	27.9	64.1	65.5	63.5
	Qwen 3 8B	19.6	43.6	43.5	**45.9**	32.6	**61.1**	19.3	50.5	56.6	48.1
oracle	+ Image Descriptions	28.1	50.0	50.7	49.7	43.2	67.2	27.1	55.7	59.9	54.0
oracle	+ Image Desc. (w/o img)	40.2	60.2	62.4	56.1	54.3	72.2	41.0	67.4	68.3	67.1
	Qwen 2.5 3B	18.3	**44.6**	**47.3**	41.9	30.9	53.7	27.7	52.2	57.3	50.4
oracle	+ Image Descriptions	27.5	52.2	54.0	48.9	46.6	63.3	34.7	58.6	61.6	57.6
oracle	+ Image Desc. (w/o img)	29.7	56.0	57.2	54.3	48.7	67.2	43.7	57.0	53.1	58.2
	GLM-4V 9B	16.7	42.5	43.4	41.8	35.3	59.2	23.8	44.6	47.3	43.7
oracle	+ Image Descriptions	23.1	47.4	48.7	45.2	43.3	63.5	33.4	45.3	43.7	45.9
oracle	+ Image Desc. (w/o img)	31.0	53.9	55.8	50.5	48.4	66.0	40.2	55.7	56.3	55.5
	Phi-3V 4.2B	16.1	40.2	40.8	39.1	38.9	57.1	14.7	48.8	53.2	47.3
oracle	+ Image Descriptions	29.9	49.9	50.6	48.7	48.1	65.9	23.9	59.8	63.6	58.4
oracle	+ Image Desc. (w/o img)	35.8	56.4	58.0	53.1	52.5	67.3	36.6	65.3	68.5	64.2
	Aya Vision 8B	15.4	37.3	37.9	37.5	28.1	59.5	7.3	45.1	52.0	42.5
oracle	+ Image Descriptions	26.7	44.3	44.9	43.6	39.8	64.7	12.1	56.0	57.4	55.5
oracle	+ Image Desc. (w/o img)	28.7	48.5	49.8	46.4	42.8	65.2	23.3	56.9	58.4	56.3
	Qwen 3 4B	14.9	37.7	37.3	40.5	28.2	59.7	12.1	41.3	43.6	40.4
oracle	+ Image Descriptions	23.0	43.7	43.5	45.0	39.8	64.4	15.3	51.2	55.7	49.4
oracle	+ Image Desc. (w/o img)	30.7	52.9	54.7	49.9	47.6	67.5	33.7	57.6	58.8	57.1
	DS-VL2 1B	13.8	40.6	41.6	39.6	34.2	53.8	**27.7**	40.2	40.3	40.1
oracle	+ Image Descriptions	19.7	45.4	46.2	44.7	38.2	60.7	31.1	44.3	40.2	45.6
oracle	+ Image Desc. (w/o img)	2.7	26.8	27.5	25.2	26.5	0.2	20.5	59.8	68.1	57.5

Table 5: Full results for models evaluated in the main paper (Table 1), including per-source and per-label Macro F1 breakdowns. Models are ordered by baseline Group Accuracy (descending) within each source group. **Bold** = best within source group; underline = second-best. Shaded rows show oracle ablations: (1) + *Image Descriptions* – image with text description; (2) + *Image Desc. (w/o img)* – text description only, no image.

