# BEYOND PERSON-CUE SUPPRESSION: THREE-VIEW MATCHED CONTRASTIVE CORRECTION FOR VISUAL CULTURAL ATTRIBUTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal large language models (MLLMs) can assign different cultural origins to the same content when depicted with different people. However, suppressing person-related cues alone does not ensure correct attribution: it may obscure relevant cultural evidence while leaving attribution errors unresolved. To address this limitation, we propose Three-View Matched Contrastive Correction (TVMCC), which uses the visual intervention itself to construct a reference for correcting the suppressed prediction. Task-aware masking produces a person-suppressed view and a complementary person-focused view from the same image while seeking to preserve the queried content. By contrasting candidate preferences in the original and person-focused views, TVMCC revises the suppressed ranking using a reference matched to the current image and intervention, without additional training or test labels. Across six MLLMs on Cultural Counterfactuals and MixCuBe, TVMCC outperforms original-image predictions in all 18 model–task settings, achieving the highest reported accuracy in 17. Nationality attribution gains on Cultural Counterfactuals range from 10.4 to 16.9 percentage points, exceeding person suppression on all six models. Reference shuffling on Cultural Counterfactuals reduces accuracy even with matched correction magnitudes, supporting the contribution of same-image correspondence beyond correction strength[1].

## 1 INTRODUCTION

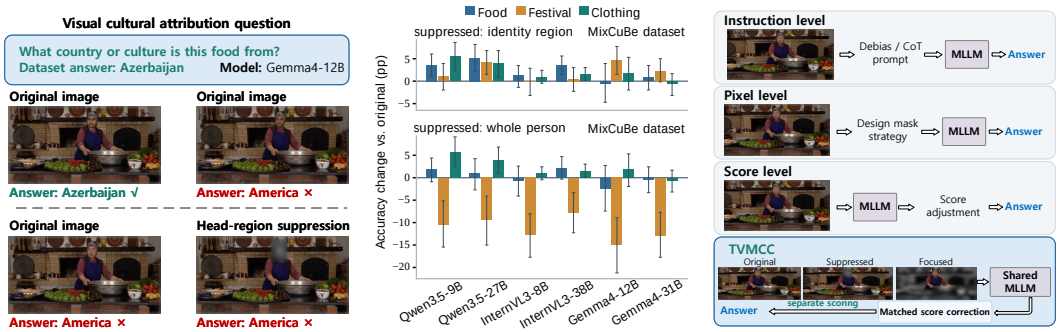
Visual cultural attribution concerns identifying the cultural origin or context of queried content in an image, including food, clothing, festivals, and cultural sites (Yayavaram et al., 2026). This task forms part of broader research on culturally aware language and multimodal models (Pawar et al., 2025), with applications in cross-cultural visual question answering and information retrieval (Nayak et al., 2024; Bhatia et al., 2024). Cultural content often appears through human activities: clothing is worn, food is eaten, and festivals involve participants. People therefore appear alongside, and sometimes overlap with, the content being queried. Reliable attribution requires multimodal large language models (MLLMs) to distinguish evidence about that content from potentially misleading cues associated with the depicted person. Yet counterfactual evaluations show that changing the depicted person can alter MLLMs' cultural attribution of the same item (Kim et al., 2025).

Reducing interference from person-related cues is important for reliable cultural attribution. Figure 1c summarizes related approaches at the instruction, pixel, and score levels alongside TVMCC. Debiasing instructions ask models to avoid demographic stereotypes (Girrbach et al., 2025), while interleaved-modal chain-of-thought combines visual and textual reasoning (Gao et al., 2025). At the pixel level, masking and blurring have been used to examine sensitivity to spurious visual features (Hirota et al., 2025). At the score level, predictions are adjusted using reference predictions or complementary visual evidence to reduce bias or improve visual grounding (Leng et al., 2024; Niu et al., 2021; An et al., 2025; Chen et al., 2026a). However, addressing person-related interference in cultural attribution requires more than cue suppression: relevant cultural evidence must be preserved, and attribution errors that persist after suppression must be corrected.

[1]The anonymous code is available at this link.



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




(a) Person-sensitive attribution (b) Suppressed region effects (c) Attribution correction strategies

Figure 1: **(a) Person-sensitive attribution.** Gemma4-12B predictions vary across person edits; head suppression leaves an error. **(b) Suppression effects.** Accuracy changes from originals; six MLLMs, 95% group-bootstrap intervals. Identity-region suppression changes little (top); whole-person suppression lowers festival accuracy on every model (bottom). **(c) Correction strategies.** Instruction-, pixel-, and score-level approaches alongside TVMCC's three-view matched correction.

Figure 1a illustrates a limitation of person-cue suppression: Gemma4-12B assigns different cultural origins to person-edited versions of the same dish, and head-region suppression leaves an incorrect answer unchanged. Figure 1b examines suppression across six MLLMs and three content categories on MixCuBe. Identity-region suppression has limited and variable effects across models and categories, whereas whole-person suppression lowers festival attribution accuracy relative to original-image predictions on all six models. These results show that broader suppression does not necessarily improve attribution and may also disrupt relevant cultural evidence. The challenge is therefore to preserve useful content while addressing errors that survive suppression, motivating a correction to the suppressed prediction rather than relying on suppression alone.

The visual intervention itself provides a reference for correcting the suppressed prediction. The mask that suppresses selected person cues also defines a complementary person-focused view, retaining those regions while attenuating their surroundings. Comparing the original and person-focused views reveals how candidate preferences differ between the full image and the retained person regions. This contrast can refine the suppressed ranking, favoring candidates supported more strongly in the original view than in the person-focused view. Deriving the reference from the same image and mask matches the adjustment to the current input and intervention. This motivates examining whether same-image correspondence improves correction beyond its magnitude alone.

Motivated by this, we propose Three-View Matched Contrastive Correction (TVMCC). Task-aware masking constructs person-suppressed and complementary person-focused views from the same image while seeking to preserve the queried content. A frozen MLLM scores all three views using the same question and candidate order. TVMCC contrasts candidate preferences in the original and person-focused views to revise the suppressed ranking, matching the correction to the current image and intervention without additional training or test labels. Across six MLLMs from three families on Cultural Counterfactuals and MixCuBe, TVMCC improves over original-image predictions in all 18 model–task settings and achieves the highest accuracy among evaluated methods in 17. Reference shuffling on CC reduces accuracy even with matched correction magnitudes, supporting the contribution of same-image correspondence beyond correction strength.

In summary, our main contributions are as follows:

- **Task-aware complementary views.** We construct person-suppressed and person-focused views through masking, targeting person-related cues while seeking to preserve cultural content.
- **Matched contrastive correction.** Our correction contrasts candidate preferences in the original and person-focused views to refine suppressed predictions, without training or test labels.
- **Cross-model effectiveness.** Across six MLLMs from three model families on Cultural Counterfactuals and MixCuBe, TVMCC achieves the highest accuracy in 17 of 18 model–task settings; controlled substitutions support the benefit of same-image matching beyond correction magnitude.

## 2 PRELIMINARIES AND PROBLEM FORMALIZATION

**Visual cultural attribution.** Given an image $x \in \mathbb{R}^{H\times W\times 3}$, a question $q$, and an ordered candidate set $\mathcal{C} = \{c_1, \ldots, c_K\}$, the task is to identify the cultural origin or context of the queried content. The dimensions $H$ and $W$ denote image height and width, and $K$ is the number of candidate answers. The reference answer $y \in \mathcal{C}$ is supplied by the benchmark and used exclusively for evaluation. The queried content may be an object, an activity, or a cultural scene. Reference labels specify the benchmark attribution without asserting a unique cultural origin. We study inference-time attribution with a frozen multimodal large language model (MLLM). Interventions modify visual inputs and candidate scores while leaving model parameters unchanged.

**Candidate scores.** Let $f_\theta$ denote the MLLM with fixed parameters $\theta$, and let $v$ be an image view derived from $x$. For candidate $c$, we extract the next-token logit $z_\theta(c; v, q)$ associated with its first answer token. We normalize over the same candidates for every view:

$$\ell_v(c) = z_\theta(c; v, q) - \log \sum_{c'\in\mathcal{C}} \exp z_\theta(c'; v, q). \tag{1}$$

The index $c'$ ranges over candidates in $\mathcal{C}$. Candidate first tokens are distinct in the evaluated configurations. The resulting vector $\ell_v \in \mathbb{R}^K$ contains closed-set log-probabilities rather than full-answer sequence likelihoods. The question, candidate order, and scoring procedure are fixed across views.

**Correction setting.** The original-image prediction is $\hat{y}o = \arg\max c \in \mathcal{C}\ell_x(c)$. The objective is to revise it without updating $\theta$ or accessing test answers. The question and category may guide view construction, but the correct label is unavailable. Correctness is assessed against $y$. Agreement across person-edited versions is evaluated separately, since consistent predictions can still be wrong.

## 3 THREE-VIEW MATCHED CONTRASTIVE CORRECTION

TVMCC revises person-suppressed predictions using a complementary reference from the same image, contrasting original and person-focused candidate preferences to correct residual errors. Figure 2 summarizes three stages: task-aware complementary view construction, matched reference construction, and contrastive score correction.

### 3.1 TASK-AWARE COMPLEMENTARY VIEW CONSTRUCTION

**Task-aware region selection.** Cultural evidence can appear in the surrounding scene, in objects held by a person, or in the clothing being worn. The suppression region is therefore selected according to the queried content. We denote the original image by $x^o = x$ and the soft suppression mask by $m \in [0, 1]^{H\times W}$, with larger values indicating stronger replacement.

For Cultural Counterfactuals (CC), the mask covers the main subject, leaving background people unmasked. For MixCuBe food questions, grounded food regions are excluded from the person mask, while heads remain included. Festival and Clothes use head suppression, retaining clothing, objects, and surrounding context. A target-protected person mask for Festival serves as a construction control. These rules seek to preserve queried content while attenuating person-related cues.

