

Whose Values Are Generated? Measuring Visual Value Leakage in Multilingual Text-to-Image Generation

Zucan Lyu^{1,*} Binwei Yao^{1,*} Rheeja Uppaal¹

Jing Yao^{2,†} Xiaoyuan Yi^{2,†} Junjie Hu^{1,†}

¹University of Wisconsin–Madison ²Microsoft Research Asia

*Equal contribution. †Corresponding authors.

Abstract

Text-to-image (T2I) systems serve users across languages and cultural contexts. However, evaluations of cultural representation focus mainly on recognizable elements such as landmarks, clothing, and activities, while overlooking the cultural values conveyed by generated images. We introduce **VisionValueBench**, a multilingual benchmark that operationalizes six cultural value dimensions from Hofstede’s framework, such as individualism versus collectivism, to investigate how T2I systems frame neutral social situations in a particular cultural orientation. We find that value-neutral prompts allowing diverse cultural interpretations nevertheless produce images that favor particular value orientations, a phenomenon we call *visual value leakage*. Across 960 social scenarios and six T2I configurations, we evaluate prompts varying in language and geographic cues. Generated images consistently favor low power distance, individualism, care and quality of life, long-term orientation, high uncertainty avoidance and indulgence. Prompt language has limited influence on value orientations, whereas geographic cues shift some dimensions toward associated cultural values while still preserving the prevailing orientations. These findings reveal substantial homogeneity in the cultural values conveyed by T2I systems worldwide.

1 Introduction

AI systems increasingly mediate how people create, learn, and communicate (Shen, 2025), and AI-generated content can shape the judgments and values of the people who encounter it (Shen et al., 2026; Glickman & Sharot, 2025). This issue is particularly important for *multilingual text-to-image (T2I) systems*, which increasingly serve users who express themselves in different languages and come from different cultural contexts. Generated images can implicitly convey cultural values (Nayak et al., 2025), potentially presenting users with particular value orientations without making them explicit. Yet existing evaluations of cultural representation in T2I systems primarily examine whether images contain recognizable cultural elements, including concepts (Saxon & Wang, 2023), artifacts (Liu et al., 2023; Kannen et al., 2024), stereotypes (Jha et al., 2024), and activities (Malakouti et al., 2026; Nayak et al., 2025). These evaluations measure *what cultural content* an image depicts, but not *whose cultural values* the depiction reflects.

Measuring cultural values presents two fundamental challenges. First, existing benchmarks are not designed to isolate value orientation from the underlying social situation: we lack curated scenarios in which the same situation can be plausibly depicted according to different cultural values. Second, cultural values are abstract and context-dependent, making them difficult to operationalize and reliably identify from visual outputs. The closest existing work, CulText2I (Ventura et al., 2025), infers value orientation from image-text similarity and broad VQA judgments over concept-level prompts, but these signals are not validated as distinguishing cultural value from visual content or style. These challenges leave open a fundamental question:

Whose values are generated by text-to-image generation systems?



Figure 1: **Six cultural orientations illustrated by generated images (top). Cultural value profiles of T2I systems benchmarked with VisionValueBench (bottom), revealing substantial homogeneity across all six systems and cultural dimensions.** HiDream-O1 and FLUX.2[dev] are shown with their prompt refiner and upsampler, and Ideogram 4.0 with Magic Prompt; Appendix Table 5 lists full model identifiers and settings.

To answer this question, we focus on **value-neutral** prompts that instruct a T2I system to visualize a social situation without specifying its value orientation. For example, a prompt describing *a teacher and students communicating in a classroom* does not specify how authority and participation are depicted: the image could show a teacher leading the class while students raise their hands to speak, or a teacher leaning toward a speaking student and listening attentively (Figure 1). Both are compatible with the prompt, but reflect different value orientations. This underdetermination allows us to study values expressed by the generator itself. We call this phenomenon **visual value leakage**: the systematic expression of a value orientation as a generator resolves the ambiguity in a value-neutral prompt.

Instantiating this idea, we construct **VisionValueBench**, a multilingual benchmark for measuring visual value leakage. Grounded in Hofstede’s six cultural dimensions (Hofstede, 2011), our benchmark covers six pairs of contrasting value orientations. Building on CDEval (Wang et al., 2023), we adapt text-based value questions into image-generation scenarios that preserve the underlying social situation while leaving its value orientation open. These scenarios provide controlled data for measuring how generators resolve value-relevant ambiguity when generating images. We then characterize each generator with a six-dimensional *value profile* based on its generated images. This benchmark allows us to not only compare value orientations across generators, but also investigate how value orientation shifts with prompt language and geographic cues that specify the country in which the scene takes place.

To reliably infer orientations of a generated image, we develop two complementary evaluation strategies:

Scenario-Anticipated Evaluation (SAE) and Image-Adaptive Evaluation (IAE) (Figure 3). SAE anticipates how a value might be expressed in the given scenario, while IAE adaptively searches the generated image for visual evidence supporting each orientation. To mitigate the bias of relying on a single model’s judgment, we leverage a multi-agent system to check if the cultural evidence is grounded in the image and that its value interpretation fits the scenario. On a human-annotated subset, our IAE orientation decisions achieve 82.92% agreement with the human reviewers.

Across six T2I systems including both commercial and open-weight models, we observe a striking degree of cross-system value consistency: images generated by all systems tend to express low power distance, individualism, care and quality of life, long-term orientation, high uncertainty avoidance, and indulgence, shown in Figure 1 (bottom). These value orientations largely persist across changes in prompt language and geographic cues. Switching from English to Chinese preserves score directions across all systems and dimensions in the no-cue comparison. Switching from US to China cues produces larger changes, primarily in orientation strength, while largely preserving the prevailing directions. These findings suggest that multilingual T2I generation can expose users across cultural contexts to broadly shared value orientations, with limited control over which orientations prevail through prompt language and geographic cues alone.

2 Related Work

Culture in text-to-image generation. Cultural evaluation of T2I models began with recognizable cultural content: whether generated artifacts match a target culture and how widely they cover it (Liu et al., 2023; Kannen et al., 2024; Bayramli et al., 2025), how faithfully long-tail concepts are represented (Rege et al., 2025), and which cultures an image can be attributed to (Yayavaram et al., 2026). A related task adapts an existing image for another culture instead of generating one (Khanuja et al., 2024). Attention then moved to people and practices: stereotypical portrayals of national identities (Jha et al., 2024), everyday activities in which nationality prompts elicit traditional attire (Alsudais, 2026), the cultural faithfulness of social activities (Malakouti et al., 2026), the expectations that local viewers hold about a depicted scene (Nayak et al., 2025), and gestures whose meaning shifts across countries (Yerukola et al., 2025). A third line varies how the prompt itself is expressed: multilingual prompts cover concepts unevenly (Saxon & Wang, 2023), generation in other languages drifts toward culturally neutral or English-dominant content (Shi et al., 2026), the surface language of a prompt can override the cultural identity it names (Holtermann et al., 2026), and commercial systems silently revise prompts before generation, which is itself a source of cultural stereotyping (Urman et al., 2026). Together these lines show that both the culture a prompt asks for and the way it is expressed shape what a model generates. This is why we hold the scenario fixed in VisionValueBench and compare value orientations across languages and geographic cues, rather than varying the requested culture itself.

Cultural values in text-to-image generation. Value-oriented evaluation is most developed for language models, where contextualized questions measure the orientations a model expresses along cultural dimensions (Wang et al., 2023), and where internalized values can shape answers that conflict with what a user asks for (Betley et al., 2026). AwarePrompt examines whether T2I models can depict explicitly specified human values in a form that viewers can recognize (Ciupinska et al., 2024). Closest to our setting, CulText2I projects generated images onto bipolar cultural dimensions for short concepts across languages and nationalities (Ventura et al., 2025). It scores these projections using common automatic T2I evaluation tools, including image-text similarity (Hessel et al., 2021), visual question answering (Hu et al., 2023; Cho et al., 2024; Lin et al., 2024), and multimodal judges (Ku et al., 2024). Related evaluation methods use fine-grained textual analysis for aesthetics (Cao et al., 2026), and multimodal judges for visualization quality (Xie et al., 2026). CulText2I uses short concepts rather than situations that put a value orientations at stake, and its judgments are not tied to verifiable visual evidence. We instead pose each value questions as a scenario with its orientation left open, tie every judgment to identified visual evidence, and record whether an image supports one orientation, both, or neither. We verify both the presence of the evidence and whether it supports the attributed orientation. Table 3 in Appendix A.2 compares the closest benchmarks.

3 VisionValueBench: Value-Neutral Prompts Construction

We construct VisionValueBench, a multilingual benchmark comprising 2,773 social scenarios across six cultural value dimensions. Each scenario specifies a value-relevant situation while leaving its orientation open. This section describes how we derive scenarios from existing value questions, filter out scenarios that prescribe a value orientation, validate that retained scenarios leave value orientations open, and vary prompt language and geographic cues. Figure 2 summarizes the construction pipeline. Figure 1 illustrates scenarios spanning all six cultural dimensions and how their visual realizations can express contrasting value orientations.

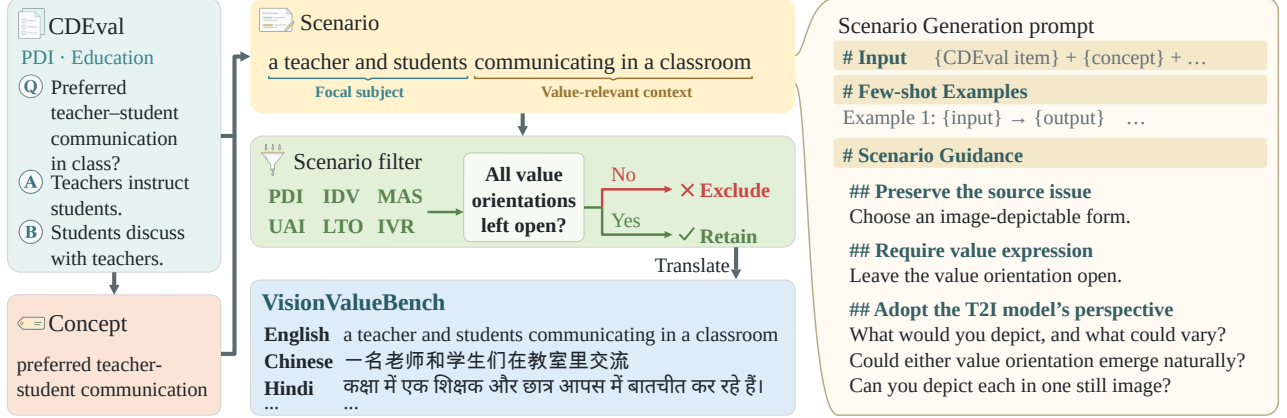


Figure 2: **VisionValueBench construction pipeline.** Building on CDEval, a text-based benchmark grounded in Hofstede’s cultural dimensions, we (1) convert contextualized questions into prompts describing value-neutral social scenarios, (2) filter out scenarios that constrain or preclude certain cultural orientations, and (3) translate the resulting English prompts into other languages.

3.1 Cultural value knowledge base

Source. Social scenarios need a broad, independently validated source of value issues rather than ones we select ourselves, which could bias what the benchmark can detect. We derive scenarios from CDEval (Wang et al., 2023), which operationalizes contextualized value questions across Hofstede’s six cultural dimensions (Hofstede, 2011) using 2,953 questions in seven domains. Each item pairs a question with two contrasting response options: the question sets the issue, and the options show how value orientation can differ.

Concept extraction. CDEval’s options already commit to specific wording, which would prescribe an orientation if carried into a prompt directly. We use a language model to extract a *concept*: the core issue shared by the question and both options, without selecting either response. The CDEval-to-Concept step in Figure 2 illustrates this extraction with the shared issue of preferred teacher–student communication.

3.2 Cultural value scenario construction

A prompt can only reveal visual value leakage if it presents a value-relevant situation without resolving it. We therefore design each underlying situation, or *scenario*, to satisfy two requirements. *Value relevance with orientation openness*: a concrete issue must be at stake, but the prompt must not prescribe which orientation is expressed, otherwise there is nothing for the generation system to decide. *Visual expressibility*: the value-related difference must be able to appear in the main visible content of a single image, otherwise it cannot be scored from the image at all.

Scenario structure. Each scenario has a *focal subject* (who or what the image depicts), a *value-relevant context* (what ties the subject to a value issue), and an *open-ended realization* (the value-relevant aspect left unspecified). The open-ended realization is usually an intentional omission, since specifying it could prescribe an orientation. In Figure 2 (Scenario), the teacher and students form the focal subject, and their classroom

communication provides the value-relevant context. How they communicate remains unspecified, allowing different power-distance orientations to emerge.

Balancing specificity and openness in scenario construction. Specificity creates a trade-off: a scenario that is too broad may lose the value-relevant tension, whereas one that is too specific may prescribe an orientation or let incidental details dominate the image. We balance this trade-off according to three criteria: **depictability**, **minimal specification**, and **openness**. First, the scenario must translate the source issue into a depictable situation without reducing it to a generic scene or explicitly naming the underlying value. Second, it should specify only the context necessary to establish the issue, while removing wording that selects an orientation or introduces visually dominant but irrelevant details. The prompt determines what must be depicted, but leaves out how it is realized to the image generator. Third, the final wording must naturally admit different value orientations through variations in the image’s main visible content, rather than relying on text, sequence, or subtle details. We operationalize these criteria in a single prompt with few-shot examples, illustrated in the Scenario Generation prompt block of Figure 2, detailed in Appendix F.2.

Filtering out prescribed orientations. The design questions guide generation but do not guarantee that every generated scenario satisfies these criteria. Hence, we filter every scenario afterward (Figure 2, Scenario filter). Content retained or introduced during construction may still prescribe an orientation, on either the target dimension or another dimension silently narrowing what the benchmark measures. We therefore check all six dimensions with the same orientation criteria used in image evaluation, requiring each judgment to quote supporting text for traceability. For example, “*a manager presenting a reward to either an individual or a team for excellent work performance.*” leaves the individualism–collectivism distinction open but already prescribes achievement orientation through “excellent work performance”. Model-assisted screening and human review remove 180 of 2,953 scenarios, leaving 2,773 (Appendix A.3).