	Condition	Group Acc.	Macro F1	Generated	Inpainting	Retrieval	Follow	Avoid	Benign$_{\text{-All}}$	Benign$_{\text{-Avoid}}$	Benign$_{\text{-Follow}}$
Open source (remaining models, ordered by baseline Group Accuracy)											
	InternVL 2B	**13.4**	36.2	36.3	36.8	31.9	58.9	10.2	39.4	44.0	37.7
oracle	+ Image Descriptions	19.8	39.4	40.0	38.5	37.6	62.2	8.7	47.3	48.1	47.0
oracle	+ Image Desc. (w/o img)	33.7	50.0	51.2	46.9	49.0	66.5	16.9	66.5	66.6	66.5
	InternVL 8B	12.9	37.7	37.9	**39.3**	27.1	**59.4**	16.4	37.4	43.9	35.0
oracle	+ Image Descriptions	22.3	46.3	47.8	44.5	38.1	64.3	24.7	49.8	57.1	46.7
oracle	+ Image Desc. (w/o img)	35.3	57.6	60.0	53.5	51.5	70.0	41.1	61.8	64.4	60.8
	Gemma 3 4B	11.7	34.6	34.7	35.1	31.0	58.7	10.5	34.7	39.6	32.9
oracle	+ Image Descriptions	14.0	36.6	37.3	35.5	35.5	60.1	11.8	37.9	43.1	35.9
oracle	+ Image Desc. (w/o img)	25.8	48.1	49.2	46.5	42.7	64.4	25.1	54.7	59.8	52.7
	LLaVA-NeXT 1.6 7B	10.8	**38.5**	**40.8**	37.1	22.8	42.6	**33.9**	39.1	36.2	39.9
oracle	+ Image Descriptions	20.5	50.5	52.3	48.5	39.2	65.8	42.4	43.2	42.1	43.6
oracle	+ Image Desc. (w/o img)	26.6	55.4	56.4	54.9	45.7	66.1	44.8	55.1	57.8	54.2
	Molmo 2 8B	5.0	31.5	32.0	32.1	24.3	56.4	31.9	6.1	10.3	4.6
oracle	+ Image Descriptions	22.4	47.2	48.8	45.8	37.3	65.8	32.8	43.2	45.2	42.5
oracle	+ Image Desc. (w/o img)	30.1	52.8	54.7	49.3	48.5	67.5	35.6	55.4	61.2	53.2
	BLIP-2 2.7B	3.8	26.3	25.7	25.5	**33.1**	7.1	13.9	**57.9**	**63.5**	**56.3**
oracle	+ Image Descriptions	4.2	26.4	25.7	25.4	36.4	11.9	9.3	58.0	62.7	56.6
oracle	+ Image Desc. (w/o img)	1.9	25.7	26.0	25.3	24.0	2.2	20.8	54.1	58.5	52.8
	PaliGemma 3B	2.0	19.6	19.3	18.5	21.9	0.0	31.9	27.8	25.5	28.4
oracle	+ Image Descriptions	1.4	20.4	19.8	20.1	21.1	0.0	32.3	29.3	25.1	30.5
oracle	+ Image Desc. (w/o img)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LLaVA 1.5 7B	1.5	22.3	23.3	20.8	17.9	34.7	32.4	0.0	0.0	0.0
oracle	+ Image Descriptions	5.3	36.5	37.3	35.3	31.1	65.0	39.7	4.7	3.0	5.3
oracle	+ Image Desc. (w/o img)	5.0	36.3	36.9	35.6	31.1	66.3	40.2	2.3	1.9	2.5
	PaliGemma 2 3B	0.0	18.8	18.6	18.5	20.8	56.3	0.0	0.0	0.0	0.0
oracle	+ Image Descriptions	0.0	18.6	18.5	18.4	20.0	56.2	0.0	0.0	0.0	0.0
oracle	+ Image Desc. (w/o img)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6: Full results for the remaining 9 open-source models not included in Table 1, ordered by baseline Group Accuracy (descending). **Bold** = best within this group. Shaded rows show oracle ablations: (1) + *Image Descriptions* – image with text description; (2) + *Image Desc. (w/o img)* – text description only, no image. Note the large gap between baseline and ablation for Molmo 2 8B (5.0 $\rightarrow$ 30.1) group accuracy with image descriptions, indicating especially poor vision-language alignment.