The primary mask pipeline combines frozen Qwen3.8-27B grounding with DeepLabV3-MobileNetV3 person segmentation. Grounded person boxes constrain the segmented regions, with a box fallback when segmentation coverage is insufficient. Head inclusion and target protection follow the task-specific rules, and the resulting mask boundary is smoothed. The grounder receives the image and localization instructions, including category descriptions where needed, but no cultural answer candidates, reference labels, or attribution scores. Appendix B provides the localization prompts, geometric parameters, and category-specific configurations.

**Complementary view construction.** Let $L(x^o)$ denote the three-channel image obtained by grayscale conversion and Gaussian blurring. The replacement operator is

$$B(x^o, m) = (1 - m) \odot x^o + m \odot L(x^o), \tag{2}$$

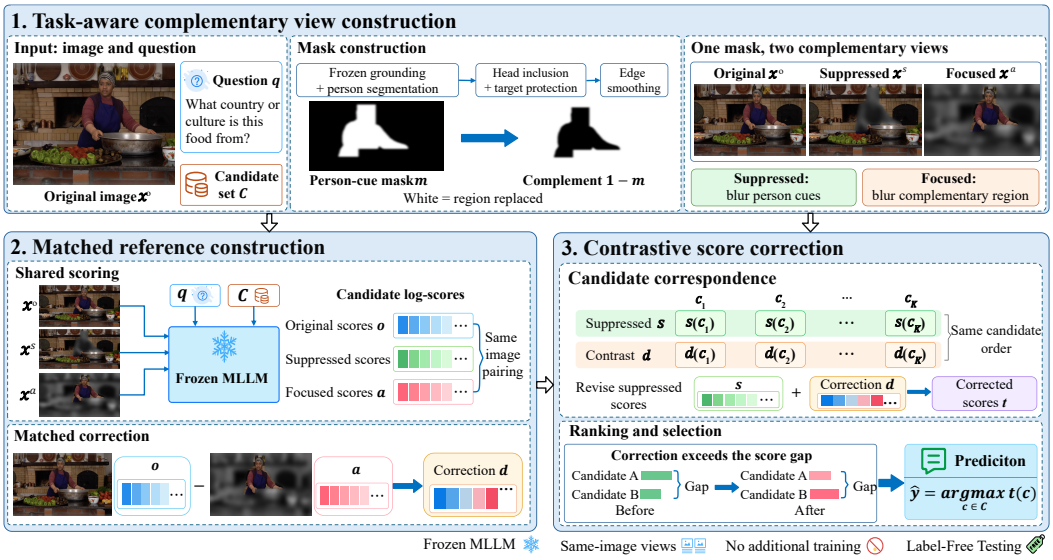



Figure 2: **Overview of TVMCC.** Task-aware masking constructs person-suppressed and person-focused views using the same mask and replacement operator. A shared frozen MLLM scores both views and the original image under the same question and candidate set. The contrast between original and focused candidate preferences supplies a matched correction to the suppressed ranking.

where $\odot$ denotes elementwise multiplication, with the mask broadcast across color channels. Pixels are unchanged where $m = 0$ and replaced by their low-frequency counterparts where $m = 1$; intermediate values provide a smooth transition.

Applying the operator with complementary masks gives

$$x^s = B(x^o, m), \qquad x^a = B(x^o, 1 - m). \tag{3}$$

The person-suppressed view $x^s$ attenuates detail in the regions. The person-focused view $x^a$ retains detail while attenuating the surroundings. For head-region interventions, the focused view is head-focused. In Figure 2, white denotes the region receiving replacement for each mask.

The two views share the image, mask geometry, and replacement operator. Their spatial correspondence ties the focused reference to the intervention producing the suppressed prediction. Because replacement retains low-frequency information, the views provide complementary visual conditions rather than strictly separating person and cultural evidence.

### 3.2 Matched Reference Construction

**Shared candidate scoring.** We apply the candidate readout defined in Eq. equation 1 to each view separately, under the shared question and fixed candidate order:

$$o = \ell_{x^o}, \qquad s = \ell_{x^s}, \qquad a = \ell_{x^a}. \tag{4}$$

The vectors $o, s, a \in \mathbb{R}^K$ contain original, suppressed, and focused candidate log-scores. All evaluations share the question $q$, ordered candidate set $\mathcal{C}$, and scoring procedure, so corresponding entries represent the same answer under different visual inputs. Appendix A provides model-specific inference settings.

**Matched candidate contrast.** The original image retains evidence that suppression may attenuate, but its scores can also favor person-associated answers. The focused view provides a reference for comparing these preferences when detail in the selected regions is preserved and surrounding detail is reduced. We define the correction vector $d \in \mathbb{R}^K$ by

$$d(c) = o(c) - a(c), \qquad c \in \mathcal{C}. \tag{5}$$

The contrast is interpreted through candidate-relative support. For an image view $v$, define the preference margin between distinct candidates $c, c' \in \mathcal{C}$ as $\Delta_v(c, c') = \ell_v(c) - \ell_v(c')$. The relative correction $d(c) - d(c')$ then equals $\Delta_{x^o}(c, c') - \Delta_{x^a}(c, c')$. Thus, the contrast favors candidates whose preference over an alternative is stronger under the complete image than the focused view.

The term *matched* refers to the reference's correspondence with the current image and intervention: $a$ is obtained from that image using the complement of the mask used to obtain $s$. This makes the comparison specific to the visual regions modified for the current prediction. Substituting another image's focused scores or an average reference changes the candidate-relative adjustment. Section 4.5 evaluates same-image correspondence through reference shuffling and correction-norm matching; Appendix D provides the control definitions and additional reference substitutions.

### 3.3 Contrastive score correction

**Corrected attribution.** The suppressed scores establish an initial ranking after selected person cues have been attenuated. TVMCC revises this ranking with the matched contrast between original and focused preferences for each candidate. Let $t \in \mathbb{R}^K$ denote the corrected score vector:

$$t(c) = \ell_{x^s}(c) + \big[\ell_{x^o}(c) - \ell_{x^a}(c)\big], \qquad \hat{y} = \arg\max_{c \in \mathcal{C}} t(c). \tag{6}$$

Here, $\hat{y}$ denotes the final attribution and $t$ the ranking scores; coefficients remain fixed across images, tasks, and models.

The suppressed view anchors the decision, while the original and focused views jointly determine its correction. Original-image scores retain the model's response to the complete visual input; the focused reference adjusts their contribution according to candidate preferences under the complementary condition. This differs from combining original and suppressed scores alone, which omits the reference, and from positive pooling, which adds support from all three views. These alternatives are evaluated using shared view scores in Appendix D.

The suppression baseline predicts $\hat{y}_s = \arg\max_{c \in \mathcal{C}} s(c)$. Comparing $\hat{y}$ with $\hat{y}_s$ under identical masks isolates the additional effect of score correction. Appendix D reports mask-construction comparisons separately from scoring-rule comparisons.

**Candidate-margin interpretation.** The corrected preference between two candidates decomposes into the suppressed margin and the matched adjustment:

$$t(c) - t(c') = \Delta_{x^s}(c, c') + \underbrace{\Delta_{x^o}(c, c') - \Delta_{x^a}(c, c')}_{\text{matched margin correction}}. \tag{7}$$

If the original and focused margins agree, the correction cancels and the suppressed preference is preserved. Otherwise, the ranking shifts toward the candidate whose relative support increases under the complete image. A competing answer overtakes the suppression-selected answer when this adjustment exceeds the existing score deficit, as illustrated in Figure 2. Candidate-independent normalization constants cancel from these margin comparisons.

Reference shuffling can alter both the candidate-relative adjustment and its magnitude. The norm-matched control in Section 4.5 rescales each centered shuffled correction to its same-image counterpart's norm, assessing correspondence beyond correction strength. Appendix D specifies the centering, rescaling, and reference-assignment procedures.

### 3.4 Inference and computational scope

Inference constructs the complementary views, obtains three candidate-score vectors, and applies Eq. equation 6. All models remain frozen, and neither view construction nor score correction uses test answers. The final prediction uses the corrected argmax without a confidence-based fallback. Computation includes grounding, segmentation, view construction, and three MLLM scoring passes; closed-set attribution does not require generating a reasoning trace. Masks can be cached and reused across attribution models for the same image and intervention specification. Appendix G reports component timings and composed cost estimates. Section 4.3 evaluates an alternative segmenter-only pipeline on both CC tasks to assess sensitivity to mask construction.