3.3 Prompt language and geographic cue

Translation. VisionValueBench is meant to reveal whether leakage shifts with prompt language, so translation must preserve the scenario, not just its topic; drift would be indistinguishable from a genuine language effect. Our rubric requires natural target-language wording that retains the original meaning, detail, and intentional omissions, while excluding cultural adaptation, which could supply the orientation the scenario is meant to leave open. We translate English (EN) scenarios into 3 languages: Chinese (ZH), Thai (TH), and Hindi (HI), as illustrated by the Translate step in Figure 2.

Geographic cue. Prompt language and geographic cues are two distinct ways a user can signal culture, so we vary them independently. A geographic cue specifies the country in which the scene takes place (e.g., “Set the scene in China”). Each language has a no-cue condition as a baseline. English is additionally paired with cues for the United States (US), China, Thailand, and India; Chinese, Thai, and Hindi are each paired with their corresponding country and the United States, letting us compare a matching cue against a mismatched one. This yields 14 conditions (Appendix Table 6), varying language and geographic cue while holding the scenario fixed. We denote the three shared geographic-cue conditions as *no cue*, *US cue*, and *China cue*. Complete prompt examples and rendering templates appear in Appendix F.2.

4 Evaluating Value Orientations in Generated Images

Value-neutral prompts leave orientation open; evaluation must establish which orientations the generated image expresses. We assess all six cultural dimensions because a scenario may involve multiple value issues, and the generator may introduce additional ones.

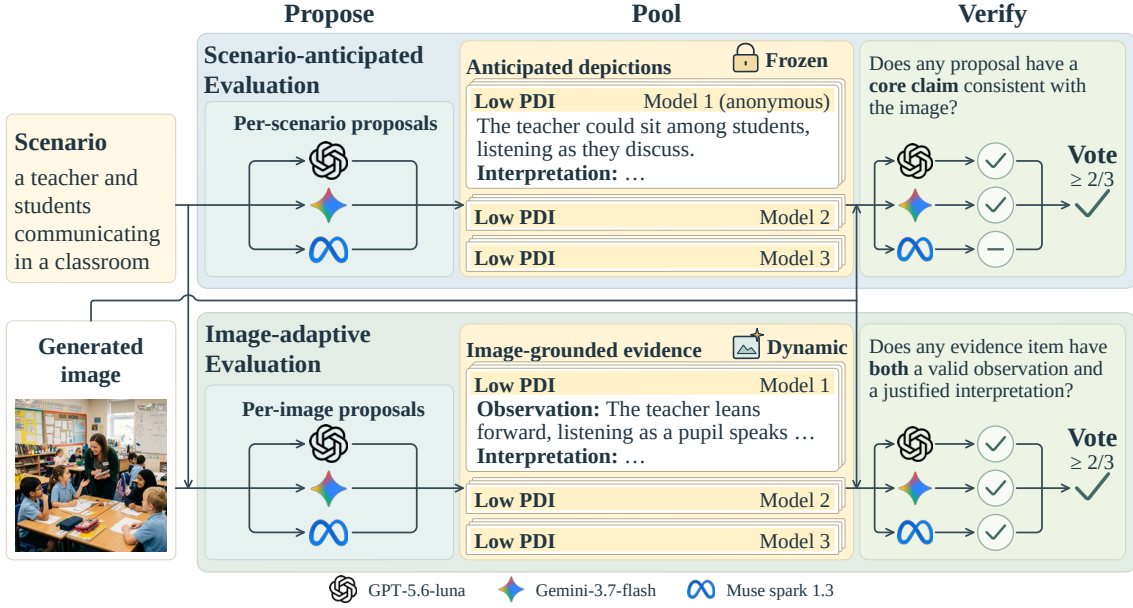


Figure 3: **Evaluation strategies.** Each model call jointly processes all 12 orientations.

4.1 Guiding value evaluation with the focal situation

Recognizing visual content alone does not establish an image’s value orientation: VLMs may attend to irrelevant details or misinterpret a situation despite recognizing its objects (Bae et al., 2026; Song et al., 2025). We therefore use the *focal situation* to guide evidence selection and interpretation. This task-guided selection is consistent with findings that human gaze patterns vary with the viewing task (Borji & Itti, 2014). We interpret each visual detail through its role in the situation when assessing the value orientation it supports.

For each cultural dimension, models identify the issue at stake and assess how concrete visual content expresses an orientation within that situation. Evidence may include composition, people’s appearance and behavior, background details, or value-relevant content introduced by the generator. We apply this principle throughout evidence proposal and verification.

4.2 Discovering and verifying candidate evidence

Figure 3 summarizes how SAE and IAE propose, pool, and verify evidence to infer value orientations. Both follow the focal-situation principle and the same criteria for the two contrasting orientations of each dimension (Appendix F.5).

Evaluation strategies. We consider two complementary strategies for proposing value-relevant evidence: (1) *Scenario-Anticipated Evaluation (SAE)* establishes a shared evaluation reference across images generated from the same scenario (Figure 3, upper row). Before observing any generated image, proposal models use the scenario and dimension-specific criteria to anticipate possible visual evidence and the orientations it would support. The resulting candidate set is reused across systems, languages, and prompt configurations, improving comparability and reducing proposal cost. Each image is still verified independently. (2) *Image-Adaptive Evaluation (IAE)* derives candidate evidence directly from each generated image (Figure 3, lower row). Without access to the SAE candidates, proposal models examine the image together with its scenario to identify concrete visual evidence and the orientations it may support. IAE can therefore capture image-specific evidence, including background details and value-relevant content introduced.

Evidence verification. The *Verify* stage in Figure 3 checks both whether a candidate is visually grounded and whether its interpretation supports the proposed orientation. Within each method, the *Pool* stage combines candidates from multiple proposal models and removes proposer identities. The same candidate bank is then provided to three independent verifiers. A candidate is accepted only if the relevant content is visible and

satisfies the orientation criteria in the focal situation; IAE verifiers may additionally propose missing evidence. Instructions appear in Appendix F.4.

Orientation judgments. A verifier supports an orientation if it accepts at least one candidate or, for IAE, adds sufficient evidence. An orientation is supported when at least two of three verifiers agree, as shown by the *Vote* step in Figure 3. Opposing orientations are judged independently, producing four possible states per dimension: either orientation alone, both, or neither. *Neither* indicates insufficient evidence for either orientation, rather than value neutrality.

Value orientation scores. For each cultural dimension, we summarize orientation support across generated images as

$$S = \frac{n_{A \text{ only}} - n_{B \text{ only}}}{n_{A \text{ only}} + n_{B \text{ only}} + n_{\text{both}}}, \quad (1)$$

where the counts denote images supporting orientation A only, orientation B only, or both.¹ The score ranges from -1 to $+1$, with positive values indicating greater support for A and negative values indicating greater support for B. Images supporting both orientations contribute no net direction; those supporting neither are excluded. The score is undefined when no image supports either orientation. We also report the proportion of images supporting at least one orientation to distinguish balanced support from limited evidence.

For a fixed total of 100 supported images, a $+0.1$ score change corresponds to an increase of ten in the difference between the A-only and B-only counts. Figure 4b illustrates this magnitude with three examples. Further aggregation details appear in Appendix B.

4.3 Validating the evaluation framework

We validate automatic verification against human evidence review of 20 images. Five reviewers, blind to model identity and automatic decisions, accept or reject candidate evidence, flag omissions, and may add missing evidence. Human consensus supports an orientation when at least three of five reviewers do. Human reviewers and VLM verifiers assess the same candidate bank. Across 240 image–orientation pairs (20 images $\times$ 12 orientations), mean pairwise reviewer agreement under IAE review is 87.25%. We compare both strategies against two human-majority references, SAE-H and IAE-H, derived from the SAE and IAE evidence reviews, respectively (Figure 4a). IAE achieves higher overall agreement under both human references and is therefore used for most subsequent analyses. Additional protocol and agreement details appear in Appendix C.

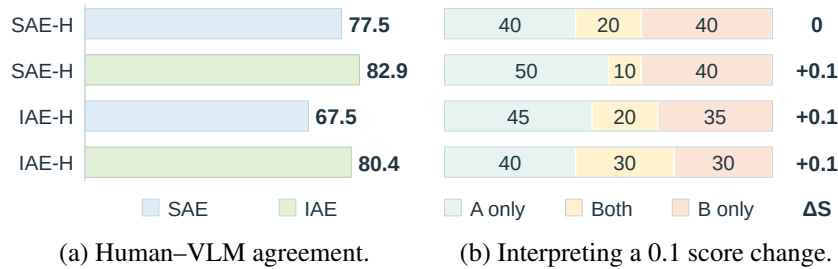


Figure 4: **Human agreement and score interpretation.** (a) IAE achieves higher agreement with both human-majority references, SAE-H and IAE-H, derived from the corresponding evidence reviews. (b) Hypothetical distributions of 100 supported images: each of the last three rows yields $\Delta S = +0.1$ relative to the first.

5 Results and Analysis

We evaluate six T2I systems, including their associated prompt-processing components where applicable, and examine how their value orientations vary with prompt language and geographic cues. We report value profiles

¹Orientation A denotes high power distance (PDI), individualism (IDV), achievement and competition (MAS), high uncertainty avoidance (UAI), long-term orientation (LTO), and indulgence (IVR); B denotes the corresponding opposite orientation (Appendix Table 1).

(a) A traveler choosing a vacation destination for recreation or serenity (English; no cue).



(b) A traveler choosing a vacation destination for recreation or serenity (Chinese; no cue).



(c) A guest interacting with a higher-status person at a social gathering (Nano Banana 2).



Figure 5: **Case studies of cross-system consistency, prompt-language effects, and geographic-cue effects.** (a,b) Within-row comparisons show shared orientations across six systems; vertical comparisons show stability across English and Chinese prompts. Both evaluation strategies support individualism, care and quality of life, and indulgence in all twelve images.

for SAE and IAE separately, using the score defined in Equation 1. Experimental settings and aggregation details are provided in Appendix B, with the rationale for language coverage in Appendix B.1. To summarize the magnitude of language and geographic-cue effects, we use an absolute score change of 0.10 as an interpretable magnitude threshold for substantial change, corresponding to a 10% shift in net orientation support among supported images (Figure 4b; sensitivity analysis in Appendix E.3).

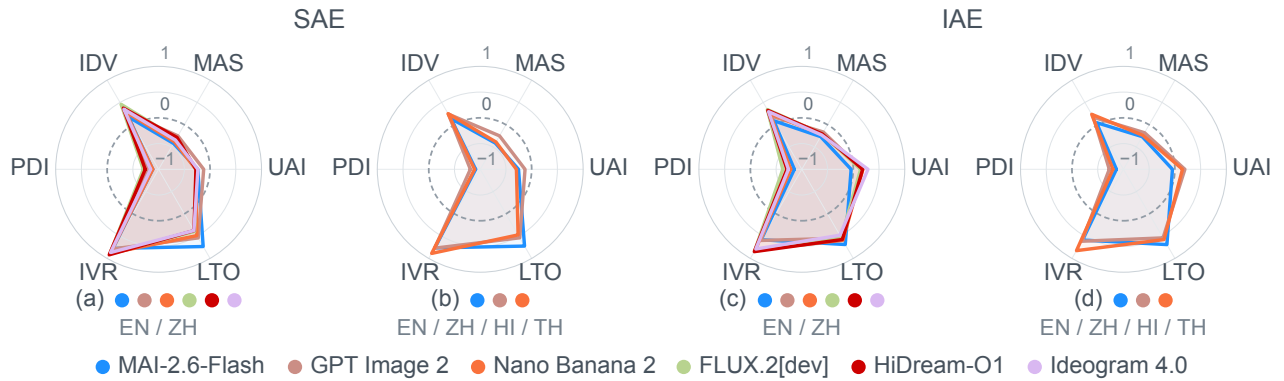


Figure 6: **T2I systems exhibit broadly consistent value orientations across dimensions.** Panels (a,c) show all six systems under the six shared conditions; panels (b,d) show the three commercial API systems under all 14 conditions (Section 3.3).

Value orientations are homogeneous across systems. Figure 6 compares the six-dimensional value profiles of T2I systems under SAE and IAE. Across the six shared conditions, all six systems exhibit broadly homogeneous value orientations. Both evaluation strategies show consistent preferences for low power distance, individualism, care for others and quality of life, long-term orientation, and indulgence. These shared directions

persisted despite differences in their strength.

UAI was the main exception to agreement between the two strategies. SAE favored low uncertainty avoidance, whereas IAE generally favored high uncertainty avoidance. Human review showed that IAE recovered UAI support missed by SAE, supporting its use for identifying evidence on this dimension (Appendix C.2).

IAE is preferred when budget permits; SAE offers a lower-cost alternative. For practitioners assessing the value orientations of their own T2I systems, we recommend IAE when the evaluation budget permits. SAE offers a lower-cost option for an initial value profile. It recovers the same broad orientations on the other five dimensions, although UAI should be treated as method-sensitive.