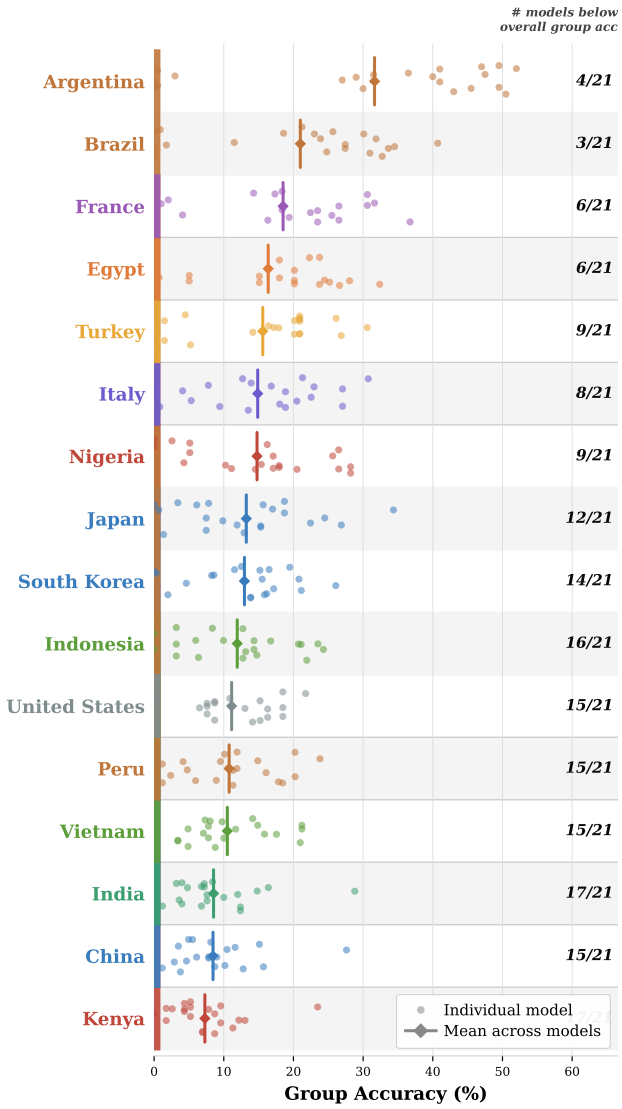


Figure 14: Group accuracy of 21 vision-language models (VLMs) across 16 countries. Each dot represents an individual model's group accuracy (%) for a given country, and the diamond marker shows the mean across all models. Countries are ordered from lowest (bottom) to highest (top) mean group accuracy. The fraction on the right (e.g., 17/21) indicates how many models score below their own overall group accuracy. For example, a model with an overall mean group accuracy of 15% that scores 8% on India would count toward this fraction. Higher fraction indicates that the country is consistently harder for most models relative to their average performance across all countries. Countries are color-coded by geographic region.

F Improving Cultural Visual Reasoning

F.1 NORMVIZ-TRAIN Construction Additional Details

We construct NORMVIZ-TRAIN from CANDLE (Nguyen et al., 2023). Norm decontamination between training and benchmark data is crucial, so we perform explicit norm-level semantic deduplication between CANDLE and NORMVIZ-BENCH's evaluation set before constructing NORMVIZ-TRAIN. For each of the 16 countries, we provide GPT-4o with each CANDLE candidate norm and the full set of test norms from that country (23–93 norms per country; mean 57.5, std. 22.4), and prompt it to identify semantic matches regardless of surface phrasing, discarding any candidate flagged as a match. This involves 62,323 GPT-4o calls across all countries. Please find the prompt used in Figure 15. To validate the efficacy of GPT-4o filtering, one author manually validated 50 randomly sampled judgments: 48 were in perfect agreement with the model, and the remaining 2 were false positives (over-flagged as matches when they were not), with no false negatives identified—indicating the filter is conservative in the direction of removing more candidate overlap rather than less.

On the resulting decontaminated set, we apply the same pipeline as §3: structuring norms into *norm–contrast* pairs, extracting metadata, and filtering for visual detectability. To prioritize data scale over source diversity, we rely solely on text-to-image generation. After filtering CANDLE across 16 countries for visual detectability, we retain 15,331 cultural descriptors of highly visual social behaviors, each tagged as Follow or Avoid. For each norm, we generate 10 T2I prompts for both the Follow/Avoid and its corresponding Benign counterpart (Benign_Follow or Benign_Avoid), then generate 10 images per prompt using `Stable Diffusion 3 (SD3)` (Esser et al., 2024), yielding 100 images per norm variant and 100 for the corresponding benign version. Next, we score each norm image against both its own prompt and its benign counterpart using VQA (as in §3.4), and vice versa for the benign images, yielding four cross-evaluated scores per image pair. We then compute a discriminability score per pair: $= (\textit{norm\_image\_on\_norm\_prompt} + \textit{benign\_image\_on\_benign\_prompt}) - (\textit{benign\_image\_on\_norm\_prompt} + \textit{norm\_image\_on\_benign\_prompt})$. This rewards pairs where each image aligns strongly with its intended prompt while remaining semantically distant from the other. We retain only pairs exceeding a discriminability threshold of 0.8, and select the top $n \in \{10, 20\}$ pairs per norm, treating n as a control variable in our experiments. Through this rigorous filtering, we ensure data quality of NORMVIZ-TRAIN. We also run a targeted training experiment using only the NORMVIZ-BENCH norms, generating a different training set using it, and measure performance difference only to find only marginal improvements (later in §F.4).