Table 1: Cultural attribution accuracy (%) across six MLLMs. All methods use the same evaluation images, candidate sets, and closed-set answer readout. † denotes our adaptations of existing methods. **Bold** marks the highest reported accuracy in each row.

| Model | Original | Debias Prompt | CoT | CF-VQA† | VCD† | AGLA† | OA-VCD† | **TVMCC** |
|---|---|---|---|---|---|---|---|---|
| *Cultural Counterfactuals — nationality (21,352 images)* | | | | | | | | |
| Qwen3.5-9B | 64.47 | 65.94 | 62.28 | 63.38 | 65.65 | 62.70 | 72.04 | **79.06** |
| Qwen3.5-27B | 62.30 | 68.97 | 58.57 | 62.85 | 62.71 | 60.53 | 70.31 | **79.11** |
| InternVL3-8B | 61.75 | 64.04 | 57.13 | 60.85 | 61.90 | 59.62 | 69.76 | **76.74** |
| InternVL3-38B | 65.72 | 70.74 | 60.91 | 66.34 | 66.63 | 63.14 | 74.50 | **82.63** |
| Gemma4-12B | 53.12 | 55.72 | 49.52 | 54.05 | 53.63 | 52.28 | 56.77 | **64.46** |
| Gemma4-31B | 71.91 | 75.92 | 69.68 | 72.25 | 72.34 | 71.38 | 75.12 | **82.27** |
| *Cultural Counterfactuals — religion (30,978 images)* | | | | | | | | |
| Qwen3.5-9B | 78.57 | 78.69 | 78.90 | 78.82 | 79.27 | 75.02 | 81.51 | **83.08** |
| Qwen3.5-27B | 78.02 | 78.64 | 76.66 | 78.26 | 78.55 | 74.99 | 81.68 | **83.10** |
| InternVL3-8B | 78.41 | 78.52 | 76.90 | 78.07 | 78.70 | 76.19 | 80.27 | **80.81** |
| InternVL3-38B | 80.49 | 81.67 | 78.64 | 81.51 | 81.06 | 77.84 | 83.17 | **83.18** |
| Gemma4-12B | 69.71 | 70.64 | 69.45 | 69.97 | 69.86 | 68.90 | 71.28 | **75.13** |
| Gemma4-31B | 78.52 | 79.01 | 79.04 | 80.36 | 79.67 | 77.60 | 80.43 | **83.65** |
| *MixCuBe — person-edited versions (1,980 images)* | | | | | | | | |
| Qwen3.5-9B | 88.18 | 86.97 | 81.16 | 90.51 | 90.56 | 86.87 | 90.91 | **92.58** |
| Qwen3.5-27B | 87.63 | 88.89 | 78.69 | 88.38 | 88.03 | 86.97 | 89.29 | **91.67** |
| InternVL3-8B | 89.75 | 89.24 | 82.98 | 89.80 | 89.95 | 88.89 | **90.91** | 90.56 |
| InternVL3-38B | 91.57 | 91.91 | 90.81 | 93.09 | 92.33 | 90.22 | 93.09 | **93.89** |
| Gemma4-12B | 77.17 | 65.20 | 74.19 | 72.07 | 74.70 | 75.61 | 79.04 | **81.77** |
| Gemma4-31B | 92.37 | 93.38 | 91.31 | 92.47 | 92.83 | 92.68 | 93.43 | **94.44** |

# 4 EXPERIMENTS

## 4.1 SETUP

**Datasets.** We evaluate TVMCC on Cultural Counterfactuals (CC; Howard et al., 2026) and MixCuBe (Kim et al., 2025). CC covers country and religious-place attribution; MixCuBe covers food, festivals, and clothing with four person-edited versions per item. Dataset details are in Appendix A.

**Experimental settings.** We evaluate six frozen MLLMs: Qwen3.5-9B/27B (Qwen Team, 2026), InternVL3-8B/38B (Zhu et al., 2025), and Gemma4-12B/31B (Abd et al., 2026). Experiments use bfloat16 inference on a server with eight NVIDIA A100 GPUs (80 GB each). All views use the same task prompt, candidate order, and first-token scoring, with log probabilities normalized over the candidate set. Masks combine Qwen3.8-27B localization with DeepLab segmentation (Chen et al., 2017) and task-aware region selection on MixCuBe. TVMCC applies the fixed scoring rule without additional training, test-label-based tuning, or confidence fallback. Following prior work (Howard et al., 2026; Kim et al., 2025), we report cultural attribution accuracy (%) against dataset reference labels, using the same images and candidate sets across methods. Paired differences are reported with 95% bootstrap confidence intervals, resampling matched groups rather than individual images. Implementation and statistical details appear in Appendix A.

**Baselines.** We compare TVMCC with original-image scoring and person suppression using the same masks as TVMCC. Debias Prompt explicitly instructs the model to disregard person-related cues, while CoT asks it to reason about visible cultural evidence before answering; both prompt templates are provided in Appendix A.1. We also evaluate four score-level baselines: CF-VQA (Niu et al., 2021) discounts language-driven priors using a blank-image reference; VCD (Leng et al., 2024) contrasts original and noise-corrupted inputs to reduce reliance on visual priors; AGLA (An et al., 2025) combines global and local visual evidence; and OA-VCD (Chen et al., 2026a) uses an attention-guided reference for contrastive correction. These methods are adapted to our closed-set scoring protocol, with implementation details in Appendix C.

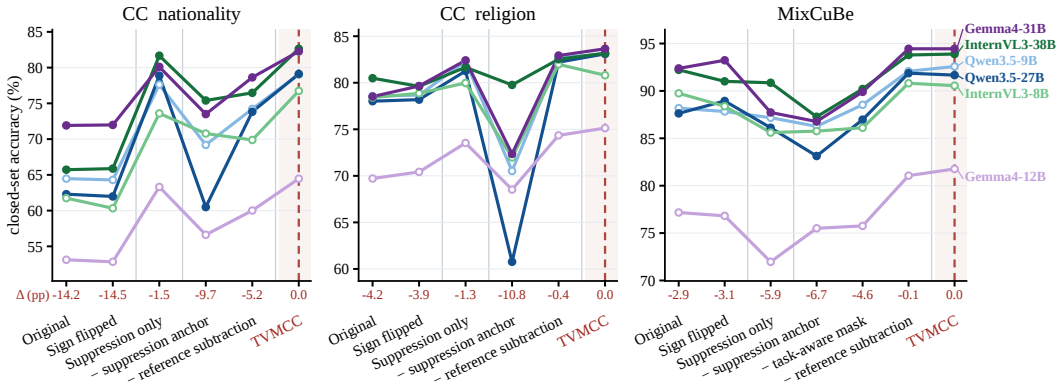



Figure 3: **Ablation study of TVMCC.** Each line denotes one MLLM. Left to right: original image, sign control, person suppression, TVMCC ablations, and the full method.

### 4.2 Main Results

We compare TVMCC with original-image scoring, prompt-based baselines, and score-level corrections across six MLLMs on CC and MixCuBe (Table 1). Three main observations emerge from the results. **(1) TVMCC consistently improves attribution.** It outperforms original-image predictions in all 18 model–task settings and achieves the highest accuracy in 17. The largest gains occur on CC nationality, ranging from 10.36 to 16.91 percentage points over original-image predictions across all models. TVMCC also consistently surpasses OA-VCD on this task, supporting the proposed three-view combination beyond original–focused contrast alone. **(2) Debiasing prompts provide partial improvements.** They improve accuracy across all CC settings but have mixed effects on MixCuBe. TVMCC outperforms this baseline throughout, indicating that explicit instructions alone leave substantial room for correction. **(3) CoT does not reliably improve attribution.** It reduces accuracy in 16 of the 18 settings, including every CC nationality setting. Under the evaluated setup, additional reasoning therefore offers no consistent advantage for identifying the correct cultural attribution.

### 4.3 Sensitivity to Mask Construction

We assess the dependence of TVMCC's gains on mask construction by replacing the grounded pipeline with DeepLab alone while retaining the view operator and scoring rule. Table 2 compares suppression and TVMCC using identical masks. Across the three models, TVMCC improves nationality attribution under both pipelines, whereas DeepLab-only results are mixed on religion. Thus, correction gains extend to an alternative mask pipeline but depend on the task and mask construction.

Table 2: Mask sensitivity on CC. Entries report suppression → TVMCC accuracy (%) using identical masks.

| Model | Mask | Nationality | Religion |
|---|---|---|---|
| Qwen3.5-9B | Grounded | 77.62 → 79.06 | 82.31 → 83.08 |
| | DeepLab | 78.77 → 79.84 | 84.89 → 84.66 |
| InternVL3-8B | Grounded | 73.59 → 76.74 | 79.98 → 80.81 |
| | DeepLab | 74.80 → 77.05 | 82.90 → 83.25 |
| Gemma4-12B | Grounded | 63.31 → 64.46 | 73.53 → 75.13 |
| | DeepLab | 64.67 → 65.35 | 77.33 → 77.25 |

### 4.4 Ablation Study

We examine the contributions of TVMCC's scoring components and task-aware masking through controlled changes to the score combination and suppression region (Figure 3). Task-aware masking is evaluated on MixCuBe, where queried content may overlap the person; CC uses a subject mask to preserve the scene. Three main observations emerge from the results. **(1) Suppression provides an effective starting point.** Replacing suppressed scores with original-image scores reduces accuracy across all models and tasks. TVMCC further improves over suppression alone in all twelve CC settings. **(2) Subtraction contributes beyond score pooling.** Omitting the focused reference lowers

accuracy for every model on CC nationality, but has mixed effects on religion and MixCuBe. Adding rather than subtracting the reference substantially weakens performance despite using the same three views. **(3) Task-aware masking improves attribution.** Replacing task-aware masks with whole-person masks lowers TVMCC accuracy across all six models on MixCuBe, supporting suppression regions adapted to the queried content.

### 4.5 Contribution of Same-Image Reference Matching

We examine whether a focused reference from the same image improves attribution beyond the magnitude of the resulting correction. Keeping the original and suppressed scores fixed, we replace the reference with one from a different matched group within the CC task, preserving the score combination and candidate order. We repeat this shuffling 100 times without selecting donors by ground-truth label. A second control rescales each centered shuffled correction to the $\ell_2$ norm of its centered same-image counterpart while retaining its shuffled direction, with centering performed across candidates, to separate pairing from magnitude. In Figure 4, bars report accuracy differences over the 100 shuffles; error bars denote marginal 95% intervals from 2,000 joint resamples of shuffle repetitions and matched groups. Across Qwen3.5-9B, InternVL3-8B, and Gemma4-12B on both CC tasks, same-image matching yields gains of 1.03–6.05 percentage points over ordinary shuffling and 1.01–2.41 points after correction-norm matching. All six intervals for the latter comparison lie above zero, supporting the contribution of input-specific reference matching beyond correction magnitude alone.

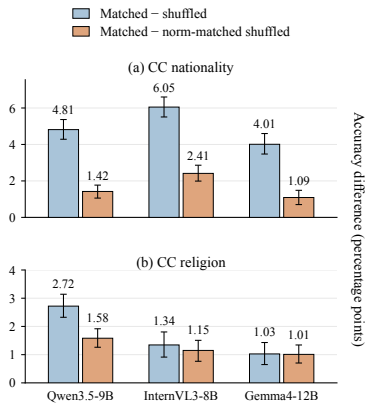



Figure 4: Matched minus shuffled accuracy on CC. Panels use different vertical scales.