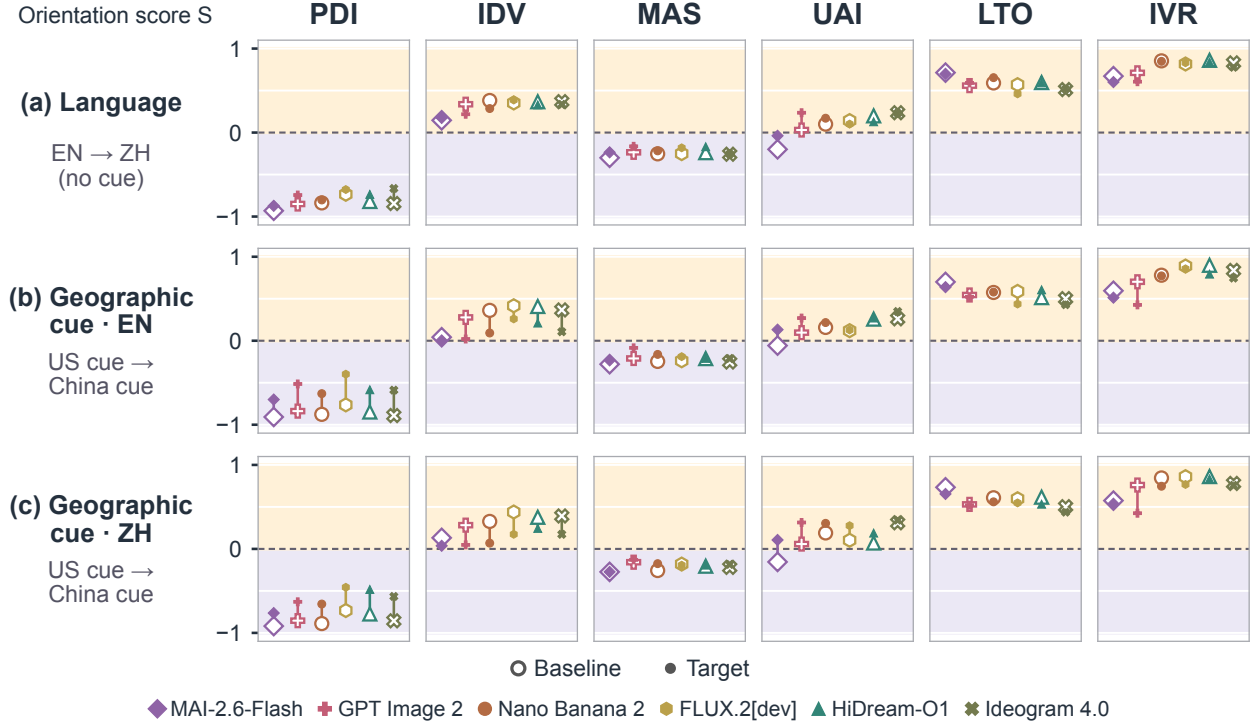


Figure 7: **Language switching preserves score directions; geographic cues reverse the direction only for MAI’s UAI.** Open and filled markers show baseline and target IAE scores on matched scenarios; lines connect scores from the same system. Confidence intervals for score changes appear in Appendix Figure 12.

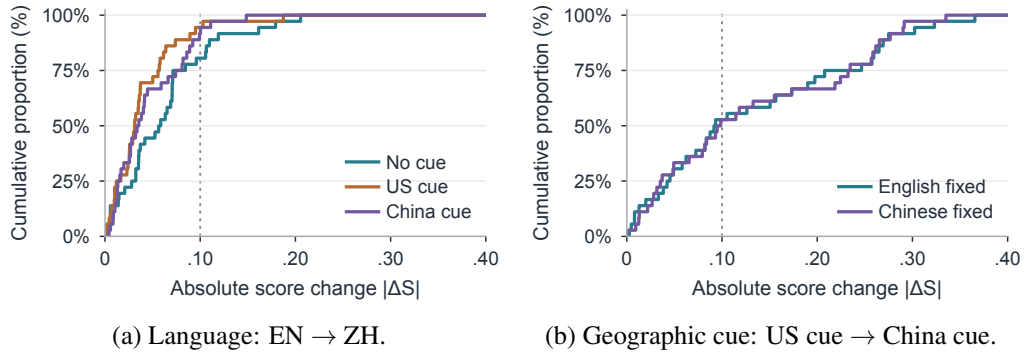


Figure 8: **Language changes are typically small; US-to-China cue changes are larger.** Empirical cumulative distributions of absolute IAE score changes across all six systems.

Prompt language has limited effect on value orientations. Figure 7a compares IAE scores before and after switching from English to Chinese in the no-cue setting. All 36 system–dimension pairs remain on the same side of zero, indicating changes in orientation strength rather than direction. SAE shows the same directional

stability (Appendix D.2). Across the three fixed geographic-cue conditions, 81–94% of comparisons across all six systems have absolute changes below 0.10 (Figure 8a). Figure 5a,b illustrates this stability in a matched scenario.

Geographic cues produce larger shifts while largely preserving orientations. Figures 7b and 7c compare IAE scores under US and China cues in English and Chinese. In each language, 35 of 36 system–dimension pairs retain their score directions. China cues consistently raise PDI and lower IDV scores across all six systems. These directions agree with China’s higher power distance and lower individualism in the 2023 country-level reference (The Culture Factor, n.d.a; n.d.b). The only reversals occur in MAI’s UAI scores, which change from negative to positive in both languages (Appendix D.1). Overall, switching from US to China cues produces larger changes than switching prompt language (Figure 8), but still largely preserves the prevailing orientations.

Adding a US cue produces small changes in value profiles. Absolute IAE score changes are below 0.10 in 94% of English and 83% of Chinese system–dimension comparisons. Across all six systems and both languages, no-cue profiles are closer to US-cue profiles than to China-cue profiles (Appendix D.3). Overall, adding a US cue largely preserves the no-cue value profiles.

Implications for culturally diverse users. Country-level reference profiles differ across cultural contexts (The Culture Factor, n.d.a; n.d.b) (Appendix Table 2), yet the evaluated systems express broadly shared value orientations that largely persist across prompt languages and geographic cues. For culturally diverse users, choosing a different T2I system, writing in another language, or naming a country may therefore offer limited control over which value orientations prevail in generated images. These findings motivate evaluating cultural responsiveness through expressed values alongside recognizable cultural content, and developing ways for users to specify the orientations they intend to depict.

6 Conclusions

We introduce VisionValueBench, a multilingual benchmark for measuring visual value leakage in text-to-image generation across Hofstede’s six cultural dimensions. The benchmark combines value-neutral social scenarios with two evaluation strategies, SAE and IAE, that assess visual evidence and its contextual interpretation. Across 960 scenarios and six T2I systems, generated images exhibit broadly homogeneous value orientations. Prompt language has limited influence on these orientations, while geographic cues primarily change their strength but still largely preserve the prevailing directions. These findings suggest that multilingual T2I systems can expose users across linguistic and cultural contexts to a broadly shared set of cultural values. This homogeneity highlights the importance of evaluating expressed value orientations alongside recognizable cultural content.

Ethics Statement

Human reviewers participated voluntarily and gave informed consent (Appendix C). Our benchmark evaluates model-generated value expressions and should not be used to rank cultures or infer individuals’ values.

Reproducibility Statement

We describe benchmark construction in Appendix A.3 and model configurations and evaluation measures in Appendix B. Prompts and human-review instructions are provided in Appendix F. We will publicly release the benchmark data, evaluation code, and analysis scripts to support reproducibility.

AI use statement

We use generative AI models to construct and translate benchmark scenarios, generate images, and propose and verify visual evidence (Appendices A.3, B, and F.4). Human review of flagged scenarios and image evidence is described in Appendices A.3 and C. The authors take responsibility for the final text, analyses, claims, and artifacts.

References

- Abdulkareem Alsudais. Text-to-image models and their representation of people from different nationalities engaging in activities. *arXiv preprint arXiv:2504.06313v5*, 2026. URL <https://arxiv.org/abs/2504.06313v5>.
- Jiyun Bae, Hyunjong Ok, Sangwoo Mo, and Jaeho Lee. Understanding the effects of distractors on reasoning vision-language models. *arXiv preprint arXiv:2511.21397v3*, 2026. doi: 10.48550/arXiv.2511.21397. URL <https://arxiv.org/abs/2511.21397v3>. EMNLP 2026.
- Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. Diffusion models through a global lens: Are they culturally inclusive? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31137–31155, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1503. URL <https://aclanthology.org/2025.acl-long.1503/>.
- Jan Betley, Johannes Treutlein, Jan Dubinski, Harry Mayne, Karol Galazka, Niels Warncke, Anna Sztzyber-Betley, and Owain Evans. Value leakage: An LLM’s answers are silently shaped by its own values. *arXiv preprint arXiv:2607.14345*, 2026. doi: 10.48550/arXiv.2607.14345. URL <https://arxiv.org/abs/2607.14345>.
- Ali Borji and Laurent Itti. Defending Yarbus: Eye movements reveal observers’ task. *Journal of Vision*, 14(3): 29, 2014. doi: 10.1167/14.3.29. URL <https://doi.org/10.1167/14.3.29>.
- Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajun Yao, and Yihao Liu. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15313–15322, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cao_ArtiMuse_Fine-Grained_Image_Aesthetics_Assessment_with_Joint_Scoring_and_Expert-Level_CVPR_2026_paper.html.
- Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/446905abfec9db8856f7f7465bbd5552-Abstract-Conference.html.
- Kinga Ciupinska, Serena Marchesi, Giulio Abbo, Tony Belpaeme, and Agnieszka Wykowska. AwarePrompt: Using diffusion models to create methods for measuring value-aware AI architectures. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, pp. 1436–1443. SCITEPRESS, 2024. doi: 10.5220/0012596400003636. URL <https://doi.org/10.5220/0012596400003636>.
- Moshe Glickman and Tali Sharot. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9:345–359, 2025. doi: 10.1038/s41562-024-02077-2. URL <https://www.nature.com/articles/s41562-024-02077-2>.

- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL <https://aclanthology.org/2021.emnlp-main.595/>.
- Geert Hofstede. Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture*, 2(1), 2011. doi: 10.9707/2307-0919.1014. URL <https://scholarworks.gvsu.edu/orpc/vol2/iss1/8/>.
- Carolin Holtermann, Florian Schneider, and Anne Lauscher. SoS: Analysis of surface over semantics in multi-lingual text-to-image generation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3955–3995, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.185. URL <https://aclanthology.org/2026.eacl-long.185/>.
- Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20406–20417, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hu_TIFA_Accurate_and_Interpretable_Text-to-Image_Faithfulness_Evaluation_with_Question_Answering_ICCV_2023_paper.html.
- Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. ViSAGE: A global-scale analysis of visual stereotypes in text-to-image generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12333–12347, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.667. URL <https://aclanthology.org/2024.acl-long.667/>.
- 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. URL <https://arxiv.org/abs/2407.06863>.
- Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10258–10279, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.573. URL <https://aclanthology.org/2024.emnlp-main.573/>.
- Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhua Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.663. URL <https://aclanthology.org/2024.acl-long.663/>.
- 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 *Computer Vision – ECCV 2024*, pp. 366–384. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72673-6_20. URL https://doi.org/10.1007/978-3-031-72673-6_20.

- Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. On the cultural gap in text-to-image generation. *arXiv preprint arXiv:2307.02971*, 2023. URL <https://arxiv.org/abs/2307.02971>.
- Sina Malakouti, Boqing Gong, and Adriana Kovashka. Culture in action: Evaluating text-to-image models through social activities. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=opG4m2U0Oo>.
- Rashid Mushkani, Shravan Nayak, Hugo Berard, Allison Cohen, Shin Koseki, and Hadrien Bertrand. LIVS: A pluralistic alignment dataset for inclusive public spaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45311–45341. PMLR, 2025. URL <https://proceedings.mlr.press/v267/mushkani25a.html>.
- Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20918–20953. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1141. URL <https://aclanthology.org/2025.findings-emnlp.1141/>.
- Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu, Aditya Kusupati, Yong Jae Lee, and Ramya Korlakai Vinayak. CuRe: Cultural gaps in the long tail of text-to-image systems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15680–15691, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Rege_CuRe_Cultural_Gaps_in_the_Long_Tail_of_Text-to-Image_Systems_ICCV_2025_paper.html.
- Michael Saxon and William Yang Wang. Multilingual conceptual coverage in text-to-image models. *arXiv preprint arXiv:2306.01735*, 2023. URL <https://arxiv.org/abs/2306.01735>.
- Hua Shen. Bidirectional human–AI alignment in education for trustworthy learning environments. *arXiv preprint arXiv:2512.21552*, 2025. doi: 10.48550/arXiv.2512.21552. URL <https://arxiv.org/abs/2512.21552>.
- Hua Shen, Tiffany Kneare, Divy Thakkar, Pat Pataranutaporn, Anoop Sinha, Yike Shi, Jenny T. Liang, Lama Ahmad, Tanu Mitra, Brad A. Myers, and Yang Li. Human–AI interaction alignment: Designing, evaluating, and evolving value-centered AI for reciprocal human–AI futures. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2026. doi: 10.1145/3772363.3778710. URL <https://doi.org/10.1145/3772363.3778710>.
- Chuancheng Shi, Shangze Li, Shiming Guo, Simiao Xie, Wenhua Wu, Jingtong Dou, Chao Wu, Canran Xiao, Cong Wang, Zifeng Cheng, Fei Shen, and Tat-Seng Chua. Where culture fades: Revealing the cultural gap in text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14702–14712, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Shi_Where_Culture_Fades_Revealing_the_Cultural_Gap_in_Text-to-Image_Generation_CVPR_2026_paper.pdf.
- Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, and Yanyi Chen. A cognitive evaluation benchmark of image reasoning and description for large vision-language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6392–6409, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.324. URL <https://aclanthology.org/2025.naacl-long.324/>.

- The Culture Factor. Country comparison tool. <https://www.theculturefactor.com/country-comparison-tool>, n.d.a. Country-score reference incorporating the October 16, 2023 IDV and LTO update; update provenance documented in the official FAQ. Accessed September 24, 2026.
- The Culture Factor. Frequently asked questions. <https://www.theculturefactor.com/frequently-asked-questions>, n.d.b. Sections on renaming MAS and the October 16, 2023 IDV and LTO score update. Accessed September 24, 2026.
- Aleksandra Urman, Elsa Lichtenegger, Salima Jaoua, Azza Bouleimen, Robin Forsberg, Corinna Hertweck, Stefania Ionescu, Nicolò Pagan, Ancsa Hannak, and Joachim Baumann. Prompt revision as a source of cultural bias in text-to-image systems. *arXiv preprint arXiv:2609.11532*, 2026. URL <https://arxiv.org/abs/2609.11532>.
- Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *Transactions of the Association for Computational Linguistics*, 13:142–166, 2025. doi: 10.1162/tacl_a_00732. URL <https://aclanthology.org/2025.tacl-1.10/>.
- Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. CDEval: A benchmark for measuring the cultural dimensions of large language models. *arXiv preprint arXiv:2311.16421*, 2023. URL <https://arxiv.org/abs/2311.16421>.
- Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, and Yuyu Luo. VisJudge-Bench: Aesthetics and quality assessment of visualizations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=lG1HWWdEbN>.
- Arnav Yayavaram, Siddharth Yayavaram, Simran Khanuja, Michael Saxon, and Graham Neubig. CAIRE: Cultural attribution of images with retrieval. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8320–8338, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.389. URL <https://aclanthology.org/2026.eacl-long.389/>.
- Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. Mind the gesture: Evaluating AI sensitivity to culturally offensive non-verbal gestures. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25041–25080, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1218. URL <https://aclanthology.org/2025.acl-long.1218/>.