NORMVIZ-TRAIN-NORMVIZ-BENCH norm decontamination

```
System Prompt: "You are a helpful assistant that analyzes whether social norms are
↪ semantically similar. You will be given a candidate norm and a list of
↪ reference norms. Your task is to identify if the candidate norm semantically
↪ matches any of the reference norms. Return a JSON with three fields: 1.
↪ "is_match": true if there's a semantic match, false otherwise 2.
↪ "matching_norms": list of indices of matching norms (empty if no match) 3.
↪ "explanation": brief explanation of your reasoning"

User Prompt: "Candidate norm: {train_norm}. Reference norms:
↪ {formatted_test_norms}. Determine if the candidate norm semantically matches
↪ any of the reference norms. Consider that norms can be phrased differently but
↪ have the same meaning. Focus on the core social rule rather than superficial
↪ wording differences. Return your analysis in JSON format as specified."
```

Figure 15: Prompt used for excluding NORMVIZ-BENCH norms from NORMVIZ-TRAIN

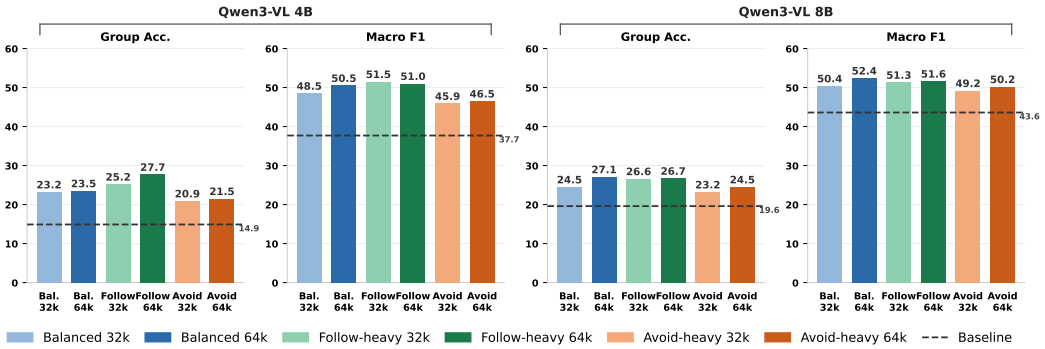


Figure 16: Finetuning results for Qwen3 VL 4B and 8B models across label distribution configurations (balanced, follow-heavy, avoid-heavy) and data scales (32k and 64k)

F.2 Experimental Setup

We experiment with three label distribution configurations: *balanced* (25% each of *Follow*, *Avoid*, $Benign_{Follow}$, $Benign_{Avoid}$), *follow-heavy* (35% *Follow*, 35% $Benign_{Follow}$, 15% *Avoid*, 15% $Benign_{Avoid}$), and *avoid-heavy* (the inverse). We additionally vary data scale by selecting top $n = 10, 20$ images per cultural norm ranked by VQA score (out of the generated 100 images), yielding 32,144 and 64,288 training examples respectively (32k and 64k). We finetune Qwen3 VL 4B and 8B models on both 32k and 64k. We also experiment with Qwen2.5 VL 7B model on 32k, and see minimal gains.

While we agree that these experiments are limited to the Qwen model family, there are meaningful differences between Qwen 2.5 VL and Qwen 3 VL models. Qwen2.5-VL model framework uses a Vision Transformer (ViT) vision encoder trained from scratch on DataComp with a Qwen 2.5 LLM as the language decoder, whereas Qwen 3 VL switches to SigLIP-2 vision encoder pretrained on WebLI data and Qwen 3 LLM as the decoder. These differ not only in architecture but also in vision encoder training data, and we observe consistent gains across all variants (Appendix F.3), suggesting the training signal in NormViz-Train is not specific to a single architecture.

F.3 Full Results discussing effects of label distributions and data scales.

Finetuning Qwen3 VL 4B and 8B models on balanced, follow-heavy and avoid-heavy configurations and data scales of NormViz-Train consistently helps across the board. *Balanced* and *follow-heavy* training perform competitively across all label types, while *avoid-heavy* configuration yields relatively lower gains likely due to the test set being skewed toward the *Follow* label. Scaling from 32*k* to 64*k* yields almost consistent but modest gains across all configurations and both models, (group accuracy: +0.1-2.6, F1: +0.3-2.0, except *follow-heavy* 64*k* which drops by -0.5). Absolute group accuracy still remains low across the board, indicating that meaningful gains likely require training objectives that go beyond label prediction, perhaps with additional supervision such as bounding boxes to better localize cultural nuances, or alternative contrastive training objectives such as CLIP (Radford et al., 2021). See Figure 16 for complete results. We also initially experiment with Qwen 2.5VL 7B (see Figure 17) and see moderate gains.