### 4.6 Case Study

We inspect selected Gemma4-31B predictions on CC and MixCuBe to illustrate where matched correction improves on person suppression and where attribution errors remain (Figure 5).

**(1) Distinguishing cultural contexts with the same person.** In the CC examples, the person remains the same within each pair while the scene changes. The reference answer therefore changes with the scene, requiring the model to distinguish the cultural contexts despite the shared person. Original predictions, debiasing prompts, and OA-VCD are incorrect on all four images. Person suppression corrects the Brazil and Synagogue examples, but still predicts Brazil for the South Africa scene and Buddhist temple for the Shinto shrine. TVMCC answers all four correctly, showing how the matched correction can resolve errors that remain after suppression.

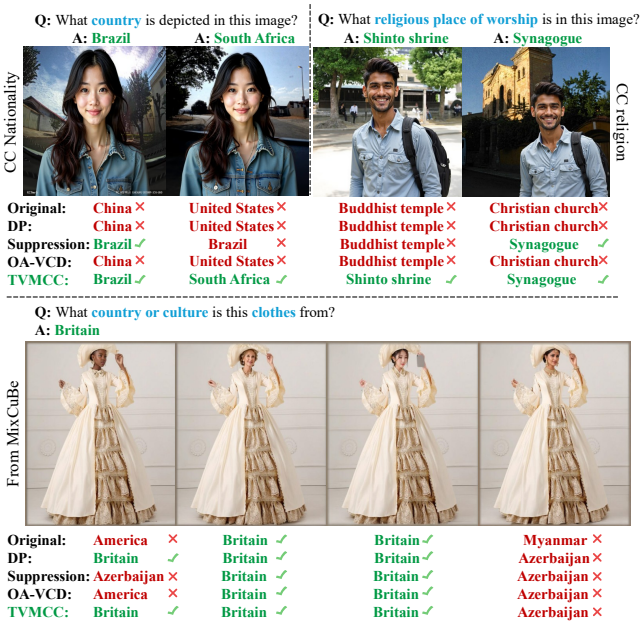



Figure 5: Qualitative results with Gemma4-31B on CC (top) and MixCuBe (bottom). Checks and crosses indicate prediction correctness. DP denotes Debias Prompt.

**(2) Preserving correct attribution across person edits.** The MixCuBe example holds the clothing fixed while changing the depicted person. On the first version, the original and OA-VCD favor America, while suppression predicts Azerbaijan. TVMCC selects the reference answer, Britain, and retains correct predictions on the next two versions. This illustrates the benefit of correcting the suppressed ranking rather than relying on person-cue removal alone.

**(3) A remaining attribution error.** On the fourth clothing version, TVMCC still predicts Azerbaijan, as do suppression, debiasing prompts, and OA-VCD. The correction therefore improves accuracy across these person edits without making attribution fully invariant to them. This remaining failure points to errors that the current three-view correction does not resolve.

## 5 Related work

**Cultural Understanding in Vision–Language Models.** Cultural understanding in vision–language models requires linking visual content to culturally specific knowledge. CulturalVQA and CVQA evaluate culturally grounded concepts across regions, languages, and visual contexts (Nayak et al., 2024; Romero et al., 2024). IndicVisionBench extends this evaluation to English and ten Indic languages, while MMAC examines cross-lingual and cross-modal cultural grounding across text, vision, and speech (Faraz et al., 2026; Zheng et al., 2026). These benchmarks mainly assess cultural coverage and understanding rather than sensitivity to the depicted person. MixCuBe directly studies this factor by changing the person's ethnic appearance while preserving cultural content (Kim et al., 2025). Cultural Counterfactuals instead places the same person in different cultural contexts (Howard et al., 2026). Related evidence also appears across modalities, where visual racial cues can alter MLLM judgments of identical speech (Kim et al., 2026). We focus on correcting cultural attribution at inference time and distinguishing correctness from consistency.

**Bias Mitigation in Visual Question Answering.** VQA models can exploit statistical shortcuts instead of relevant multimodal evidence. RUBi addresses language priors with a question-only branch that modulates training (Cadène et al., 2019), while Counterfactual Samples Synthesizing creates training examples by masking critical image regions or question words (Chen et al., 2023). Counterfactual VQA models language bias as the direct effect of the question on the answer and removes it from the total effect (Niu et al., 2021). More recently, Adaptive Logit Adjustment performs post-hoc debiasing on output logits using bias measurements and token-level importance (Jung et al., 2026). These methods intervene on biased predictions in different ways, but target different bias sources and assumptions. TVMCC instead performs inference-time correction on person-related visual evidence with a frozen MLLM, without assigning causal meaning to the score adjustment.

**Visual Grounding in Multimodal Large Language Models.** Visual grounding aims to make predictions depend on relevant image evidence. AGLA identifies prompt-relevant content through image–prompt matching and calibrates predictions with original and augmented views (An et al., 2025), while PADE extracts core visual regions from internal attention dynamics and strengthens them at inference time (Lyu et al., 2026). VCD contrasts the original image with a noise-distorted view (Leng et al., 2024); Object-Aligned VCD instead masks salient object evidence (Chen et al., 2026a), and DaID derives two contrastive anchors from internal visual attention patterns without external perturbation (Wu et al., 2026). Recent work also improves visual reliance through text-derived steering, attention intervention, or auxiliary visual clues during decoding (Gan et al., 2026; Wang et al., 2026; Yu et al., 2026; Chen et al., 2026b). TVMCC instead combines complementary same-image views for inference-time correction and tests the role of matched pairing.

## 6 Conclusion

We presented TVMCC, an inference-time method for visual cultural attribution that uses complementary views from the same image to correct person-suppressed predictions. Task-aware masking preserves queried content where possible, while the contrast between original and person-focused preferences guides correction without additional training or test labels. Across six MLLMs on Cultural Counterfactuals and MixCuBe, TVMCC improves attribution accuracy over original-image predictions. Controlled reference substitutions support the value of same-image matching beyond correction magnitude. These findings show how a visual intervention can provide both a basis for prediction and a matched reference for correcting errors that suppression alone leaves unresolved.

## Ethics Statement

We use the publicly available Cultural Counterfactuals (Howard et al., 2026) and MixCuBe (Kim et al., 2025) benchmarks with existing images and annotations, without collecting new personal data or recruiting participants. Our task concerns cultural content, not individuals' nationality, ethnicity, or religious beliefs. Reference labels do not imply exclusive cultural ownership. We aim to reduce reliance on person-related cues, while acknowledging the benchmarks' limited cultural coverage and synthetic-image artifacts.

## AI Use Statement

The authors developed the core research ideas and method, conducted the experiments, and prepared the initial manuscript. Large language models were used only for language editing and consistency checks of the manuscript. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content, results, and conclusions.

## References

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Ju-yeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya O. Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adámek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean-Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle K. Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, and Arthur Brazinskas. Gemma 4 technical report. *CoRR*, abs/2607.02770, 2026. doi: 10.48550/ARXIV.2607.02770. URL https://doi.org/10.48550/arXiv.2607.02770.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 29915–29926. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02784. URL https://openaccess.thecvf.com/content/CVPR2025/html/An_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_with_Assembly_of_CVPR_2025_paper.html.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 6763–6782. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.385. URL https://doi.org/10.18653/v1/2024.emnlp-main.385.

Rémi Cadène, Corentin Dancette, Hédi Ben-Younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In Hanna M. Wallach, Hugo Larochelle,

Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 839–850, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/51d92be1c60d1db1d2e5e7a07da55b26-Abstract.html.

Boqi Chen, Xudong Liu, and Jianing Qiu. Mask what matters: Mitigating object hallucinations in multimodal large language models with object-aligned visual contrastive decoding. *CoRR*, abs/2602.11737, 2026a. doi: 10.48550/ARXIV.2602.11737. URL https://doi.org/10.48550/arXiv.2602.11737.

Guoqing Chen, Fu Zhang, Bingqian Liu, Chenglong Lu, and Jingwei Cheng. VCGD: visual clue guided decoding with caption model for mitigating hallucination in multimodal large language models. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 20041–20049. AAAI Press, 2026b. doi: 10.1609/AAAI.V40I24.39089. URL https://doi.org/10.1609/aaai.v40i24.39089.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. URL http://arxiv.org/abs/1706.05587.

Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13218–13234, 2023. doi: 10.1109/TPAMI.2023.3290012. URL https://doi.org/10.1109/TPAMI.2023.3290012.

Ali Faraz, Akash, Shaharukh Khan, Raja Kolla, Akshat Patidar, Suranjan Goswami, Abhinav Ravi, Chandra Khatri, and Shubham Agarwal. Indicvisionbench: Benchmarking cultural and multilingual understanding in vlms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 147604–147646, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/eecdbf1beb9a006f69811b0663b45552-Paper-Conference.pdf.

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 40056–40087. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1861. URL https://doi.org/10.18653/v1/2026.acl-long.1861.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19520–19529. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01818. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gao_Interleaved-Modal_Chain-of-Thought_CVPR_2025_paper.html.

Leander Girrbach, Stephan Alaniz, Yiran Huang, Trevor Darrell, and Zeynep Akata. Revealing and reducing gender biases in vision and language assistants (vlas). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=oStNAMWELS.

Yusuke Hirota, Ryo Hachiuma, Boyi Li, Ximing Lu, Michael Ross Boone, Boris Ivanovic, Yejin Choi, Marco Pavone, Yu-Chiang Frank Wang, Noa Garcia, Yuta Nakashima, and Chao-Han Huck Yang. Bias in gender bias benchmarks: How spurious features distort evaluation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 8634–8644. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00808. URL https://doi.org/10.1109/ICCV51701.2025.00808.