Appendix Guide

The appendices follow the benchmark, measurement, validation, and results, followed by prompts, annotation materials, and limitations.

- Appendix A: Benchmark construction and filtering.
- Appendix B: Experimental settings and measures.
- Appendix C: Human validation and annotation diagnostics.
- Appendix D: Extended results.
- Appendix E: Robustness and sensitivity analyses.
- Appendix F: Prompts and annotation materials.
- Appendix G: Limitations and future directions.

A Benchmark Construction and Filtering

A.1 Value Dimensions and Orientations

We use Hofstede’s six cultural value dimensions (Hofstede, 2011), summarized in Table 1. For MAS, we adopt The Culture Factor’s updated terminology, *Motivation towards Achievement and Success*. The renaming leaves the dimension’s interpretation and country scores unchanged (The Culture Factor, n.d.b). For each dimension, orientation A corresponds to the higher end of the country-score scale and orientation B to the lower end. Appendix F.5 provides the image annotation criteria.

Table 1: Six cultural value dimensions and their contrasting orientations (A and B).

Code	Value dimension	Orientation A	Orientation B
PDI	Power distance	Acceptance of unequal authority	Equal standing and influence
IDV	Individualism–collectivism	Personal autonomy	In-group obligation
MAS	Motivation towards achievement and success	Achievement and competition	Care and quality of life
UAI	Uncertainty avoidance	Need for certainty	Tolerance of ambiguity
LTO	Long-term orientation	Perseverance and pragmatic adaptation	Quicker returns and normative continuity
IVR	Indulgence–restraint	Relatively free enjoyment	Restraint by norms and duties

Table 2 reports country-level reference scores from The Culture Factor’s Country Comparison Tool (The Culture Factor, n.d.a). The scores incorporate the October 16, 2023 update to IDV and LTO (The Culture Factor, n.d.b).

Table 2: Country reference scores (The Culture Factor, 2023 update; retrieved September 24, 2026). Higher values correspond to orientation A.

Country	PDI	IDV	MAS	UAI	LTO	IVR
United States	40	60	62	46	50	68
China	80	43	66	30	77	24
India	77	24	56	40	51	26
Thailand	64	19	34	64	67	45

Table 3: Comparison of cultural benchmarks and evaluations for text-to-image systems.

Benchmark	Cultural focus	Prompt design	Multilingual	Geographic cues	Evaluation method
CUBE (2024)	Cultural awareness and diversity	Artifacts / broad concepts	✗	Country / no cue	Human ratings; diversity metrics
ViSAGe (2024)	Visual stereotypes	National identities and attributes	✗	Nationality	Human labels; embedding / caption metrics
CuRe (2025)	Long-tail cultural representativeness	Artifacts with varying detail	✗	Region / no cue	Embedding / perceptual metrics; human ratings
CultDiff (2025)	Cultural fidelity	Architecture, clothing, food	✗	Country	Human ratings; learned similarity
Nationalities audit (2025 / rev. 2026)	Attire stereotypes in activities	Everyday activities	✗	Nationality	Human labels; embedding / VLM alignment
CulturalFrames (2025)	Explicit / implicit expectations	Social practices and scenes	✗	Country context	Human ratings
MC-SIGNS (2025)	Gesture meaning and offensiveness	Gestures and social intentions	✗	Country context	Rejection rates; VLM gesture checks
LIVS (Mushkani et al., 2025)	Community value preferences	Public-space scenes	✗	Local places	Human pairwise preferences
CulText2I (2025)	Cultural axes and associations	Short concept templates	✓	Nationality / no cue	Embedding metrics; VQA
CAIRE (2026)	Cultural relevance	Universal cultural concepts	✗	Country	VLM-as-a-judge with retrieval
CultureBench (2026)	Cultural consistency	Descriptions with / without cultural modifiers	✓	Optional cultural modifiers	VLM cultural attribution
SoS (2026)	Language vs. requested culture	Cultural identity templates	✓	Cultural identity	Image-embedding similarity
WORLDVIEW (2026)	Revision-induced cultural bias	Shared everyday topics	✓	Geographic context / no cue	Embedding metrics; lexical analysis
CULTIVate (2026)	Cultural faithfulness in social activities	Social activities and interactions	✗	Country	Descriptor-based metrics
Ours	Cultural value orientations	Social scenarios; value relevance with orientation openness; visual expressibility	✓	Country / no cue	VLM-as-a-judge; evidence proposal and verification

Notes. Multilingual refers to the languages of image-generation prompts in the reported experiments: ✓ indicates multiple languages, and ✗ indicates no multilingual generation experiment reported. Geographic cues summarize reported settings, not necessarily a full factorial design. Evaluation methods list the main procedures. VQA denotes visual question answering. The CAIRE row covers its universal-concept generated-image evaluation, and MC-SIGNS its T2I branch.

Country scores provide society-level references, whereas S summarizes orientation support in generated images. For comparisons between two geographic cues, we compare the sign of ΔS with the sign of the corresponding country-score difference, even when both country scores lie on the same side of the scale midpoint.

A.2 Comparison with Related Benchmarks

A.3 Scenario Construction and Filtering

We derive scenarios from the 2,953 CDEval questions, which cover six cultural value dimensions and seven domains: Arts, Education, Family, Lifestyle, Scientific, Wellness, and Work. Concept extraction identifies the value-relevant issue shared by each question and its two response options. Scenario generation then frames this issue as a depictable situation while leaving its value orientation open (Figure 9; prompts in Appendix F.2).

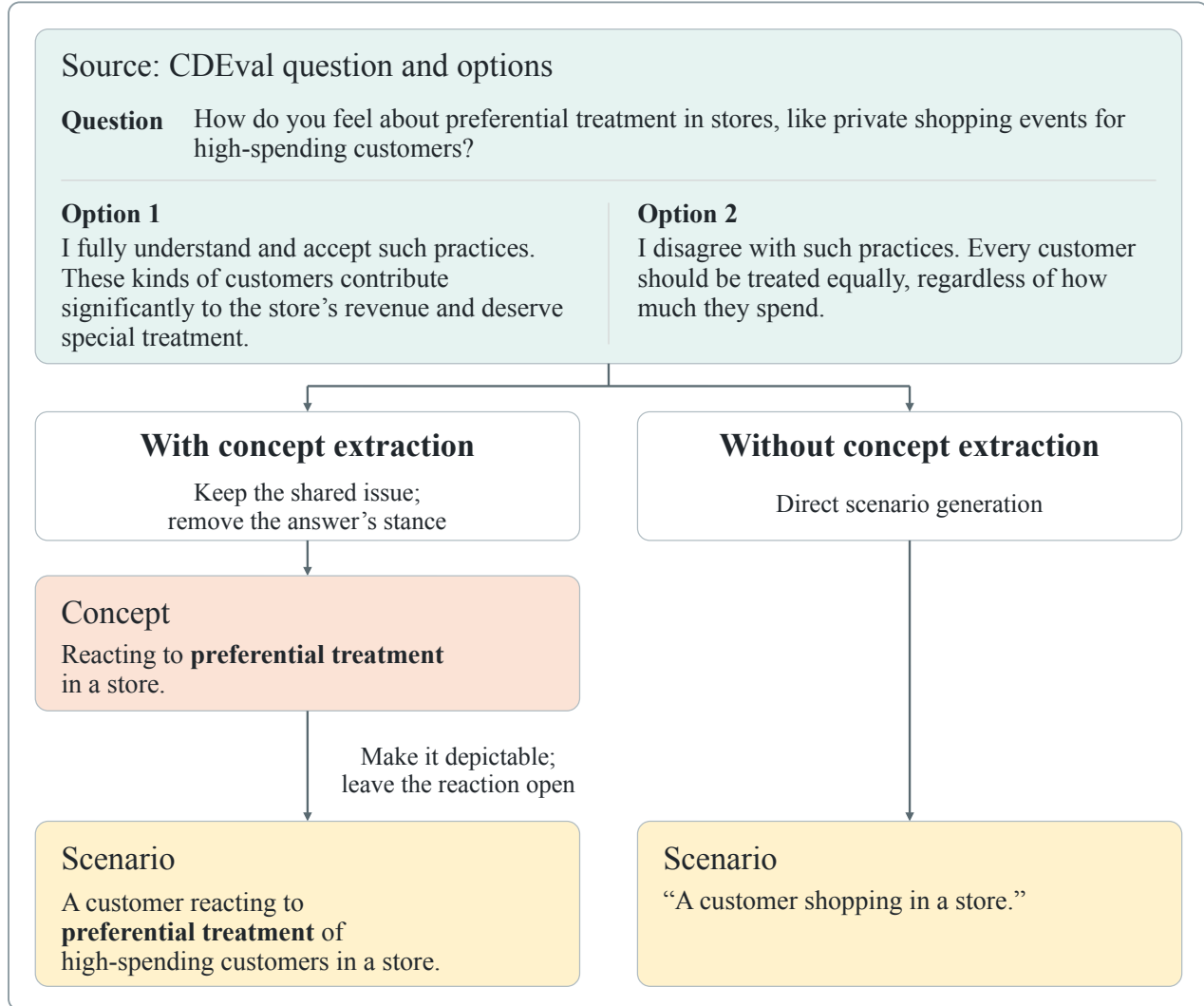


Figure 9: Case study of scenario generation with and without concept extraction using CDEval item `pdi_0052` (Lifestyle) (Wang et al., 2023).

Filtering rules and procedure. We use Gemini 3.7 Flash and Claude Opus 5 to screen all 2,953 scenarios across all six cultural value dimensions. A scenario is flagged only when quoted wording requires any faithful depiction to express a particular value orientation; for example, “a person indulging in their favorite dessert” prescribes indulgence. We review scenarios flagged by either model and remove those confirmed to prescribe an orientation.

Filtering removes 180 scenarios, leaving 2,773 (Table 4). Of the removed scenarios, 90 are flagged on their source dimension, including eight also flagged on another dimension; the other 90 are flagged only on other dimensions. We evaluate 160 retained scenarios per source dimension, for 960 in total.

Table 4: Scenario counts by source dimension before and after filtering.

	PDI	IDV	MAS	UAI	LTO	IVR
Initial	512	472	452	530	485	502
Removed	25	28	44	27	25	31
Retained	487	444	408	503	460	471

B Experimental Settings and Measures

B.1 Image-generation Systems and Prompt Processing

Table 5: Six image-generation models under eight evaluation configurations. Condition counts refer to benchmark prompts before any system-level prompt processing.

Model	Prompt processing	Conditions	Output resolution
Commercial API systems			
◆ Nano Banana 2	Built-in thinking	14	1024 × 1024
⊞ GPT-Image-2	Internal processing	14	Auto
■ MAI-Image-2.6-Flash	Internal processing	14	1024 × 1024
Open-weight image models			
▢ HiDream-O1-Image	Without Prompt Refiner	6	2048 × 2048
▢ HiDream-O1-Image	With Prompt Refiner	6	2048 × 2048
⚡ FLUX.2-dev	Without Prompt Upsampler	6	1024 × 1024
⚡ FLUX.2-dev	With Prompt Upsampler	6	1024 × 1024
⊞ ideogram-4 (FP8)	Magic Prompt API, enabled	6	1024 × 1024

The main analysis uses the prompt-processed HiDream and FLUX configurations (Table 5); their unrefined counterparts serve as paired controls (Appendix D.4). We evaluate ideogram only with Magic Prompt, as short prompts sent directly to the local FP8 model produced refusal screens in preliminary tests.

Language coverage. In preliminary tests, the open-weight models followed Hindi and Thai prompts less reliably than English and Chinese ones (Figure 10). We therefore evaluate them only under the six shared English and Chinese conditions; the three commercial systems cover all 14 conditions (Table 6).

Scenario: A married couple spending their leisure time.



(a) EN

(b) ZH

(c) TH

(d) HI

Figure 10: **Scenario adherence in a multilingual HiDream pilot.** English and Chinese prompts yield a couple at leisure, while Thai and Hindi prompts do not. HiDream-O1 without refinement, no cue.

Table 6: Language and geographic-cue conditions in VisionValueBench. ✓ marks configured conditions and a dash unconfigured ones; * marks the six conditions shared by all configurations.

Prompt language	No cue	US cue	China cue	Thailand cue	India cue
English	✓*	✓*	✓*	✓	✓
Chinese	✓*	✓*	✓*	–	–
Thai	✓	✓	–	✓	–
Hindi	✓	✓	–	–	✓

Generation settings and prompt processing. We submit benchmark prompts unchanged. Any rewriting occurs within the system’s pipeline: HiDream’s Prompt Refiner, FLUX’s Prompt Upsampler, or ideogram’s Magic Prompt. We keep original and processed prompts separately; failed HiDream refinements do not fall back to unrefined prompts. Local generations use seed 20260728 (Table 7).

Table 7: Generation and prompt-processing settings. “Unspecified” marks parameters absent from the recorded API requests.

Component	Model identifier	Settings
Nano Banana 2	gemini-3.1-flash-image	Minimal thinking; seed 20260728
GPT-Image-2	gpt-image-2	Auto size
MAI	MAI-Image-2.6-Flash	$n = 1$; seed unspecified
HiDream	HiDream-01-Image-Dev-2604	With Refiner: flash scheduler, 28 image steps
HiDream Refiner	HiDream-ai/Prompt-Refine	$T = 1$, top- $p = .95$, top- $k = 64$; 2,048 output tokens
FLUX	black-forest-labs/FLUX.2-dev	50 steps; guidance 4; bfloat16
FLUX Upsampler	Mistral-Small-3.2-24B-Instruct-2506	$T = .15$; 512 output tokens
ideogram Magic	ideogram-4-fp8	Magic Prompt enabled; 48 steps

B.2 Evaluation Sample and Completion

Every image is evaluated on all six dimensions. We retain one image per available scenario–condition–configuration with complete SAE and IAE judgments from all three verifiers (Table 8).

Table 8: Images with complete annotations / planned images. Shared conditions are EN/ZH $\times$ no/US/China cue; only the three commercial systems have the remaining conditions.