F.4 Probing for training-test leakage

We additionally ran a preliminary targeted experiment to specifically probe whether leakage is a meaningful concern: we constructed a separate training subset using only the test set norms directly (i.e., explicitly training on the same norms as NormViz-Bench), generated images using the same SD3 pipeline, and finetuned both Qwen3-VL 4B and 8B on this set. We observe only marginal improvements on balanced label distribution and 32k datasize: Qwen3-VL 4B improves from 23.2 group accuracy (F1: 48.5) to 23.9 (F1: 52.0), and Qwen3-VL

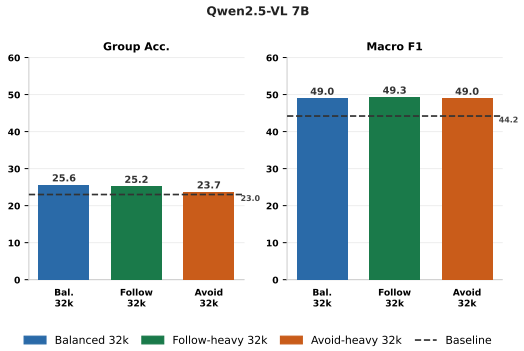


Figure 17: Finetuning results for Qwen 2.5 VL 7B model across label distribution configurations (balanced, follow-heavy, avoid-heavy) and 32k data size

8B from 24.5 (F1: 50.4) to 26.7 (F1: 53.4), suggesting that the lack of norm-level information is not the bottleneck for the current finetuning setup, but rather that meaningful gains likely require richer training signals, better vision-language alignment for cultural content, or architectures better suited to perceiving and reasoning about cultural behaviors.

F.5 Improvement Across Countries and Semantic Cultural Nuances

We see significant improvement from culturally underrepresented non-Western countries, especially for the smaller finetuned Qwen3 VL 4B model. Qwen3 VL 4B model has very poor performance on NORMVIZ-BENCH, and potentially less biases. Hence finetuning on NORMVIZ-TRAIN significantly helps improve visual norm understanding, even surpassing finetuned Qwen3 VL 8B and Gemini 3.0 Flash. Figure 18 and Figure 19 show performance across countries, using group accuracy and single image F1 score. Finetuning also consistently improves cultural knowledge across several social norm domains, as well compositionality across different visual contrast types. Figures 20, 21, 22, 23 show improvements for Qwen3 VL 4B and Qwen3 VL 8B models (follow-heavy 64k).