Phillip Howard, Xin Su, and Kathleen C. Fraser. Cultural counterfactuals: Evaluating cultural biases in large vision-language models with counterfactual examples. *CoRR*, abs/2603.02370, 2026. doi: 10.48550/ARXIV.2603.02370. URL https://doi.org/10.48550/arXiv.2603.02370.

Hoin Jung, Junyi Chai, and Xiaoqian Wang. Adaptive logit adjustment for debiasing multimodal language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 53666–53692, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/581066637403171708903 3d95f9973ea-Paper-Conference.pdf.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. WHEN TOM EATS KIMCHI: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Yong Cao, Li Zhou, Laura Cabello, and Ife Adebara (eds.), *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pp. 143–154, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-237-4. doi: 10.18653/v1/2025.c3nlp-1.11. URL https://aclanthology.org/2025.c3nlp-1.11/.

Kyusik Kim, Hyunwoo Yoo, Jaehoon Choi, Gail Rosen, and Bongwon Suh. Visual interference in speech evaluation: Cultural asymmetry and cross-modal bias in mllms. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 27332–27358. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.1362. URL https://doi.org/10.18653/v1/2026.findings-acl.1362.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://doi.org/10.1109/CVPR52733.2024.01316.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 15214–15230. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.748. URL https://doi.org/10.18653/v1/2026.findings-acl.748.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5769–5790. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.329. URL https://doi.org/10.18653/v1/2024.emnlp-main.329.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12700–12710. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01251. URL https://openaccess.thecvf.com/content/CVPR2021/html/Niu_Counterfactual_VQA_A_Cause-Effect_Look_at_Language_Bias_CVPR_2021_paper.html.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004, September 2025. doi: 10.1162/coli.a.14. URL https://aclanthology.org/2025.cl-3.7/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, Munkh-Erdene Otgonbold, Munkhjargal Gochoo, Guido Ivetta, Luciana Benotti, Laura Alonso Alemany, Hernán Maina, Jiahui Geng, Tiago Timponi Torrent, Frederico Belcavello, Marcelo Viridiano, Jan Christian Blaise Cruz, Dan John Velasco, Oana Ignat, Zara Burzo, Chenxi Whitehouse, Artem Abzaliev, Teresa Clifford, Grainne Caulfield, Teresa Lynn, Christian Salamea Palacios, Vladimir Araujo, Yova Kementchedjhieva, Mihail Mihaylov, Israel Abebe Azime, Henok Biadglign Ademtew, Bontu Fufa Balcha, Naome A. Etori, David Ifeoluwa Adelani, Rada Mihalcea, Atnafu Lambebo Tonja, Maria Camila Buitrago Cabrera, Gisela Vallejo, Holy Lovenia, Ruochen Zhang, Marcos Estecha-Garitagoitia, Mario Rodríguez-Cantelar, Toqeer Ehsan, Rendi Chevi, Muhammad Farid Adilazuarda, Ryandito Diandaru, Samuel Cahyawijaya, Fajri Koto, Tatsuki Kuribayashi, Haiyue Song, Aditya Khandavally, Thanmay Jayakumar, Raj Dabre, Mohamed Fazli Mohamed Imam, Kumaranage Ravindu Yasas Nagasinghe, Alina Dragonetti, Luis Fernando D'Haro, Olivier Niyomugisha, Jay Gala, Pranjal A. Chitale, Fauzan Farooqui, Thamar Solorio, and Alham Fikri Aji. CVQA: culturally-diverse multilingual visual question answering benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1568882ba1a50316e87852542523739c-Abstract-Datasets_and_Benchmarks_Track.html.

Yujun Wang, Aniri, Jinhe Bi, Sören Pirk, and Yunpu Ma. ASCD: attention-steerable contrastive decoding for reducing hallucination in MLLM. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 10306–10314. AAAI Press, 2026. doi: 10.1609/AAAI.V40I12.38000. URL https://doi.org/10.1609/aaai.v40i12.38000.

Yebo Wu, Han Jin, Zhijiang Guo, and Li Li. Spotlight and shadow: Attention-guided dual-anchor introspective decoding for MLLM hallucination mitigation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 13219–13233. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.646. URL https://doi.org/10.18653/v1/2026.findings-acl.646.

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, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pp. 8320–8338. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.EACL-LONG.389. URL https://doi.org/10.18653/v1/2026.eacl-long.389.

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in lvlms. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 36021–36029. AAAI Press, 2026. doi: 10.1609/AAAI.V40I42.40918. URL https://doi.org/10.1609/aaai.v40i42.40918.

Weihua Zheng, Zhengyuan Liu, Tanmoy Chakraborty, Weiwen Xu, Xiaoxue Gao, Bryan Chen Zhengyu Tan, Bowei Zou, Chang Liu, Yujia Hu, Xing Xie, Xiaoyuan Yi, Jing Yao, Chaojun Wang, Long Li, Rui Liu, Huiyao Liu, Koji Inoue, Ryuichi Sumida, Tatsuya Kawahara, Fan

Xu, Lingyu Ye, Wei Tian, Dongjun Kim, Jimin Jung, Jaehyung Seo, Nadya Yuki Wangsajaya, Pham Minh Duc, Ojasva Saxena, Palash Nandi, Xiyan Tao, Wiwik Karlina, Tuan Luong, Keertana Arun Vasan, Roy Ka-Wei Lee, and Nancy F. Chen. MMAC: A multilingual, multimodal alignment framework for cultural grounding evaluation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 21667–21709. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.989. URL https://doi.org/10.18653/v1/2026.acl-long.989.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025. doi: 10.48550/ARXIV.2504.10479. URL https://doi.org/10.48550/arXiv.2504.10479.

# A Evaluation Protocol and Implementation

**Datasets and evaluation units.** CC nationality contains 2,669 matched person groups, each with eight cultural contexts, giving 21,352 images. CC religion contains 5,163 groups with six contexts each, giving 30,978 images. Context labels vary within a CC group. MixCuBe contains 495 cultural items, divided equally among food, festivals, and clothing. Each item has one original-person image and four person-edited versions. Evaluation uses the four edited versions: 165 groups and 660 images per category, or 1,980 images per model. The original-person images are excluded from these accuracy estimates. Within each MixCuBe group, the reference attribution is shared across the four versions. These group structures determine the resampling units used below.

**Candidate readout.** The same ordered candidate list is used for every view of a task. Each candidate is tokenized without special tokens, and its first token identifies the corresponding next-token score. The implementation checks that these token identifiers are distinct. Scores are normalized over the candidate set and combined before taking the argmax. Ties are resolved by the first candidate in the fixed order. Each view is evaluated separately; the three views are not placed in a shared multimodal prompt. Reference labels enter only the evaluation and statistical analysis.

**Model inputs and inference.** All six MLLMs are frozen and use bfloat16 inference on the eight-A100-80 GB server described in the main text. The Qwen and Gemma scoring interface converts inputs to RGB, preserves aspect ratio, and limits the longest edge to 896 pixels before processor-specific preprocessing. Thinking is disabled through the chat-template interface where supported. InternVL3 uses aspect-ratio-dependent tiling into up to six $448 \times 448$ tiles, with an additional global thumbnail when more than one tile is used. Its readout extracts the scores from a single generated token. Larger checkpoints can be sharded across devices; the server GPU count is not the GPU count of every individual evaluation. No parameter updates are performed.

**Scores and cached views.** TVMCC uses fixed coefficients $(1, 1, -1)$ and the raw corrected argmax. The same mask and replacement operator produce the suppression and focused views in a paired comparison. Masks can be reused across attribution models for a given image and intervention specification. Mask-independent baseline scores are joined to the corresponding evaluation records by image identifier. The numerical tables in this appendix retain the full precision of the stored scores during combination; rounding is applied only when displaying the results.

**Paired uncertainty estimates.** For methods $A$ and $B$, let $u_{g,j}$ be the difference of their correctness indicators on image $j$ in group $g$. We estimate the accuracy difference by averaging $u_{g,j}$ and obtain

percentile intervals by resampling groups with replacement, retaining all images in each sampled group. Groups have equal size within each reported evaluation, so averaging group means gives the same point estimate as averaging images. CC resamples person-group identifiers; MixCuBe resamples cultural-item identifiers. All intervals are marginal 95% intervals, without adjustment for multiple comparisons.

The CC baseline-effect tables retain the recorded 5,000-draw bootstrap with base seed 0 and its per-comparison random-number sequence. The mask tables use the independently recomputed 5,000-draw audit, with seed $20260925 + j$ for cell $j$ in task–model–mask order (nationality before religion; grounded before DeepLab). The MixCuBe effects use 5,000 draws with seed 0 for each comparison. Randomized-reference intervals use a separate 2,000-draw procedure that samples both a shuffle realization and matched groups; its construction is given in Appendix D. Counts of improved model–task settings summarize these evaluations rather than independent replications.

### A.1 Task Prompts and Prompt-Based Baselines

**Classification prompt.** Let $Q$ be the question and $\mathcal{C}$ the ordered candidate list in Table 3. The classification prompt is $Q$ followed by the instruction below and the candidate list, displayed as a bracketed, comma-separated list of quoted strings.

> Answer only with one of the following choices and do not provide any additional commentary: [CANDIDATE LIST]

Table 3: Questions and candidate order. For MixCuBe, CATEGORY is replaced by food, festival, or clothes.

| Task | Question | Ordered candidates |
|---|---|---|
| CC nationality | What country is depicted in this image? | India; China; United States; Brazil; South Africa; Morocco; Germany; France. |
| CC religion | What type of religious place of worship is depicted in this image? | Christian church; Buddhist temple; Shinto shrine; Mosque; Synagogue; Hindu temple. |
| MixCuBe | What country or culture is this CATEGORY from? | Azerbaijan; Korea; Myanmar; Britain; America. |

MixCuBe's dataset labels Azerbaijani, Korean, Myanmar, UK, and US map to Azerbaijan, Korea, Myanmar, Britain, and America, respectively. All candidates are presented together; the reference answer is not distinguished in the input.