Configuration	Shared conditions	All configured conditions
MAI	5,701 / 5,760	13,302 / 13,440
GPT-Image-2	5,650 / 5,760	13,243 / 13,440
Nano Banana 2	5,753 / 5,760	13,426 / 13,440
FLUX + Upsampler	5,760 / 5,760	5,760 / 5,760
HiDream + Refiner	5,760 / 5,760	5,760 / 5,760
ideogram-4 Magic	5,753 / 5,760	5,753 / 5,760
FLUX control	5,760 / 5,760	5,760 / 5,760
HiDream control	5,760 / 5,760	5,760 / 5,760

Table 9 gives the evaluated images N and the supported-image counts U that serve as score denominators.

Table 9: Supported-image counts U (score denominators) under the six shared conditions. N counts completely annotated images and is shared across dimensions and methods.

Configuration	N	PDI	IDV	MAS	UAI	LTO	IVR
<i>SAE</i>							
MAI	5701	927	3165	2302	674	1650	923
GPT-Image-2	5650	897	2443	1911	631	1242	800
Nano Banana 2	5753	952	2245	1605	514	963	1017
FLUX + Upsampler	5760	678	1708	1005	350	561	765
HiDream + Refiner	5760	741	1812	1257	364	639	768
ideogram-4 Magic	5753	822	1965	1328	410	794	755
FLUX (control)	5760	661	1290	791	234	316	607
HiDream (control)	5760	688	1400	1092	276	433	1044
<i>IAE</i>							
MAI	5701	1827	4896	5384	2252	4558	2013
GPT-Image-2	5650	1743	4052	4916	2026	3605	1960
Nano Banana 2	5753	1644	3463	4526	1328	2573	1938
FLUX + Upsampler	5760	1318	2850	3935	1131	1766	1681
HiDream + Refiner	5760	1311	2827	4126	1094	1870	1690
ideogram-4 Magic	5753	1552	3271	4311	1199	2403	1790
FLUX (control)	5760	1188	2395	3364	905	1336	1429
HiDream (control)	5760	1240	2642	4055	790	1477	2136

B.3 Four-state Outcomes and Directional Scores

Separate support judgments for orientations A and B yield four states per dimension: A only, B only, both, and neither. Neither means that support was not established, not that the image is value-neutral. Alongside S (Equation 1), we report support prevalence,

$$C = \frac{n_{A \text{ only}} + n_{B \text{ only}} + n_{\text{both}}}{N}, \quad (2)$$

where N counts images in all four states. S measures direction among supported images, and C measures how often support is established. Images supporting both orientations enter S with weight 0. Identical S and C can still arise from orientations co-occurring within images or opposing each other across images.

B.4 Aggregation, Pairing, and Uncertainty

We compute profile scores S from state counts pooled over the six shared conditions.

Language and cue comparisons pair scenarios within each system, hold the other prompt factor fixed, and include only scenarios available in both conditions. Each condition uses its own supported-image denominator. We report $\Delta S = S_{\text{target}} - S_{\text{baseline}}$.

We estimate 95% confidence intervals for profile scores and paired score differences using 5,000 bootstrap replicates. Each replicate samples 160 scenarios with replacement per source dimension, keeping their conditions, systems, dimensions, and verifier judgments together. We then recompute both condition scores and their difference. With one image per prompt setting, the intervals reflect variation across scenarios. Pointwise intervals use bootstrap percentiles.²

We use approximate 95% simultaneous confidence intervals to account for multiple comparisons in Tables 12 and 13. Within each method (SAE or IAE), all 36 profile estimates form one group, and each prompt comparison forms another (six main configurations $\times$ six dimensions).³

²The bootstrap seed is 20260925. Intervals are reported only when at least 95% of replicates yield defined estimates.

³The critical value is the 95th percentile of the bootstrap maximum standardized absolute deviation.

C Human Validation and Annotation Diagnostics

C.1 Human Evidence-review Protocol

Five undergraduate and doctoral students from our laboratory, with cultural backgrounds spanning China, India, and the United States, reviewed the same 20 images after training and practice. Participation was voluntary and uncompensated, and all reviewers gave informed consent. They could ask the authors for clarification of the review instructions (Appendix F.6). Reviewers inspected the frozen bank of 388 SAE and 234 IAE candidates pooled from the three production proposers, with proposer identities and automatic decisions hidden, and could add their own evidence.

We derive each reviewer’s orientation labels from their evidence review: an orientation is supported if they accept at least one corresponding candidate or add complete supporting evidence. Human consensus requires support from at least three of the five reviewers. We recompute automatic orientation labels from each verifier’s accepted candidates in the same bank and, for IAE, its added evidence, retaining the two-of-three aggregation rule. Agreement therefore measures consistency between human and automatic review of this shared candidate bank.

Sample selection. We selected 20 images for human review. The sample included five Nano Banana 2 images and three each from HiDream, HiDream + Refiner, FLUX, FLUX + Upsampler, and ideogram-4 Magic (no GPT or MAI). It covered four PDI scenarios, four IVR scenarios, and three from each other source dimension. Two images the authors judged to lack value content were replaced within the same configuration, source dimension, and prompt condition.⁴

C.2 Human Validation Results

Table 10 reports agreement and Cohen’s κ by dimension against human-majority labels from the corresponding evidence reviews: SAE-H for SAE and IAE-H for IAE. Confidence intervals for agreement use 10,000 image-level bootstrap resamples (seed 20260925). Across 240 image–orientation pairs, SAE has 19 positive agreements, 54 human-only positives, no model-only positives and 167 negative agreements; IAE has 52, 45, 2 and 141.

Table 10: Agreement with human review by dimension (%), with 95% intervals on the 0–1 scale. Each dimension has 40 image–orientation pairs.

Dim.	SAE versus SAE-H			IAE versus IAE-H		
	Agree.	95% CI	κ	Agree.	95% CI	κ
All	77.5	[0.717, 0.829]	0.329	80.4	[0.738, 0.871]	0.562
PDI	82.5	[0.725, 0.925]	0.307	75.0	[0.575, 0.900]	0.322
IDV	55.0	[0.425, 0.675]	0.137	67.5	[0.525, 0.825]	0.387
MAS	75.0	[0.625, 0.875]	0.342	80.0	[0.675, 0.925]	0.614
UAI	75.0	[0.625, 0.875]	0.000	87.5	[0.775, 0.950]	0.600
LTO	87.5	[0.775, 0.950]	0.635	85.0	[0.750, 0.950]	0.652
IVR	90.0	[0.800, 0.975]	0.615	87.5	[0.775, 0.975]	0.684

Against the corresponding human-majority reference, positive/negative agreement is 41.3%/86.1% for SAE and 68.9%/85.7% for IAE.⁵ Among reviewers, mean pairwise agreement is 78.58% for SAE and 87.25% for IAE; Fleiss’ κ is 0.513 and 0.731, respectively. Table 11 compares both automatic methods against both human references.

⁴Initial selection used seed 20260916; replacements used seed 20260917 and the first author-accepted candidate.

⁵Positive and negative agreement are $2a/(2a + b + c)$ and $2d/(2d + b + c)$, where a, d count positive and negative agreements and b, c count disagreements.

Table 11: Agreement with both human-majority references (240 image–orientation pairs overall; 40 for UAI). Detection counts automatically supported pairs among the ten human-supported UAI pairs.

Metric	Automatic method	SAE-H	IAE-H
Overall agreement (%)	SAE	77.50	67.50
	IAE	82.92	80.42
UAI agreement (%)	SAE	75.0	75.0
	IAE	87.5	87.5
Human-supported UAI pairs detected	SAE	0/10	0/10
	IAE	5/10	5/10

UAI support. IAE detects the same five human-supported UAI pairs under both human references, while SAE detects none (Table 11). These results concern orientation support in the reviewed images, not the direction of individual systems’ production UAI scores.

Evidence disagreements. Automatic majority rejects 79 of the 216 IAE candidates accepted by human majority; candidate acceptance agreement is 64.10% ($\kappa = 0.124$). For SAE, agreement on whether candidate content is shown in the image is 71.39%. At the orientation level, IAE-H supports 33 of the 35 pairs supported only by IAE, and 45 of the 186 pairs supported by neither method. Reviewers added 14 IAE evidence records.

D Extended Results

D.1 Complete Value Profiles

Table 12 reports pooled orientation scores for each main configuration under the six shared conditions. SAE and IAE agree on score signs in all dimensions other than UAI.

Table 12: Pooled orientation scores under the six shared conditions. Bold: within-method simultaneous 95% interval excludes zero.

Method	System	PDI	IDV	MAS	UAI	LTO	IVR
SAE	MAI	−0.886	0.176	−0.420	−0.239	0.733	0.777
SAE	GPT Image 2	−0.805	0.273	−0.250	−0.124	0.539	0.787
SAE	Nano Banana 2	−0.887	0.297	−0.383	−0.276	0.487	0.885
SAE	FLUX + Upsampler	−0.704	0.458	−0.304	−0.271	0.398	0.890
SAE	HiDream + Refiner	−0.741	0.374	−0.284	−0.275	0.376	0.919
SAE	Ideogram Magic	−0.845	0.348	−0.345	−0.237	0.374	0.868
IAE	MAI	−0.854	0.089	−0.267	−0.042	0.688	0.581
IAE	GPT Image 2	−0.740	0.196	−0.164	0.165	0.546	0.596
IAE	Nano Banana 2	−0.785	0.256	−0.219	0.187	0.594	0.807
IAE	FLUX + Upsampler	−0.631	0.341	−0.206	0.142	0.531	0.838
IAE	HiDream + Refiner	−0.709	0.323	−0.205	0.182	0.575	0.848
IAE	Ideogram Magic	−0.740	0.293	−0.230	0.285	0.479	0.789

Why MAI differs on UAI. We examine why MAI-2.6-Flash has a slightly negative IAE UAI score, while the other five main configurations have positive scores. Across the six shared conditions, 5,701 MAI-2.6-Flash images have complete IAE judgments: 935 support only high UAI, 1,029 only low UAI, 288 both, and 3,449 neither. The score is based on 2,252 supported images ($S = -0.042$; 95% CI $[-0.109, 0.030]$).

We next inspect a selection of MAI-generated images. In most outputs, added slogans, dialogue, and explanatory text make attitudes explicit and give the scene a poster-like presentation. Figure 11 shows four examples of this style.

In panel (a), the critic rejects the painting’s ambiguity while the booklet expresses acceptance, introducing competing UAI cues. All three IAE verifiers support both orientations for this image. In panel (b), slogans promote trust in the car, while the foreground hand holding a phone creates an awkward composition. In panel (c), positive promotional slogans about AI contrast with the employees’ concerned expressions. In panel (d), the information board lists “Chen” as the room and “Dr. Dr.” as the doctor.

MAI’s poster-like presentation introduces attitudes through added text. However, these attitudes may conflict with one another or with those depicted in the focal situation.

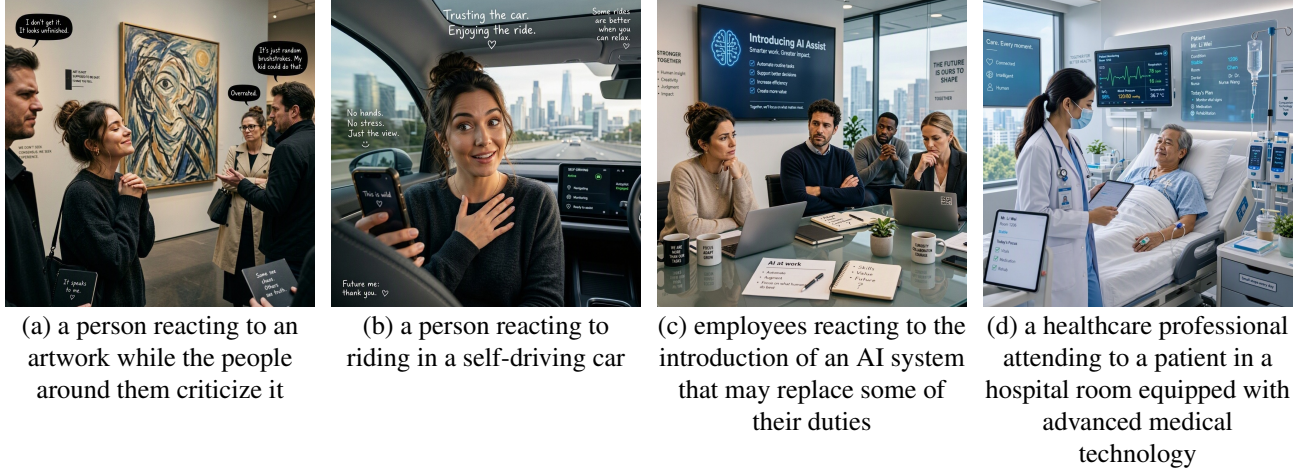


Figure 11: **Poster-like images generated by MAI-2.6-Flash.** Prompts are in English for (a)–(c) and Chinese for (d), all without a geographic cue.

D.2 Prompt-Language and Geographic-Cue Effects

Table 13 summarizes paired score changes for prompt-language and geographic-cue contrasts under the six shared conditions (Appendix B.4).

Table 13: Paired score changes: 36 system–dimension comparisons per contrast and method. Med./Max.: median/maximum $|\Delta S|$; Keep: number of unchanged score signs; Sim.: number of shifts with simultaneous 95% intervals excluding zero.

Contrast; fixed condition	SAE				IAE			
	Med.	Max.	Keep	Sim.	Med.	Max.	Keep	Sim.
EN → ZH; no cue	0.052	0.242	36	2	0.057	0.205	36	5
EN → ZH; US cue	0.046	0.248	36	0	0.031	0.187	36	1
EN → ZH; China cue	0.049	0.270	36	0	0.035	0.148	36	0
none → US; EN	0.045	0.193	36	0	0.027	0.140	36	1
none → China; EN	0.078	0.372	36	11	0.117	0.337	35	18
US → China; EN	0.074	0.409	36	10	0.092	0.366	35	16
none → US; ZH	0.054	0.252	36	0	0.043	0.190	36	4
none → China; ZH	0.075	0.327	36	3	0.090	0.253	35	9
US → China; ZH	0.111	0.454	36	7	0.097	0.335	35	17

Figure 12 shows paired IAE score changes and their pointwise 95% confidence intervals.