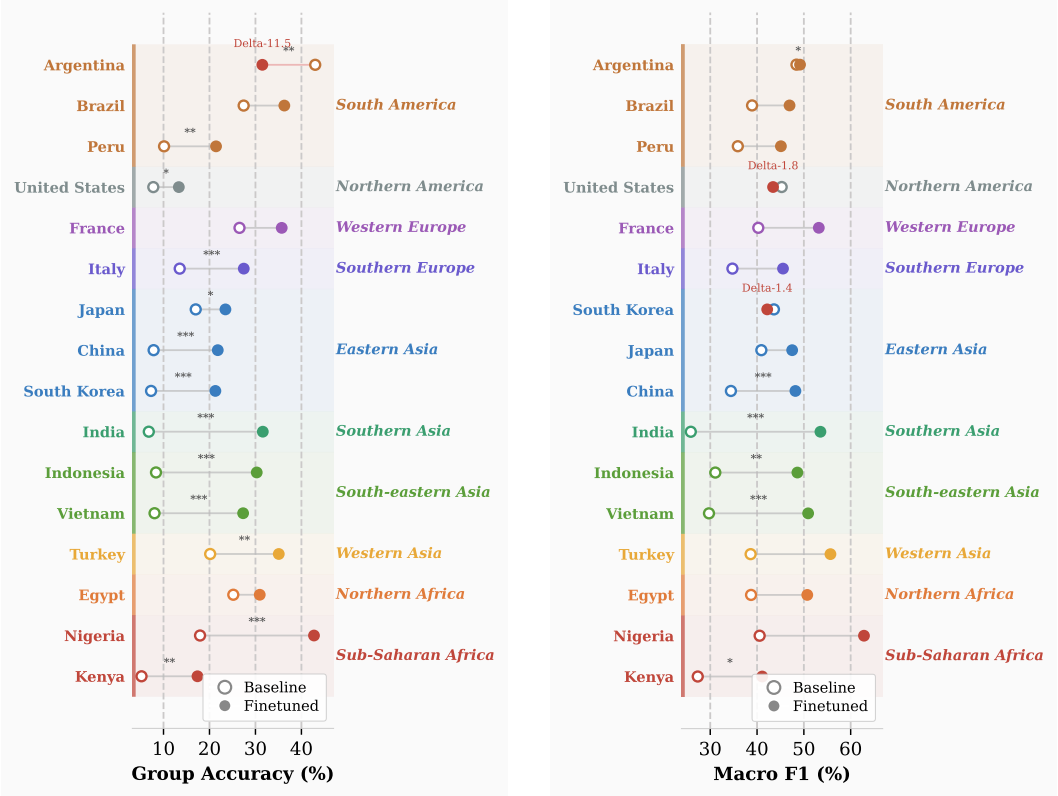


Figure 18: Country-wise performance of finetuned Qwen3VL 4B model, using group accuracy and single image classification F1 score. Results for follow-heavy 64k finetuned Qwen3VL 4B on NORMVIZ-TRAIN

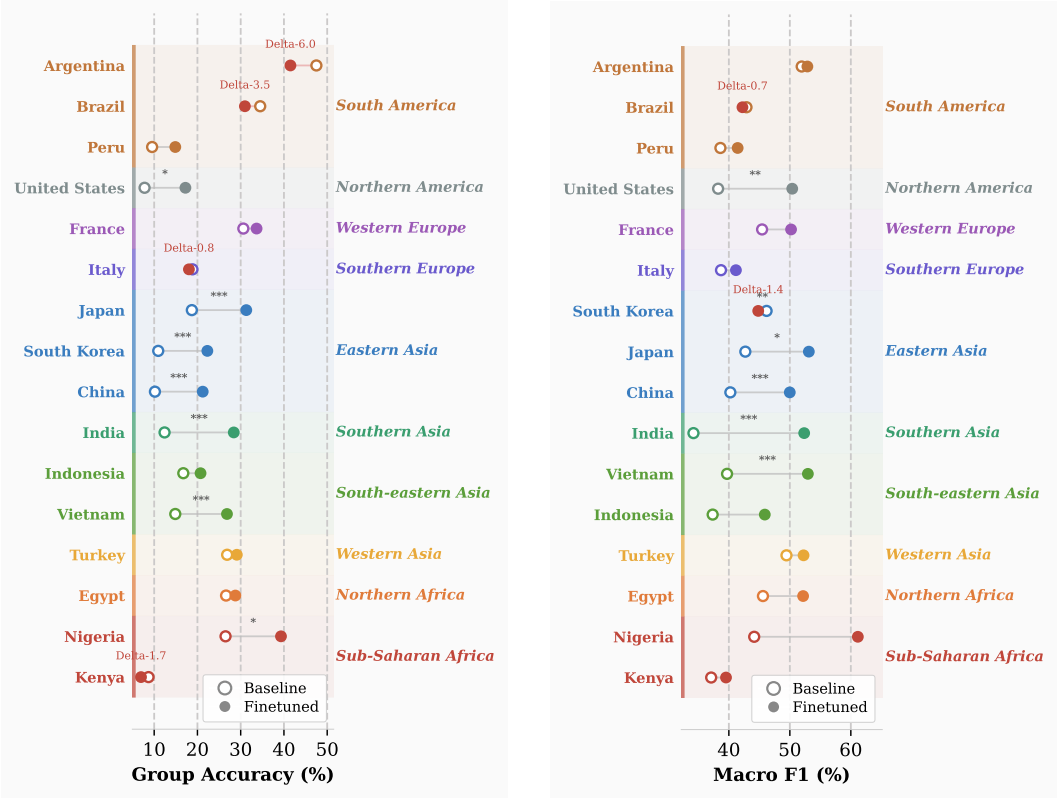


Figure 19: Country-wise performance of finetuned Qwen3VL 8B model, using group accuracy and single image classification F1 score. Results for follow-heavy 64k finetuned Qwen3VL 8B on NORMVIZ-TRAIN