**Debias Prompt.** This baseline uses the original image and prepends the following instruction to the classification prompt:

> Ignore the apparent ethnicity, skin tone, face, and appearance of any person, and ignore any cultural background hints outside the specific target of the question.

The final answer uses the same candidate first-token readout as original-image scoring.

**CoT prompt and answer extraction.** CoT first generates one trace greedily, with a limit of 320 new tokens. The reasoning prefix is:

> Let's think step by step.
> Step 1: Briefly describe the visible cultural cues in the image (architecture, artifacts, clothing, food, or scene layout).
> Step 2: Reason about the culture, religion, country, or income level these cues most likely belong to.
> Step 3: Give the final short answer as a single label.
>
> Question: [QUESTION INPUT]

For CC, QUESTION INPUT is the complete classification prompt, including its candidate list and answer-only instruction. The second-stage prompt concatenates the first-stage prompt, the generated

trace, and `Answer:`. For MixCuBe, the first stage receives the task question alone. The second stage concatenates that reasoning prompt, the trace, the full classification prompt, and `Answer:`. Both stages receive the original image. The final prediction is selected by candidate first-token scoring conditioned on this text, rather than string matching against the trace. The reported CoT comparison therefore evaluates this two-stage prompting procedure.

## B Mask Construction and Complementary Views

**Grounding inputs.** The frozen Qwen3.8-27B grounder receives an image and localization instructions. Coordinates are expressed on a $[0, 1000]$ scale. It receives neither candidate cultural answers nor attribution scores. CC images are reduced to a maximum edge of 768 pixels for grounding; MixCuBe uses a maximum edge of 1,024 pixels. The masks are rasterized at the original image resolution.

The CC localization prompt is:

> Locate the main person in this image (the largest, most prominent human subject).
> Output ONLY a JSON object, no other text, with relative coordinates in the range 0-1000:
> {`"person": [x1, y1, x2, y2], "head": [x1, y1, x2, y2]`}
> "person" = full visible extent of that person including hair, arms, hands and clothing;
> "head" = that person's head including hair and any head covering.
> If there is no human in the image, output {`"person": null, "head": null`}.

For MixCuBe, the prompt requests a JSON object with `persons`, `heads`, and `targets`, each containing a list of boxes. The full template is:

> Detect objects in this image and output ONLY a JSON object, no other text.
> Use relative coordinates in the range 0-1000 as [x1, y1, x2, y2].
> {
> `"persons": [[x1,y1,x2,y2], ...],` // every human, full visible extent incl. partially visible people, arms and hands
> `"heads": [[x1,y1,x2,y2], ...],` // head + hair of every human
> `"targets": [[x1,y1,x2,y2], ...]` // [TARGET DESCRIPTION]
> }
> Use an empty list if there is none.

TARGET DESCRIPTION is category-specific:

- Food: "the food: every dish, bowl, plate, cup or drink containing food/beverage (include the food even when a person is holding it)".
- Festival: "the festival-specific objects: decorations, lanterns, instruments, props, offerings, flags, stalls or displayed items that indicate the festival (objects only, not the people)".
- Clothes: "the clothing/garments the people are wearing".

Coordinates are clipped to $[0, 1000]$, and boxes with either side no greater than two coordinate units are discarded. Generation is greedy. The localization runners default to 80 new tokens for CC and 600 for MixCuBe; malformed MixCuBe outputs can be reparsed from complete coordinate lists or regenerated with a larger budget.

**Segmentation and mask geometry.** DeepLabV3-MobileNetV3 uses the frozen COCO-with-VOC-labels weights and person class 15. Its person map is resized to the image resolution with nearest-neighbor interpolation and dilated by $r = \max(1, \lfloor 0.015 \min(W, H) \rfloor)$ pixels. Grounded person boxes are expanded by 5% of their width and height on each side. Within each expanded box, segmented person pixels are retained; if they occupy less than 15% of the box, the entire box is used. Head boxes are expanded by 10% and included in the mask. CC retains the largest person box and heads whose centers fall inside it. If no person box is available, the implementation uses at most the first available head box.

**Category-specific interventions.** For MixCuBe Food, the mask contains the segmented person regions, excluding grounded target boxes, with expanded head regions included. Clothes uses the

union of expanded head boxes. The cached task-aware Festival evaluation also uses head-region suppression and its complementary focused view; its target-protected person-mask counterpart is retained as a construction control. In the numerical comparisons below, the category-specific pipeline therefore uses target-protected person masks for Food and head masks for Festival and Clothes. The construction control uses the cached q38 masks in all categories: target-protected person masks for Food and Festival, and head masks for Clothes. This distinction identifies the actual intervention in each category rather than treating every cached q38 mask as a whole-person mask.

**Smoothing and view replacement.** Mask edges use Gaussian smoothing with radius $\max(3, \lfloor 0.010 \min(W, H) \rfloor)$. The replacement image is formed by grayscale conversion, replication to RGB, and Gaussian blur with radius $\max(8, \lfloor 0.035 \min(W, H) \rfloor)$. The suppression view replaces pixels according to the soft mask; the focused view applies the identical operation with the complementary mask. All score comparisons within a mask setting share these views.

**Segmenter-only alternative.** DeepLab-only masks retain the same segmenter, dilation, edge smoothing, and replacement operator. They omit grounded box constraints, main-subject selection, box fallback, and explicit head inclusion. Consequently, this is a comparison of two mask pipelines, including their region-selection rules. Tables 4 and 5 report all six models on both CC tasks.

Table 4: CC nationality: accuracy (%) and paired TVMCC-minus-suppression effects (pp), with marginal 95% intervals. Each pair uses identical masks.

| Model | Mask | Suppress | TVMCC | Δ | 95% interval |
|---|---|---|---|---|---|
| Qwen3.5-9B | Grounded | 77.62 | 79.06 | +1.44 | [+1.15, +1.74] |
| Qwen3.5-9B | DeepLab | 78.77 | 79.84 | +1.07 | [+0.76, +1.38] |
| Qwen3.5-27B | Grounded | 78.87 | 79.11 | +0.24 | [-0.07, +0.54] |
| Qwen3.5-27B | DeepLab | 79.50 | 79.88 | +0.38 | [+0.08, +0.69] |
| InternVL3-8B | Grounded | 73.59 | 76.74 | +3.15 | [+2.77, +3.51] |
| InternVL3-8B | DeepLab | 74.80 | 77.05 | +2.25 | [+1.90, +2.60] |
| InternVL3-38B | Grounded | 81.67 | 82.63 | +0.96 | [+0.68, +1.26] |
| InternVL3-38B | DeepLab | 81.39 | 82.32 | +0.93 | [+0.65, +1.23] |
| Gemma4-12B | Grounded | 63.31 | 64.46 | +1.15 | [+0.81, +1.48] |
| Gemma4-12B | DeepLab | 64.67 | 65.35 | +0.68 | [+0.36, +1.00] |
| Gemma4-31B | Grounded | 80.10 | 82.27 | +2.16 | [+1.86, +2.47] |
| Gemma4-31B | DeepLab | 81.89 | 83.31 | +1.42 | [+1.14, +1.70] |

Table 5: CC religion: accuracy (%) and paired TVMCC-minus-suppression effects (pp), with marginal 95% intervals. Each pair uses identical masks.

| Model | Mask | Suppress | TVMCC | Δ | 95% interval |
|---|---|---|---|---|---|
| Qwen3.5-9B | Grounded | 82.31 | 83.08 | +0.77 | [+0.53, +1.00] |
| Qwen3.5-9B | DeepLab | 84.89 | 84.66 | -0.23 | [-0.47, +0.00] |
| Qwen3.5-27B | Grounded | 81.26 | 83.10 | +1.84 | [+1.59, +2.09] |
| Qwen3.5-27B | DeepLab | 83.96 | 85.55 | +1.59 | [+1.36, +1.83] |
| InternVL3-8B | Grounded | 79.98 | 80.81 | +0.83 | [+0.54, +1.11] |
| InternVL3-8B | DeepLab | 82.90 | 83.25 | +0.34 | [+0.08, +0.60] |
| InternVL3-38B | Grounded | 81.66 | 83.18 | +1.52 | [+1.28, +1.76] |
| InternVL3-38B | DeepLab | 83.90 | 84.79 | +0.89 | [+0.65, +1.13] |
| Gemma4-12B | Grounded | 73.53 | 75.13 | +1.60 | [+1.33, +1.87] |
| Gemma4-12B | DeepLab | 77.33 | 77.25 | -0.08 | [-0.33, +0.17] |
| Gemma4-31B | Grounded | 82.41 | 83.65 | +1.24 | [+1.00, +1.48] |
| Gemma4-31B | DeepLab | 85.90 | 85.88 | -0.01 | [-0.22, +0.20] |

On nationality, both pipelines yield positive TVMCC-minus-suppression point estimates for all six models. All six DeepLab-only intervals are above zero, whereas the grounded Qwen3.5-27B interval

includes zero. On religion, the grounded pipeline yields positive effects across all six models, while DeepLab-only yields three positive and three slightly negative estimates. The negative DeepLab-only estimates have intervals reaching or crossing zero. DeepLab-only nevertheless gives higher absolute TVMCC accuracy in five nationality settings and all six religion settings. Mask quality and the additional benefit of score correction are therefore distinct aspects of the comparison.

## C Baseline Adaptations and Paired Comparisons

**Score combinations.** Let $o, s, a$ denote the original, suppressed, and focused candidate-score vectors. Let $n$, $b$, and $l$ denote scores from the noised, blank, and relevance-selected local views. The score-level adaptations use

$$\begin{aligned} \text{CF-VQA-style}: \quad & o - b, \qquad \text{VCD-style}: \quad 2o - n, \\ \text{OA-style}: \quad & 2o - a, \quad \text{AGLA-style}: \quad (1-w)o + wl. \end{aligned}$$

All final predictions use the same closed-set argmax. These adaptations implement the reference-subtraction or view-combination operations in the evaluated scoring setting; they do not introduce the original methods' training stages or token-by-token decoding constraints.