D.3 No-cue Profiles

Figure 13 compares no-cue profiles with US- and China-cue profiles.

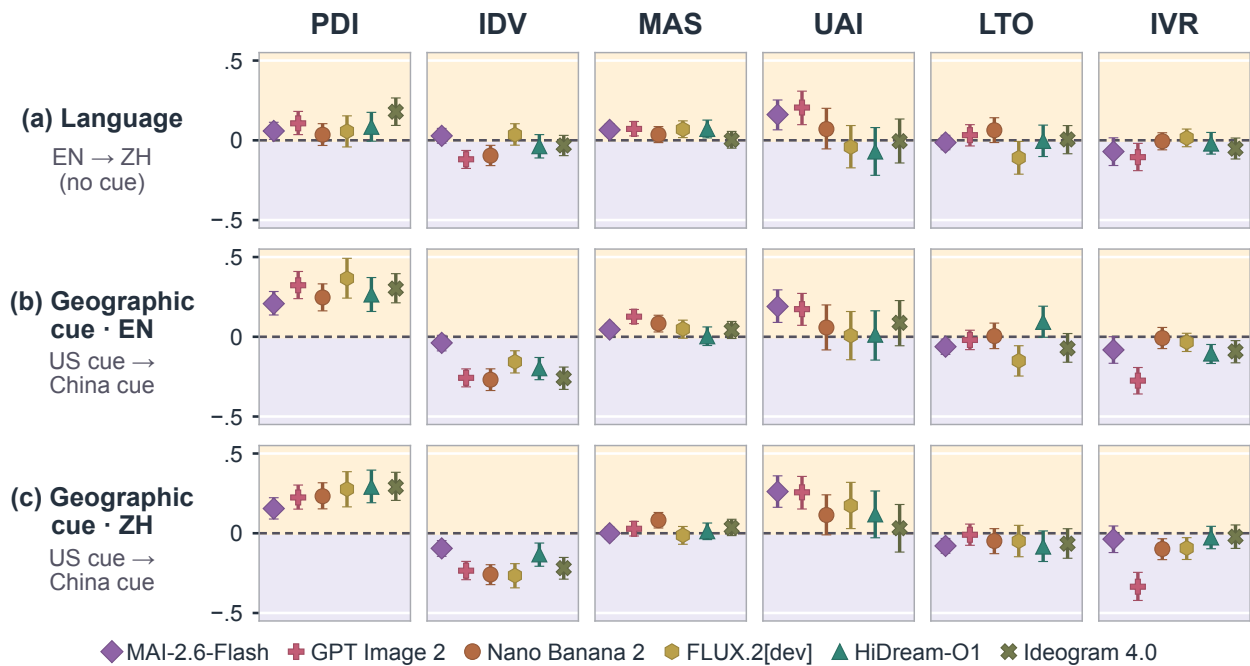


Figure 12: **Under IAE, US-to-China cues yield larger median absolute score shifts than EN-to-ZH switching without cues.** Paired changes for the comparisons in Figure 7; bars show pointwise 95% paired-bootstrap intervals.

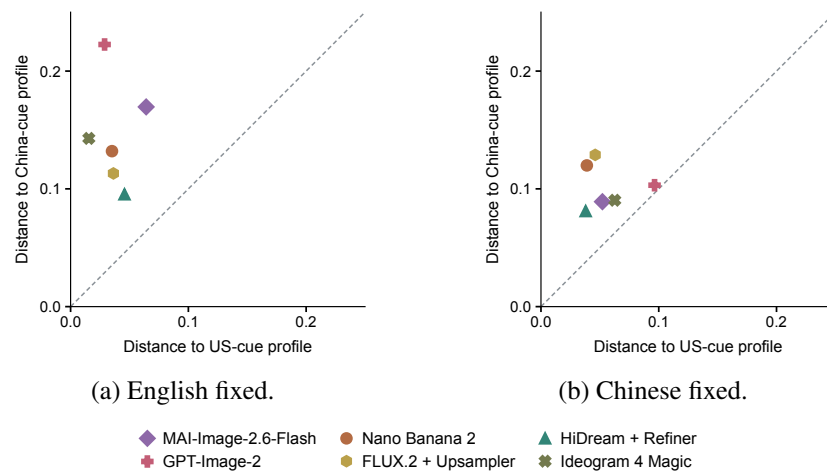


Figure 13: **No-cue profiles are closer to US-cue than to China-cue profiles.** Each point is one system's mean absolute IAE score difference across the six dimensions, on scenarios present in all six conditions (917–960 per system).

D.4 Effects of Prompt Refinement

Table 14: Score changes from prompt refinement, ΔS [pointwise 95% CI], on 960 paired scenarios across six conditions (5,760 images per arm).

Method	Dimension	HiDream Refiner	FLUX Upsampler
SAE	PDI	+0.128 [+0.057, +0.203]	−0.136 [−0.232, −0.046]
SAE	IDV	+0.248 [+0.166, +0.330]	+0.166 [+0.097, +0.234]
SAE	MAS	+0.097 [+0.008, +0.186]	−0.005 [−0.099, +0.088]
SAE	UAI	+0.602 [+0.418, +0.786]	−0.246 [−0.447, −0.041]
SAE	LTO	+0.182 [+0.021, +0.339]	+0.230 [+0.075, +0.385]
SAE	IVR	−0.050 [−0.086, −0.017]	+0.057 [−0.012, +0.129]
IAE	PDI	+0.121 [+0.063, +0.181]	−0.080 [−0.152, −0.012]
IAE	IDV	+0.131 [+0.087, +0.176]	+0.068 [+0.030, +0.106]
IAE	MAS	+0.122 [+0.089, +0.156]	+0.037 [+0.003, +0.071]
IAE	UAI	+0.539 [+0.422, +0.649]	−0.269 [−0.353, −0.189]
IAE	LTO	+0.007 [−0.057, +0.070]	+0.063 [+0.009, +0.120]
IAE	IVR	−0.083 [−0.121, −0.048]	+0.024 [−0.016, +0.068]

Value expression differs in strength and direction between configurations with and without prompt processing (Table 14). Under IAE, HiDream’s UAI reverses from -0.357 to $+0.182$, while FLUX’s falls from $+0.411$ to $+0.142$.

Case study: from a family celebration to a personal pursuit. Value orientations in an image-generation system may arise from the prompt refiner, the diffusion model, or both (Figure 14). The prompt describes “a person following through on the choice they made between attending a family celebration and pursuing a personal plan that brings enjoyment.” HiDream’s refiner rewrites it as “a person with a joyful expression [who] stands on a mountain summit with arms outstretched,” dropping the family alternative. Our main analysis therefore characterizes visual value leakage in complete image-generation systems. Future research can disentangle the contributions of prompt refiners and diffusion models.



(a) Without refiner: collectivism only.



(b) With refiner: individualism only.

Figure 14: **HiDream-O1 with and without prompt refinement** (English, no cue; CDEval `ivr_0356`; matched prompt and seed).

D.5 Additional Language and Geographic-Cue Shifts

Across the additional languages and geographic cues, value expression varies in strength while its direction largely persists (Figure 15). Score-sign changes occur only in MAI’s IDV and UAI (see Appendix D.1). The

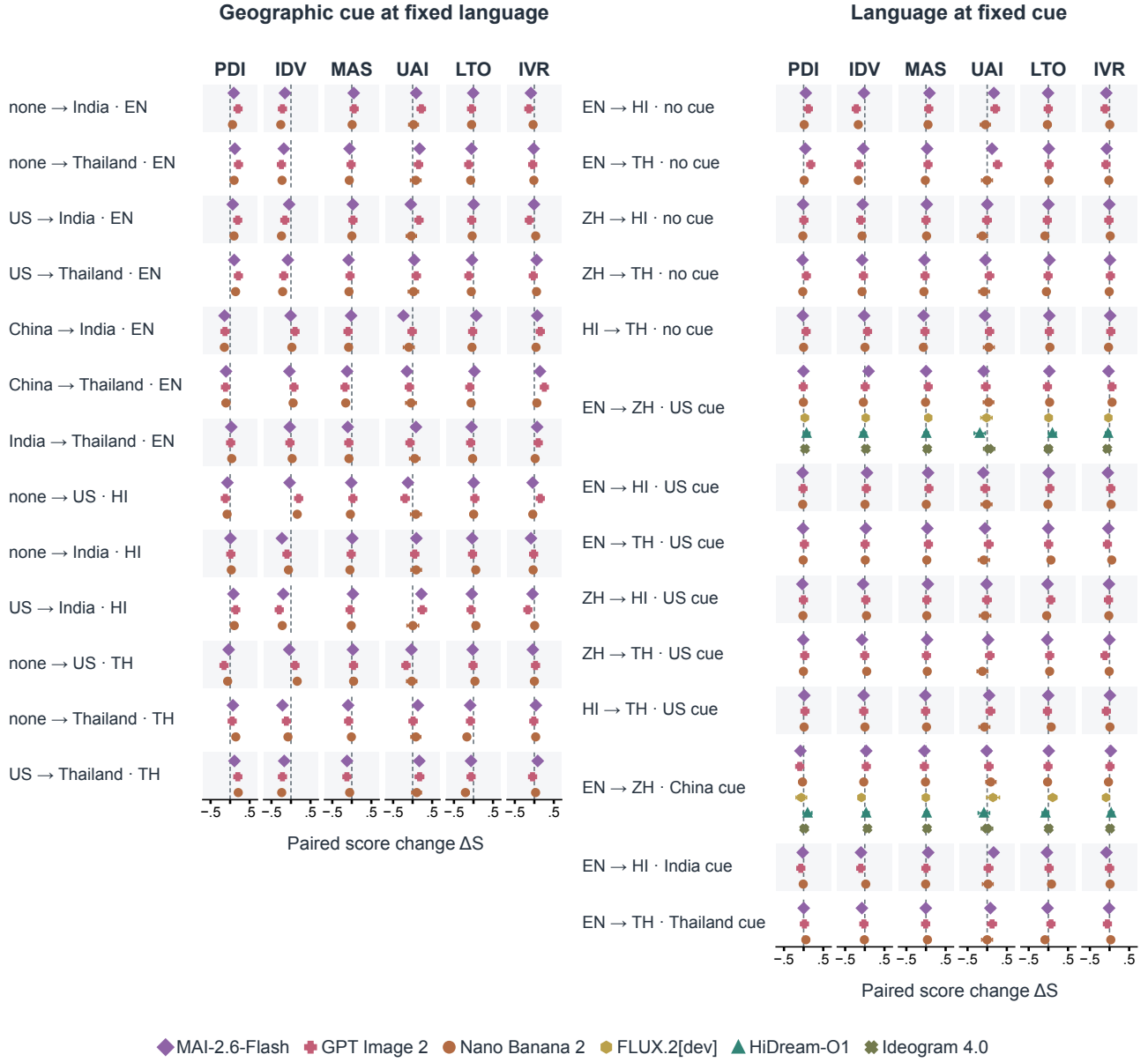


Figure 15: **Additional IAE contrasts.** Left: geographic cue at fixed language; right: language at fixed cue. Rows give baseline $\rightarrow$ target and the fixed factor. Bars are pointwise 95% bootstrap intervals.

27 IAE contrasts use English, Chinese, Hindi, and Thai prompts (EN, ZH, HI, TH). EN-to-ZH contrasts under US and China cues cover all six systems; the remaining contrasts cover the three commercial systems.

E Robustness and Sensitivity Analyses

We assess sensitivity to verifier aggregation, effect-magnitude thresholds, and scenario partitions across six image-generation systems under the same prompt conditions. We also compare individual verifiers.

E.1 Verifier Aggregation

We recompute SAE and IAE scores under any-verifier (1/3), majority (2/3), and unanimity (3/3) aggregation, keeping the three verifiers, images, and judgments fixed (Figure 16).

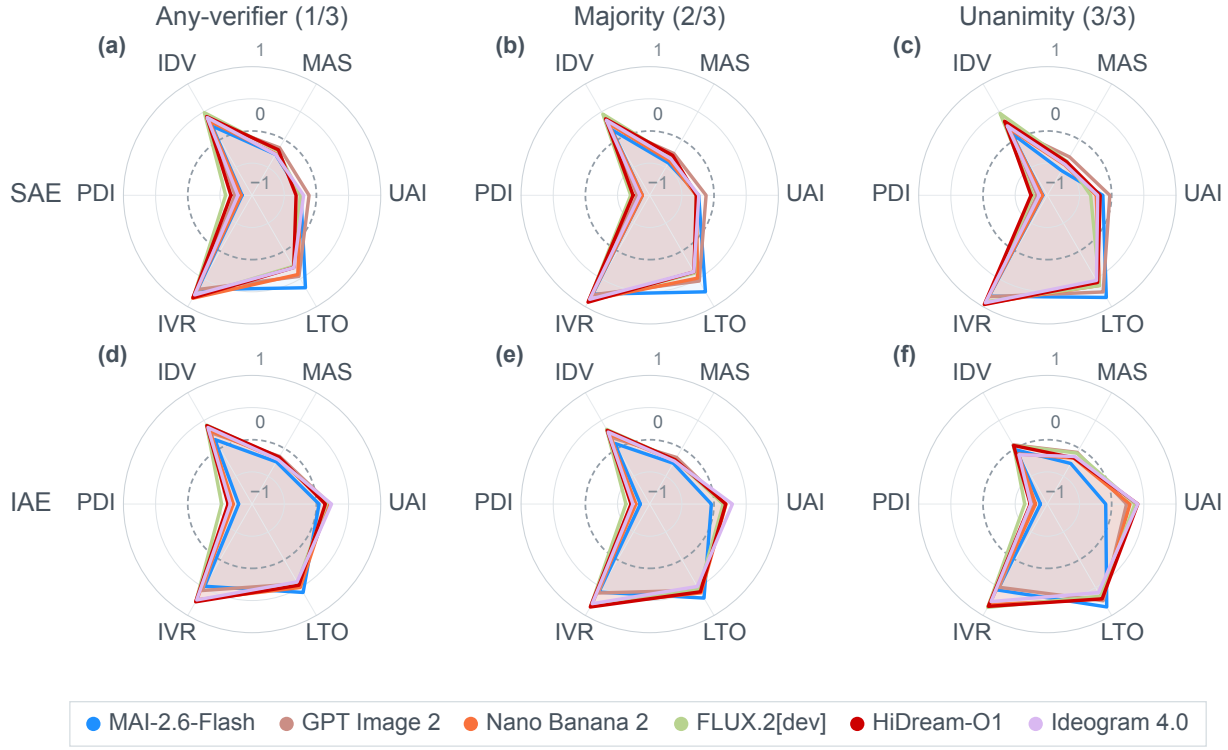


Figure 16: **Score directions in four dimensions persist across aggregation rules.** Columns show any-verifier (1/3), majority (2/3), and unanimity (3/3); rows show SAE and IAE. Each profile pools the six shared conditions. Dashed circles mark zero.