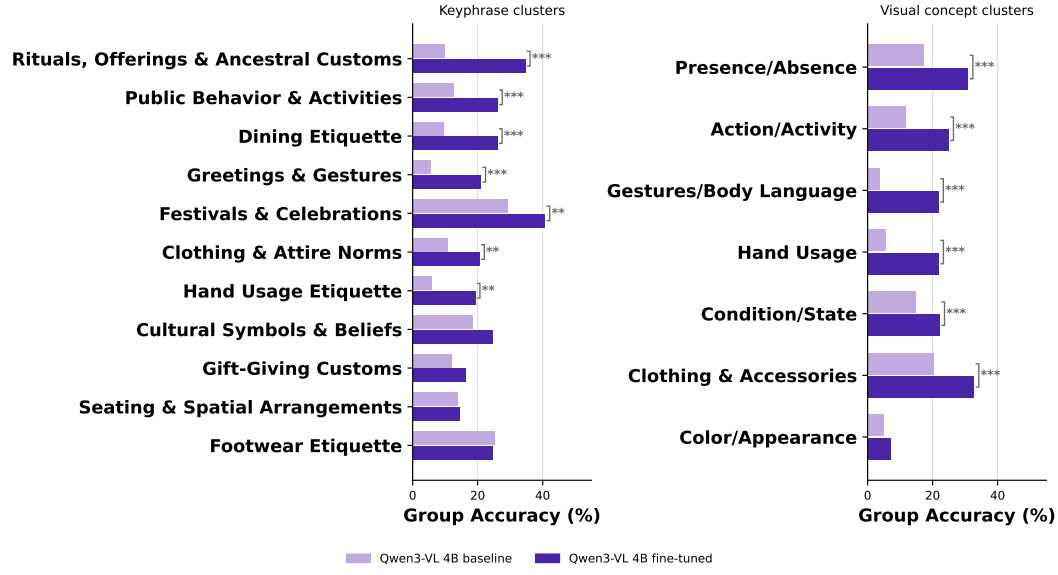


Figure 20: Improvement of finetuned Qwen3 VL 4B (follow-heavy 64k) across cultural norm domains and visual contrast types. Improvements measured using group accuracy.

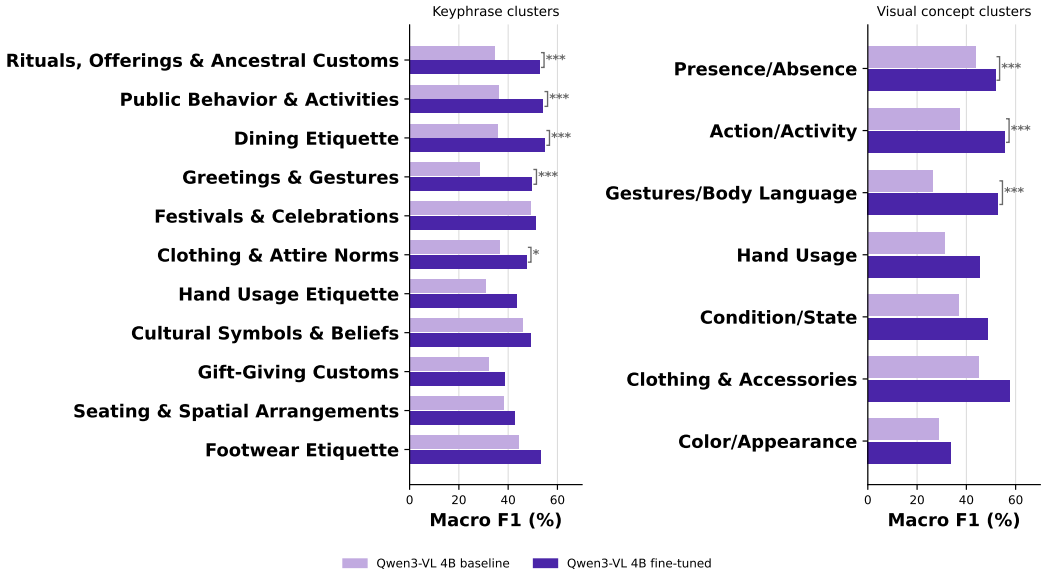


Figure 21: Improvement of finetuned Qwen3 VL 4B (follow-heavy 64k) across cultural norm domains and visual contrast types. Improvements measured using single image classification F1.