**Noise and blank views.** The blank view is a uniform RGB $(128, 128, 128)$ image at the input resolution. Both noise implementations use a 1,000-step linear schedule from $\beta_0 = 10^{-4}$ to $\beta_{999} = 0.02$, with zero-indexed step 500 and cumulative product $\bar{\alpha}_t = \prod_{k=0}^{t}(1 - \beta_k)$. The corrupted input is $\sqrt{\bar{\alpha}_t}x + \sqrt{1 - \bar{\alpha}_t}\epsilon$. CC applies this operation to RGB values in $[0, 1]$, clips to that range, and uses the runner's NumPy seed 20260818. MixCuBe operates in $[-1, 1]$, clips in that range, and initializes a PyTorch noise generator with seed 20260818 for each image. Each result uses its recorded corrupted-view scores.

**Focused-reference comparison.** On CC, OA-style and TVMCC share the same focused reference, making $2o - a$ versus $o + s - a$ a direct replacement comparison. On MixCuBe, the cached q38 reference and the category-specific reference differ for Festival. We report both reference choices in Table 6; the task-aware score-rule table uses the same focused reference for every rule. This reference-controlled comparison is separate from the original OA-VCD attention-based view-construction procedure.

Table 6: MixCuBe OA-style accuracy (%) under two focused references, with the corresponding full TVMCC result. The score rule is $2o - a$ in both OA-style columns; only the Festival reference changes. The q38-reference column corresponds to the OA-VCD adaptation in Table 1; the task-aware-reference column provides a separate reference-controlled comparison.

| Model | OA: q38 reference | OA: task-aware reference | TVMCC |
|---|---|---|---|
| Qwen3.5-9B | 90.91 | 91.21 | 92.58 |
| Qwen3.5-27B | 89.29 | 90.40 | 91.67 |
| InternVL3-8B | 90.91 | 91.26 | 90.56 |
| InternVL3-38B | 93.09 | 93.43 | 93.89 |
| Gemma4-12B | 79.04 | 79.24 | 81.77 |
| Gemma4-31B | 93.43 | 93.69 | 94.44 |

**Local-view combination.** The AGLA-style local view uses frozen CLIP ViT-L/14 patch relevance to the task question, excluding the candidate list. Projected patch embeddings are compared with the projected text embedding by cosine similarity. A $5 \times 5$ mean filter smooths the relevance map, which is then min–max normalized. Thresholding retains the selected fraction of patches; the map is bilinearly resized, its boundary is softened, and the complement receives the grayscale-blur replacement. The main configuration retains 50% of patches and uses $w = 0.25$. Tables 7, 8, and 9 give the additional evaluated weights and retention settings. The local-view-only result corresponds to $w = 1$. A 75%-retention comparison is available for CC nationality.

Table 7: AGLA-style CC nationality accuracy (%). The first four settings retain 50% of patches; the last three retain 75%.

| Model | 50% local | $w = .25$ | $w = .50$ | $w = .75$ | 75% local | $w = .25$ | $w = .50$ |
|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | 44.31 | 62.70 | 57.84 | 49.85 | 55.27 | 63.31 | 60.87 |
| Qwen3.5-27B | 42.48 | 60.53 | 55.45 | 48.52 | 52.24 | 61.00 | 58.48 |
| InternVL3-8B | 42.25 | 59.62 | 55.80 | 48.91 | 52.20 | 60.32 | 58.34 |
| InternVL3-38B | 46.54 | 63.14 | 58.60 | 51.90 | 56.66 | 64.10 | 61.58 |
| Gemma4-12B | 36.01 | 52.28 | 49.29 | 40.57 | 45.13 | 52.72 | 51.03 |
| Gemma4-31B | 48.70 | 71.38 | 67.59 | 55.70 | 62.53 | 71.79 | 69.55 |

Table 8: AGLA-style CC religion accuracy (%) at 50% patch retention.

| Model | Local only | $w = .25$ | $w = .50$ | $w = .75$ |
|---|---|---|---|---|
| Qwen3.5-9B | 41.62 | 75.02 | 67.49 | 53.48 |
| Qwen3.5-27B | 40.76 | 74.99 | 64.80 | 49.13 |
| InternVL3-8B | 39.23 | 76.19 | 71.87 | 59.80 |
| InternVL3-38B | 45.19 | 77.84 | 72.39 | 62.41 |
| Gemma4-12B | 33.57 | 68.90 | 64.53 | 48.89 |
| Gemma4-31B | 36.99 | 77.60 | 73.94 | 62.76 |

Table 9: AGLA-style MixCuBe accuracy (%) at 50% patch retention. The $w = 0.25$ column corresponds to the AGLA-style baseline in Table 1.

| Model | $w = .25$ | $w = .50$ |
|---|---|---|
| Qwen3.5-9B | 86.87 | 83.33 |
| Qwen3.5-27B | 86.97 | 83.79 |
| InternVL3-8B | 88.89 | 86.01 |
| InternVL3-38B | 90.22 | 87.88 |
| Gemma4-12B | 75.61 | 72.32 |
| Gemma4-31B | 92.68 | 90.96 |

**Paired accuracy differences.** Tables 10 and 11 report TVMCC's effects relative to original-image scoring, same-mask suppression, and the shared-reference OA-style rule on CC. The nationality gain over suppression is small for Qwen3.5-27B and its interval includes zero. On religion, InternVL3-38B is nearly tied with OA-style. These paired estimates distinguish the numerical ordering of methods from the uncertainty in their differences.

Table 10: CC nationality: paired accuracy advantage of TVMCC (pp), with marginal 95% intervals.

| Model | vs. Original | vs. Suppression | vs. OA-style |
|---|---|---|---|
| Qwen3.5-9B | +14.58 [+13.83, +15.31] | +1.44 [+1.14, +1.74] | +7.01 [+6.48, +7.56] |
| Qwen3.5-27B | +16.81 [+16.04, +17.59] | +0.24 [-0.07, +0.55] | +8.80 [+8.23, +9.40] |
| InternVL3-8B | +15.00 [+14.28, +15.71] | +3.15 [+2.78, +3.52] | +6.98 [+6.45, +7.54] |
| InternVL3-38B | +16.91 [+16.14, +17.69] | +0.96 [+0.67, +1.26] | +8.13 [+7.57, +8.69] |
| Gemma4-12B | +11.34 [+10.73, +11.93] | +1.15 [+0.80, +1.49] | +7.69 [+7.15, +8.24] |
| Gemma4-31B | +10.36 [+9.74, +10.96] | +2.16 [+1.86, +2.47] | +7.15 [+6.62, +7.68] |

Table 11: CC religion: paired accuracy advantage of TVMCC (pp), with marginal 95% intervals.

| Model | vs. Original | vs. Suppression | vs. OA-style |
|---|---|---|---|
| Qwen3.5-9B | +4.51 [+4.09, +4.92] | +0.77 [+0.53, +1.00] | +1.57 [+1.24, +1.89] |
| Qwen3.5-27B | +5.08 [+4.66, +5.52] | +1.84 [+1.59, +2.09] | +1.42 [+1.07, +1.79] |
| InternVL3-8B | +2.39 [+1.97, +2.82] | +0.83 [+0.55, +1.12] | +0.53 [+0.19, +0.88] |
| InternVL3-38B | +2.69 [+2.31, +3.07] | +1.52 [+1.28, +1.76] | +0.01 [-0.29, +0.32] |
| Gemma4-12B | +5.42 [+4.99, +5.86] | +1.60 [+1.33, +1.87] | +3.85 [+3.45, +4.25] |
| Gemma4-31B | +5.13 [+4.74, +5.53] | +1.24 [+1.00, +1.48] | +3.22 [+2.86, +3.58] |

## D Scoring Components and Reference Controls

**Scoring-rule definitions.** The complete score grid evaluates original-image scoring $o$, suppression $s$, additive combination $o + s$, deletion of the original term $s - a$, deletion of the suppressed term $o - a$, original-for-suppressed replacement $2o - a$, suppressed-for-original replacement $2s - a$, positive pooling $o + s + a$, and TVMCC $o + s - a$. Deleting the suppressed term and replacing it with the original scores are different controls. In the ablation figure, the "suppression anchor" removal corresponds to $o - a$, while reference-subtraction removal corresponds to $o + s$. Tables 12–14 give the numerical grid.

Table 12: Complete scoring-rule grid on CC nationality. Accuracy (%); all rules use the primary grounded views.

| Model | $o$ | $s$ | $o+s$ | $s-a$ | $o-a$ | $2o-a$ | $2s-a$ | $o+s+a$ | $o+s-a$ |
|---|---|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | 64.47 | 77.62 | 74.22 | 59.59 | 69.19 | 72.04 | 76.18 | 64.29 | 79.06 |
| Qwen3.5-27B | 62.30 | 78.87 | 73.82 | 58.16 | 60.50 | 70.31 | 75.49 | 61.98 | 79.11 |
| InternVL3-8B | 61.75 | 73.59 | 69.88 | 61.70 | 70.78 | 69.76 | 71.76 | 60.33 | 76.74 |
| InternVL3-38B | 65.72 | 81.67 | 76.47 | 66.70 | 75.40 | 74.50 | 79.14 | 65.87 | 82.63 |
| Gemma4-12B | 53.12 | 63.31 | 60.01 | 57.35 | 56.62 | 56.77 | 62.84 | 52.84 | 64.46 |
| Gemma4-31B | 71.91 | 80.10 | 78.63 | 70.81 | 73.50 | 75.12 | 79.43 | 71.98 | 82.27 |

Table 13: Complete scoring-rule grid on CC religion. Accuracy (%); all rules use the primary grounded views.