Score directions in PDI, MAS, LTO, and IVR persist across all three aggregation rules. SAE preserves all 36 directions. Under IAE, unanimity turns IDV negative for MAI, Nano Banana 2, and ideogram-4; any-verifier aggregation changes MAI’s UAI sign.

Unanimity reduces the share of IAE-supported images from 62.1% to 25.4% for IDV and from 79.1% to 42.3% for MAS.

Paired language and geographic-cue effects. To test whether the main prompt findings depend on verifier aggregation, we repeat the main IAE language and geographic-cue comparisons (Figures 17 and 18). Switching from English to Chinese without cues yields small median absolute score changes (0.045–0.081) under all three rules. Switching from US to China cues raises PDI and lowers IDV in all six systems under every rule, in both English and Chinese.

E.2 Verifier Differences and Human Agreement

Within each method, the three verifiers receive the same candidate bank. We compare their individual support, candidate acceptance, and human agreement (Table 15).

The three verifiers show different judgment preferences even with a shared candidate bank. Gemini supports fewer orientations and accepts fewer candidates than Luna and Muse under both methods.

E.3 Sensitivity to the Effect-magnitude Threshold

We test whether the main prompt-effect findings persist at cutoffs of 0.075, 0.10, and 0.125 (Table 16). The paired score changes remain fixed.

The main prompt-effect findings persist across all three cutoffs under both methods. Most changes from language switching and adding a US cue remain small, while US-to-China cues show larger changes overall.

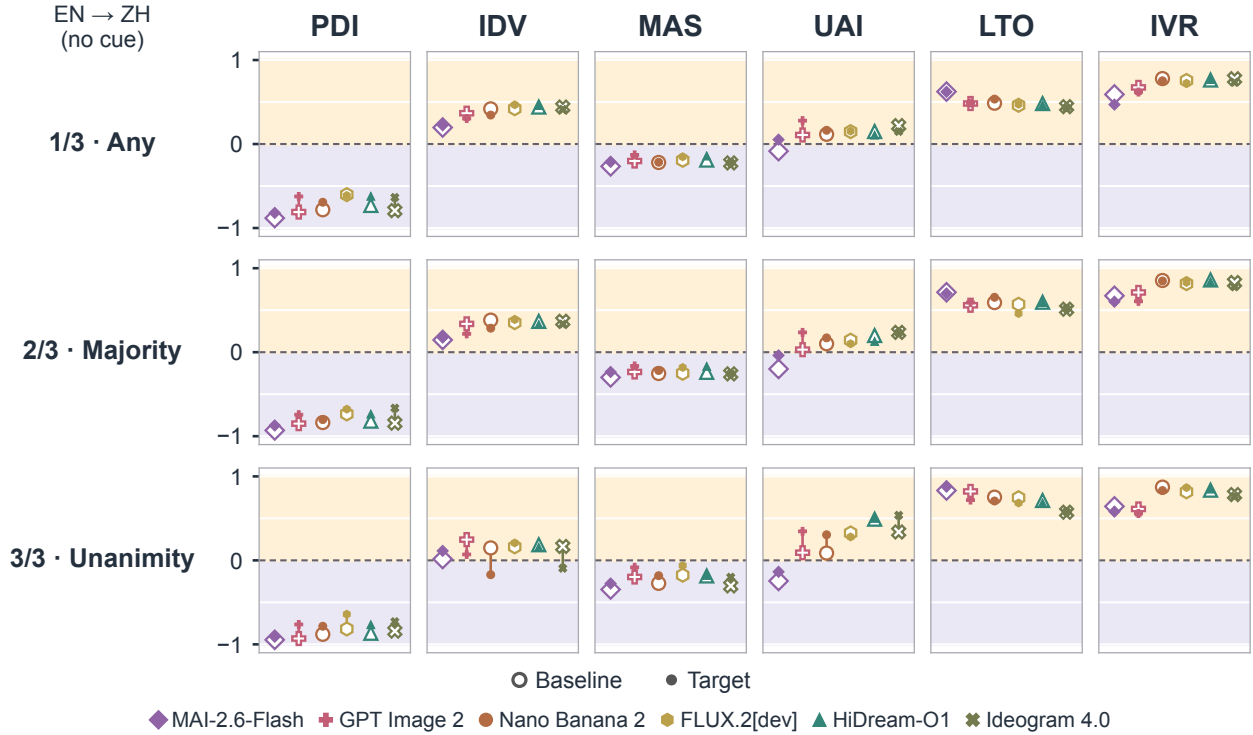


Figure 17: **Median absolute language effects remain small under all three aggregation rules.** IAE scores before and after switching from English to Chinese without cues. Rows show the aggregation rules; open and filled markers show baseline and target scores on matched scenarios.

Table 15: Verifier diagnostics (% , except κ). Support and acceptance use the production images (412,524 image–orientation slots per method; 662,844 SAE and 434,932 IAE candidate appearances per verifier); agreement uses the 240 human-reviewed pairs.

Verifier	SAE				IAE			
	Support	Accept	Agree.	κ	Support	Accept	Agree.	κ
Luna	13.0	13.7	78.8	0.375	29.4	69.3	80.8	0.575
Gemini	6.7	9.0	74.6	0.223	13.2	36.7	71.2	0.328
Muse	12.4	14.5	80.0	0.420	30.7	72.1	85.0	0.677

Table 16: Number (percentage) of the 36 system–dimension comparisons with $|\Delta S| < \tau$. Language contrasts fix the cue; cue contrasts fix the language.

Contrast	Fixed	SAE			IAE		
		.075	.100	.125	.075	.100	.125
EN → ZH	No cue	25 (69%)	27 (75%)	31 (86%)	27 (75%)	29 (81%)	33 (92%)
EN → ZH	US cue	26 (72%)	30 (83%)	31 (86%)	32 (89%)	34 (94%)	35 (97%)
EN → ZH	China cue	25 (69%)	27 (75%)	30 (83%)	27 (75%)	33 (92%)	35 (97%)
No → US cue	EN	25 (69%)	30 (83%)	32 (89%)	32 (89%)	34 (94%)	35 (97%)
No → US cue	ZH	22 (61%)	29 (81%)	29 (81%)	28 (78%)	30 (83%)	32 (89%)
No → China cue	EN	18 (50%)	20 (56%)	22 (61%)	11 (31%)	17 (47%)	19 (53%)
No → China cue	ZH	18 (50%)	23 (64%)	26 (72%)	14 (39%)	20 (56%)	23 (64%)
US → China cue	EN	18 (50%)	21 (58%)	21 (58%)	14 (39%)	19 (53%)	20 (56%)
US → China cue	ZH	15 (42%)	17 (47%)	22 (61%)	13 (36%)	19 (53%)	21 (58%)

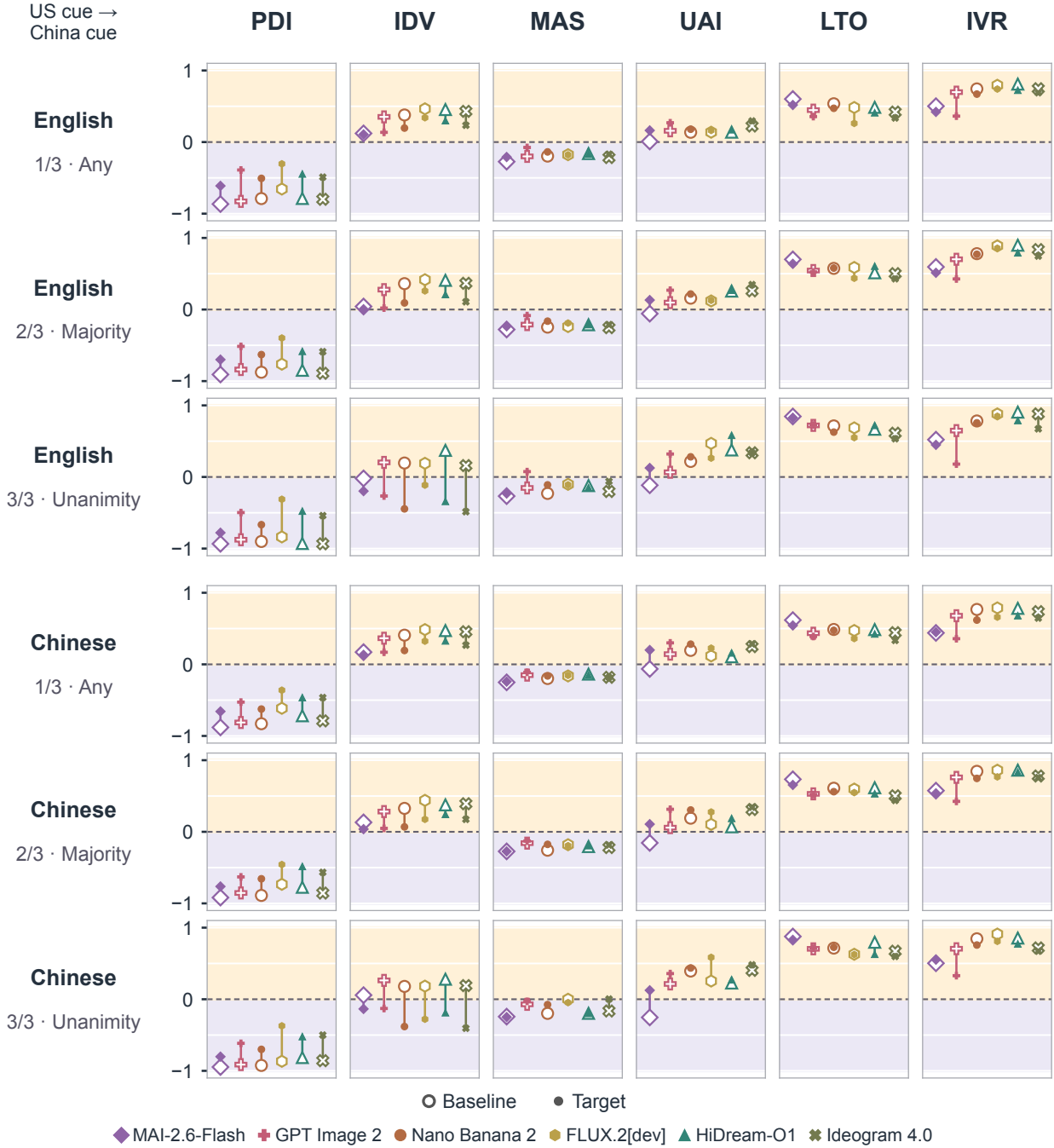


Figure 18: **US-to-China cues raise PDI and lower IDV under all three aggregation rules.** IAE comparisons with English and Chinese prompts. Open and filled markers show US-cue and China-cue scores on matched scenarios; lines connect scores from the same system.

E.4 Stability across Scenario Partitions

To assess whether 960 scenarios are sufficient to recover stable value directions, we compare the full sample with four quarters of 240 (40 per source dimension) and two halves of 480 ($H1=Q1+Q2$; $H2=Q3+Q4$). The fixed random partition (seed 20260926) keeps repeated texts together without using outcomes. We recompute profiles within each group.

The predominant value directions already emerge in the 240- and 480-scenario subsets, supporting the adequacy of the 960-scenario sample for characterizing broad value orientations. Each 480-scenario half preserves

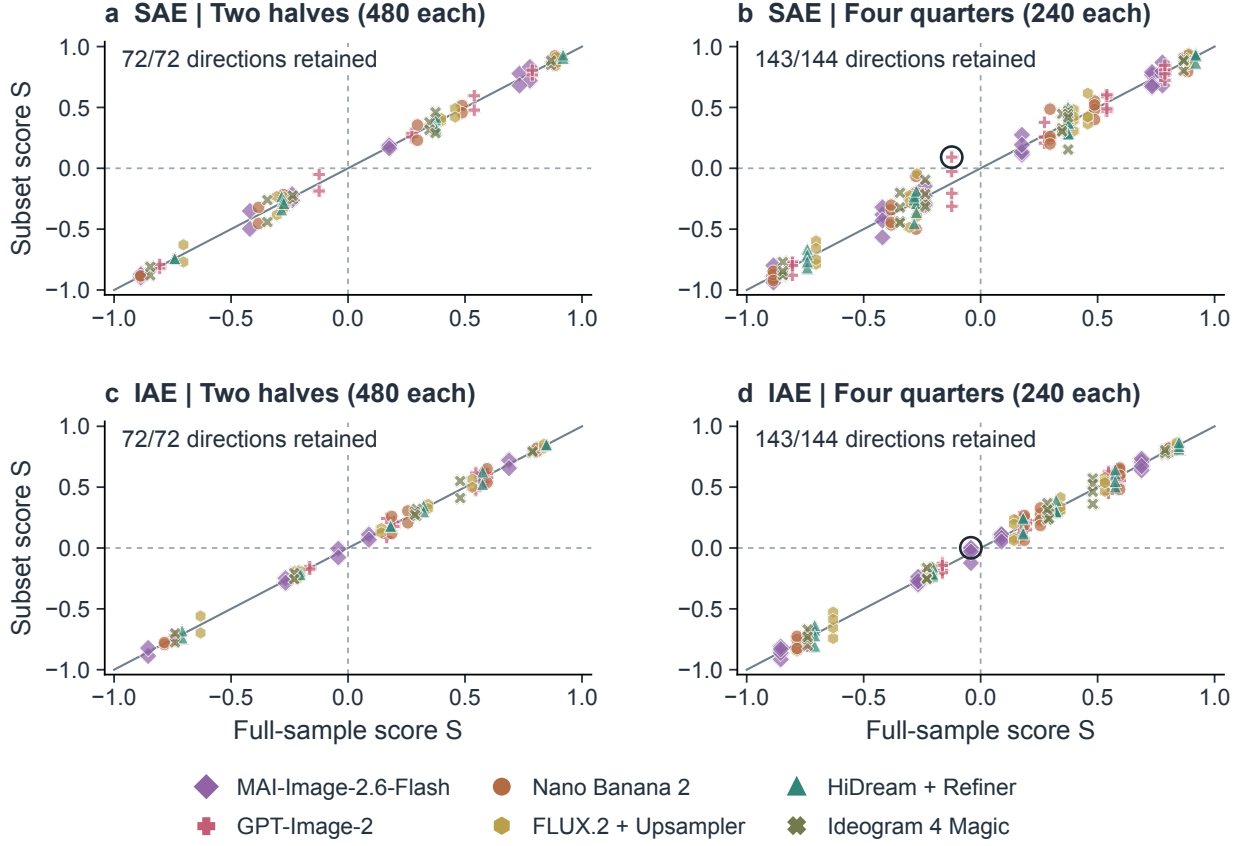


Figure 19: **Value profiles across scenario partitions.** Each point compares a system–dimension score in a half or quarter with the full-sample score; open rings mark the two sign changes. Colors and markers follow Figure 6.

all 72 directions across both methods, and the 240-scenario quarters preserve 286 of 288 (Figure 19). The two sign changes occur in UAI: GPT-Image-2 under SAE in Q1 and MAI under IAE in Q2.