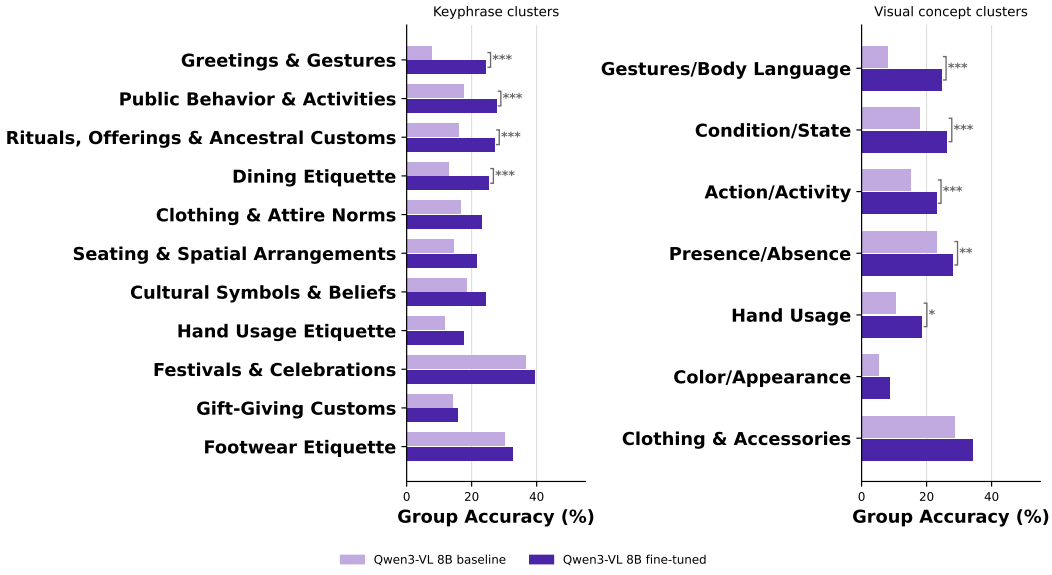


Figure 22: Improvement of finetuned Qwen3 VL 8B (follow-heavy 64k) across cultural norm domains and visual contrast types. Improvements measured using group accuracy.

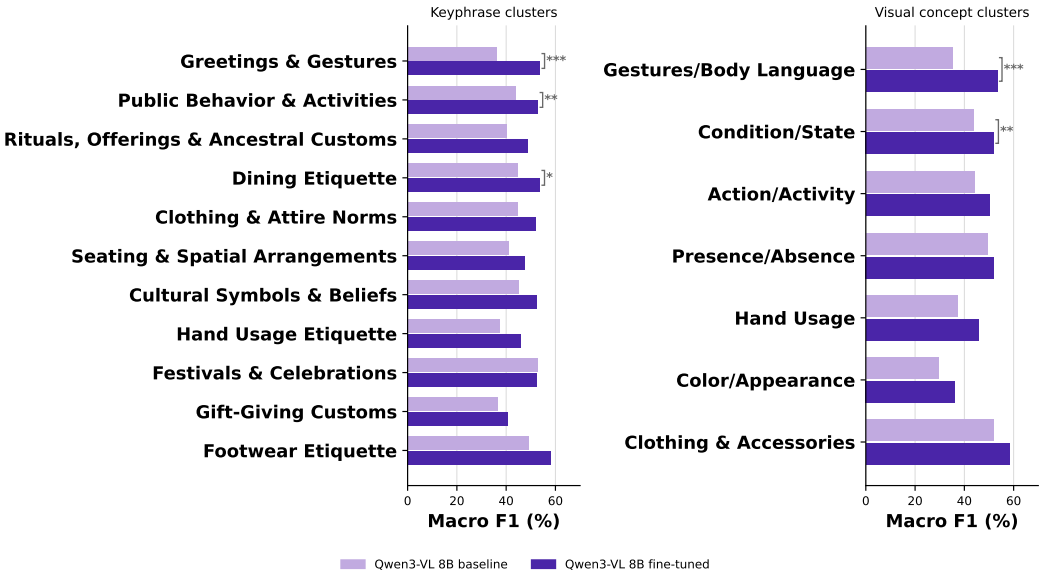


Figure 23: Improvement of finetuned Qwen3 VL 8B (follow-heavy 64k) across cultural norm domains and visual contrast types. Improvements measured using single image classification F1.