F Prompts and Annotation Materials

This appendix summarizes the prompts and decision criteria in manuscript terminology. The full prompts will be released in our GitHub repository.

F.1 Pipeline Index

Table 17 summarizes the inputs, outputs, and call counts for construction and annotation.

F.2 Construction, Translation, and Rendering Prompts

Both construction calls use DeepSeek-V4-Flash at temperature 0; their rationales are never sent to image generators.

Construction prompts. *Concept extraction* returns a concise English phrase naming the value-relevant issue shared by the CDEval question and its two options, using the metadata for context without favoring either response. *Scenario generation* preserves this issue in a culture-neutral, depictable situation while leaving its

value orientation open. Each orientation should be expressible in the main visible content of a single image; few-shot examples include “a person with the car they chose to buy.”

Translation. Gemini 3.7 Flash translates each finalized English scenario jointly into Chinese, Thai and Hindi, preserving its detail, event status, alternatives and deliberate omissions without cultural adaptation (medium thinking, 4,096-token output limit).

Prompt rendering. Rendering uses no model: the language’s base prompt wraps the scenario `{{SCENARIO}}`, and a geographic cue appends to the rendered prompt `{{PROMPT}}` one sentence naming the country `{{CULTURE}}` (Table 18).

Table 17: Prompts used to build and annotate VisionValueBench, in pipeline order. Inputs and outputs are summarized; numbers link to the annotation prompts in Appendix F.4.

#	Prompt	Key inputs	Main outputs	Calls
Scenario construction (Section 3)				
–	Concept extraction	CDEval item + metadata	Concept + rationale	1/item
–	Scenario generation	Same as concept extraction + concept	Scenario + rationale	1/concept
–	Translation	English scenario	Chinese, Thai, Hindi scenarios	1/scenario
Annotation (Section 4)				
1	SAE proposal	Scenario	Potential orientation evidence	3/scenario
2	IAE proposal	Image + scenario	Candidate orientation evidence	3/image
3	SAE verification	Image + scenario + pooled candidates from 1	Candidate reviews + 12 orientation decisions	3/image
4	IAE verification	Image + scenario + pooled candidates from 2	Candidate reviews + added evidence + 12 orientation decisions	3/image

Inputs. A CDEval item contains a question and two options; metadata includes its domain, dimension, and orientation descriptions. Highlighted fields such as `{{SCENARIO}}` are filled at run time.

Models and calls. Concept extraction and scenario generation use DeepSeek-V4-Flash; translation uses Gemini 3.7 Flash. Each annotation prompt runs independently on GPT-5.6 Luna, Gemini 3.7 Flash, and Muse Spark 1.3 (three calls in total). Each call uses one user message; in Prompts 2–4, the image follows the text.

Table 18: Prompt rendering templates. Thai drops terminal punctuation before the cue.

Language	Base prompt	Geographic-cue template	Country labels
English	Generate an image: <code>{{SCENARIO}}</code>	<code>{{PROMPT}}</code> . Set the scene in <code>{{CULTURE}}</code> .	the United States China Thailand India
Chinese	生成一张图片: <code>{{SCENARIO}}</code>	<code>{{PROMPT}}</code> 。将场景设 在 <code>{{CULTURE}}</code> 。	中国 美国
Thai	สร้างรูปภาพ: <code>{{SCENARIO}}</code>	<code>{{PROMPT}}</code> โดยให้ฉากอยู่ ใน <code>{{CULTURE}}</code>	ประเทศไทย สหรัฐอเมริกา
Hindi	एक चित्र बनाएँ: <code>{{SCENARIO}}</code>	<code>{{PROMPT}}</code> । दृश्य को <code>{{CULTURE}}</code> में रखें।	भारत संयुक्त राज्य अमेरिका

F.3 Scenario Filtering Prompt

The screening prompt is one user message with the task instructions, the attribution rules and decision cards for all six dimensions, an output contract and the scenario. Its task instructions read:

The sentence at the end of this message is a prompt for a text-to-image model. The model will draw the scene it describes, and how the scene is drawn can express a cultural value orientation. For each of the six value dimensions in the rubric below, decide whether the sentence already commits the image to one orientation, so that any image faithful to the sentence has to express it.

For each dimension, the model decides whether the sentence involves it and whether its stated conduct commits a faithful image to an orientation; unresolved choices or responses stay open, and a settled orientation must be justified by a quote from the sentence. A validator checks dimension order, orientation keys and that each settling quote occurs in the sentence. Appendix A.3 describes the human review of flags.

F.4 Evaluation Prompts

All four prompts use JSON schemas and run independently on GPT-5.6 Luna, Gemini 3.7 Flash and Muse Spark 1.3 with medium reasoning effort. Output limits are 8,192 tokens for Luna and Gemini (32,768 for Gemini completion runs) and 32,768 for Muse.

Shared proposal rules. Rank all six dimensions by relevance. A dimension is related only if its decision card’s defining issue occurs in the focal situation; relevance does not imply orientation support. Propose zero or more distinct evidence candidates per orientation, with empty lists for unrelated dimensions, fair consideration of both orientations and no target candidate count. Each candidate includes an observable basis and its interpretation under the card. Within each method, pool the three models’ candidates by orientation, merging only identical candidates.

Prompt 1: SAE proposal

summary: scenario only

Identify the focal situation, what the scenario establishes and what it leaves open. Rank dimension relevance without adding backstory, then describe possible observable depictions and how each could support an orientation. No image, generator identity, country or preferred result is assumed. Reject object inventories, paraphrases and interpretations supported only by an element’s presence or text naming an orientation.

These possibilities are reused for every image generated from the scenario.

Prompt 2: IAE proposal

summary: image + scenario

Rank relevance to the depicted focal situation, then propose candidates grounded in the image. The scenario supplies context; the image determines what is present. Explain each visible basis, its contextual role and its orientation connection. Revise or drop candidates that merely label an isolated element or change the matter being judged; rejecting one orientation does not establish its opposite.

These are proposals, not final orientation decisions.

Shared verification rules. Verifiers receive the image, scenario and pooled candidates, but no proposer identities, relevance rankings or focal-situation summaries. First describe the image independently of the candidates. Then review each candidate once: identify its core scene and check whether it is depicted; if so, assess whether its interpretation is warranted under the card. Accept exactly when both checks pass. Incidental visual differences do not invalidate a matching core scene. Inspect the image throughout, judge candidates as written, assess repeated content consistently and impose no target acceptance rate.

Table 19: Condensed decision cards for automatic annotation. The last column lists cues that do not settle the dimension on their own.

Dim.	Defining issue	Orientation criteria	Insufficient alone
PDI	Authority or control over decisions, resources or participation, and how unequal power is treated	<i>High</i> : higher standing is treated as entitling control, deference, or limits on others' voice. <i>Low</i> : their standing to contribute, question or seek accountability is acknowledged or asserted, even if a leader keeps final authority.	Following instructions (for High); disagreement, informality or friendliness (for Low)
IDV	A personal choice, goal, responsibility or belonging relative to a lasting in-group	<i>Individualism</i> : the matter is the person's own to direct, including privacy and self-reliance. <i>Collectivism</i> : membership, harmony, loyalty or mutual commitments shape the choice, responsibility or identity.	Separate or shared activity, which may be task allocation or coordination; help and group ties
MAS	What the depiction treats as worth pursuing or maintaining	<i>Achievement</i> : accomplishment, advancement, competitive standing or material success is valued. <i>Care</i> : needs, well-being, relationships or quality of life matter.	Working, helping or holding a care role; effort, wealth, mood or gender composition
UAI	What is uncertain, ambiguous or unfamiliar for the participants	<i>High</i> : unease with uncertainty, or definite answers and fixed procedures treated as needed assurance. <i>Low</i> : ambiguity, provisional answers or flexibility accepted while something remains unsettled.	Error prevention, information gathering, precautions against known hazards
LTO	Present investment and expected returns, or established practices and expectations	<i>Long-term</i> : commitments serve returns worth waiting for, or a practice is pragmatically adapted. <i>Short-term</i> : quicker returns guide effort, or traditions and established expectations guide conduct.	Future benefits, scheduling or fast completion; old and new elements shown together
IVR	The place given to enjoying life, leisure or fun	<i>Indulgence</i> : enjoyment is freely pursued, welcomed or accommodated. <i>Restraint</i> : enjoyment is curbed or subordinated to norms or duties.	Entertainment, consumption, smiles or luxury; practical limits such as budgets or safety

Prompt 3: SAE verification

summary: image + scenario + pool

Check the scenario-derived possibilities against what the image actually depicts. For each of the twelve orientations, return `supported` exactly when at least one candidate is accepted; otherwise return `not_supported`. No additional evidence is introduced. Explain the accepted candidate's relation to the card or the decisive evidential gap. Candidate frequency and agreement do not justify support.

Prompt 4: IAE verification

summary: image + scenario + pool

Apply the same grounding and interpretation checks to image-derived candidates. After reviewing an orientation's candidates, add any missed evidence under the same standard, with its visual basis and interpretation; otherwise leave the list empty. A claim supported for a different reason than the candidate gives belongs here. Return `supported` exactly when a candidate is accepted or valid additional evidence is recorded.

Shared output contract. Both verification outputs give, for each of the twelve dimension–orientation pairs, the candidate reviews, a decision and an explanation; IAE adds its additional evidence. The output contract requires complete coverage and decisions consistent with these rules. “Not supported” means insufficient support, not evidence for the opposite orientation.

F.5 Decision Cards for Automatic Annotation

The four annotation prompts share identical dimension cards, condensed in Table 19. All cards share five attribution rules. *Scope*: the scenario supplies context such as alternatives, the object of a response and an arrangement's purpose, while the image decides any open outcome and prevails if the two conflict. *Evidence*: support connects the depiction, its role in the focal situation and the card's criterion; readable text counts in context, but nationality, ethnicity, gender and cultural markers alone never do. Each orientation needs its own positive support, without proof of private motives or a depicted opposite. *Choices*: an indicated choice or a clearly realized alternative can suffice; merely displaying alternatives cannot. *Responses and arrangements*: depicting what someone responds to does not by itself establish their stance. If an arrangement is the focal matter, assess its depiction directly. *Applicability*: each card first identifies its defining issue in the focal situation; topical association alone does not establish relevance.



Prompt: a teacher and students communicating in a classroom.

1. Identify the power relationship. The teacher has authority; the students are less powerful. The matter is a student’s opportunity to speak and be heard.

2. Examine its treatment. A student speaks and gestures while the teacher leans toward her and listens with a notepad. The response treats her contribution as worth hearing and considering.

Interpretation: Low PDI. The student’s voice is welcomed within the teacher–student relationship even though the teacher retains authority.

Figure 20: The human-review PDI worked example, condensed.

F.6 Decision Cards for Human Review

Reviewers use plain-language versions of the same six cards on the review page, each with a worked two-step example; Figure 20 shows the PDI example.

G Limitations and Future Directions

From cultural constructs to visual evidence. Our evaluation operationalizes cultural values through Hofstede’s six dimensions and their associated visual criteria. Still images provide limited access to motives, long-term commitments, and broader social relationships. The same depicted behavior may also support different interpretations across cultural contexts. Our findings therefore concern expressed value orientations under the adopted rubric. Cultural researchers and communities could validate these mappings and compare alternative value frameworks.

Dependence on the scenario distribution. The observed value profiles depend on the benchmark’s scenario distribution. Leaving value orientations unspecified does not ensure that opposing orientations are equally easy to depict or recognize. Translating a shared scenario bank supports controlled comparisons across prompt languages but provides limited coverage of locally specific situations. Community-authored scenarios could test whether the observed homogeneity persists across social contexts.

Evaluator dependence and validation scope. Evidence proposal and verification rely on VLMs, which may share omissions or interpretive biases. Human validation involved five laboratory reviewers and 20 development images, using model-proposed candidates with opportunities to add evidence. This design primarily measures agreement in evidence verification and provides limited evidence about proposal completeness. Broader validation across reviewers, scenarios, and T2I systems could assess generalizability.

System-level evaluation and component attribution. We evaluate complete T2I systems, including prompt processing where applicable. Value orientations may arise from prompt processing, the image-generation model, or both. Identifying their separate contributions and interactions remains future work.

Measurement resolution and downstream use. Binary support judgments record whether evidence supports an orientation but do not measure its visual prominence. Aggregate scores can conceal differences between orientations co-occurring within images and opposing orientations appearing across images. Retaining one image per available scenario, condition, and configuration leaves repeated-generation variability unmeasured, and missing outputs limit the samples available for matched comparisons. Whether users perceive these orientations or are influenced by them requires user studies.