| Model | $o$ | $s$ | $o+s$ | $s-a$ | $o-a$ | $2o-a$ | $2s-a$ | $o+s+a$ | $o+s-a$ |
|---|---|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | 78.57 | 82.31 | 82.68 | 61.56 | 70.52 | 81.51 | 79.38 | 78.67 | 83.08 |
| Qwen3.5-27B | 78.02 | 81.26 | 82.21 | 62.62 | 60.77 | 81.68 | 80.79 | 78.19 | 83.10 |
| InternVL3-8B | 78.41 | 79.98 | 81.97 | 60.64 | 71.98 | 80.27 | 75.76 | 78.87 | 80.81 |
| InternVL3-38B | 80.49 | 81.66 | 82.53 | 68.94 | 79.77 | 83.17 | 79.17 | 79.61 | 83.18 |
| Gemma4-12B | 69.71 | 73.53 | 74.35 | 64.67 | 68.52 | 71.28 | 72.80 | 70.42 | 75.13 |
| Gemma4-31B | 78.52 | 82.41 | 82.92 | 70.80 | 72.35 | 80.43 | 81.60 | 79.65 | 83.65 |

Table 14: Complete MixCuBe score grid using category-specific views. Accuracy (%).

| Model | $o$ | $s$ | $o+s$ | $s-a$ | $o-a$ | $2o-a$ | $2s-a$ | $o+s+a$ | $o+s-a$ |
|---|---|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | 88.18 | 91.01 | 92.07 | 81.77 | 86.26 | 91.21 | 90.20 | 87.83 | 92.58 |
| Qwen3.5-27B | 87.63 | 90.66 | 91.87 | 75.25 | 83.13 | 90.40 | 88.89 | 88.94 | 91.67 |
| InternVL3-8B | 89.75 | 89.80 | 90.81 | 74.49 | 85.76 | 91.26 | 86.46 | 88.38 | 90.56 |
| InternVL3-38B | 91.57 | 93.59 | 93.79 | 80.81 | 87.27 | 93.43 | 91.46 | 91.01 | 93.89 |
| Gemma4-12B | 77.17 | 78.48 | 81.06 | 67.73 | 75.51 | 79.24 | 76.97 | 76.82 | 81.77 |
| Gemma4-31B | 92.37 | 92.78 | 94.44 | 81.62 | 86.77 | 93.69 | 91.67 | 93.23 | 94.44 |

**Separating the mask and score comparisons.** The MixCuBe category-specific score grid uses the same views for all algebraic combinations. Table 15 separately compares that pipeline with the cached q38 construction control defined in Appendix B. Reporting suppression under both pipelines separates changes due to the view construction from changes due to score correction. The plotted

MixCuBe "suppression only" reference uses the construction-control masks; the same-mask suppression comparison is also provided here.

Table 15: MixCuBe mask-pipeline comparison. Accuracy (%); base denotes the cached q38 construction control, and task denotes category-specific views. The base pipeline already uses head masks for Clothes.

| Model | Base: Supp. | Base: TVMCC | Task: Supp. | Task: TVMCC |
|---|---|---|---|---|
| Qwen3.5-9B | 87.17 | 88.54 | 91.01 | 92.58 |
| Qwen3.5-27B | 86.11 | 86.97 | 90.66 | 91.67 |
| InternVL3-8B | 85.61 | 86.11 | 89.80 | 90.56 |
| InternVL3-38B | 90.86 | 90.20 | 93.59 | 93.89 |
| Gemma4-12B | 71.97 | 75.76 | 78.48 | 81.77 |
| Gemma4-31B | 87.73 | 89.90 | 92.78 | 94.44 |

**Fixed reference substitutions.** The fixed controls retain the recipient's original and suppressed vectors and replace the focused reference with either the blank-image vector $b$ or a cross-fitted mean $\bar{a}$. For the latter, matched groups are assigned to five deterministic folds by hashing the group identifier with the prefix `qcare-e1-20260912-v1:` using SHA-256 and taking the resulting integer modulo five. The reference mean for an evaluated group is computed from images outside its fold. This construction uses group identifiers and score vectors, not reference answers. Table 16 reports these substitutions together with additive and replacement controls.

Table 16: Fixed reference controls on CC. Accuracy (%); $b$ is the blank-image vector and $\bar{a}$ the cross-fitted reference mean.

| Model | Task | $s$ | $o+s$ | $2s-a$ | $o+s-b$ | $o+s-\bar{a}$ | TVMCC |
|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | Nat. | 77.62 | 74.22 | 76.18 | 75.78 | 76.50 | 79.06 |
| InternVL3-8B | Nat. | 73.59 | 69.88 | 71.76 | 69.28 | 72.75 | 76.74 |
| Gemma4-12B | Nat. | 63.31 | 60.01 | 62.84 | 61.13 | 61.42 | 64.46 |
| Qwen3.5-9B | Rel. | 82.31 | 82.68 | 79.38 | 83.10 | 82.60 | 83.08 |
| InternVL3-8B | Rel. | 79.98 | 81.97 | 75.76 | 82.29 | 81.98 | 80.81 |
| Gemma4-12B | Rel. | 73.53 | 74.35 | 72.80 | 74.48 | 75.31 | 75.13 |

TVMCC is highest among these fixed rules in the three nationality comparisons. Religion differs: for InternVL3-8B, additive combination, blank substitution, and the cross-fitted mean all exceed TVMCC. These results motivate assessing same-image matching against explicitly shuffled references rather than treating it as optimal among every possible reference construction.

**Cross-group shuffling.** The randomized-reference experiment covers Qwen3.5-9B, InternVL3-8B, and Gemma4-12B on both CC tasks. For each repetition, the implementation randomly orders the matched groups and cyclically shifts that order by one position to assign donor groups. It then randomly permutes the rows inside each donor group. Each recipient therefore receives a reference from another group within the same task, with candidate order unchanged. Donors are not selected by context labels or demographic annotations. This procedure samples cyclic group assignments, rather than uniformly sampling all possible derangements.

**Correction-norm matching.** Let $C(v) = v - K^{-1}(\sum_{k=1}^{K} v_k)\mathbf{1}$, $d_i = C(o_i - a_i)$, and $\widetilde{d}_i = C(o_i - a_{\pi(i)})$. Ordinary shuffling scores $s_i + \widetilde{d}_i$; centering does not change its argmax. The norm-matched control scores

$$s_i + \frac{\|d_i\|_2}{\|\widetilde{d}_i\|_2}\widetilde{d}_i.$$

The rescaling matches the centered correction vector, not the focused score vector alone. When $\|\widetilde{d}_i\|_2 \le 10^{-12}$, the implementation assigns a zero correction. No such events occur in the six recorded cells.

**Randomized-reference intervals.** Each cell uses 100 shuffles. The point estimate is the mean paired accuracy advantage of TVMCC over the shuffled control. For each of 2,000 interval draws, one shuffle repetition is sampled and matched groups are then resampled with replacement. Thus, the intervals include variation across reference assignments and evaluated groups; they are not standard errors of the mean over 100 shuffles. The base seed is 20260912. With cell index $k = 2 \times$ model index + task index, shuffle $r$ uses seed $20260912 + 10^6 k + r$; interval seeds are $20260912 + 900000 + 100k + j$, where $j = 0, 1$ denotes ordinary and norm-matched shuffling. Model order is Qwen3.5-9B, InternVL3-8B, Gemma4-12B; task order is nationality, religion.

Table 17: TVMCC minus randomized-reference accuracy (pp), with marginal 95% intervals from joint shuffle/group resampling.

| Model | Task | Shuffled reference | Norm-matched shuffled |
|---|---|---|---|
| Qwen3.5-9B | Nat. | +4.81 [+4.29, +5.37] | +1.42 [+1.06, +1.77] |
| InternVL3-8B | Nat. | +6.05 [+5.51, +6.60] | +2.41 [+1.99, +2.86] |
| Gemma4-12B | Nat. | +4.01 [+3.47, +4.60] | +1.09 [+0.70, +1.48] |
| Qwen3.5-9B | Rel. | +2.72 [+2.32, +3.14] | +1.58 [+1.26, +1.92] |
| InternVL3-8B | Rel. | +1.34 [+0.91, +1.80] | +1.15 [+0.76, +1.51] |
| Gemma4-12B | Rel. | +1.03 [+0.65, +1.43] | +1.01 [+0.70, +1.34] |

The norm-matched advantage is positive in all six cells, with all marginal intervals above zero. The comparison supports input-specific reference matching beyond correction magnitude in these settings.

## E MixCuBe Correctness Across Person Edits

**Image and group measures.** We evaluate Original, same-mask Suppression, and TVMCC on the four edited versions of each item. A group is consistently correct if all four answers equal the reference, consistently wrong if all four answers equal the same incorrect label, and varying if its answers are not all identical. These three states partition the groups. Accuracy uses images as its denominator, while group-state rates use cultural items. Each model–category cell contains 660 images and 165 groups.

Table 18: MixCuBe category results. The first three numeric columns are image accuracy (%); the last three are consistently-wrong group rates (%). O: Original; S: same-mask Suppression; T: TVMCC. Each row has 660 images and 165 groups.

| Model | Category | Acc. O | Acc. S | Acc. T | Wrong O | Wrong S | Wrong T |
|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | Food | 86.82 | 88.64 | 90.61 | 0.61 | 4.85 | 3.03 |
| Qwen3.5-9B | Festival | 90.15 | 91.21 | 92.27 | 1.21 | 5.45 | 4.85 |
| Qwen3.5-9B | Clothes | 87.58 | 93.18 | 94.85 | 3.03 | 4.85 | 2.42 |
| Qwen3.5-27B | Food | 86.06 | 86.97 | 89.09 | 1.21 | 7.27 | 4.24 |
| Qwen3.5-27B | Festival | 87.42 | 91.67 | 92.88 | 1.21 | 3.64 | 1.82 |
| Qwen3.5-27B | Clothes | 89.39 | 93.33 | 93.03 | 0.61 | 4.24 | 2.42 |
| InternVL3-8B | Food | 83.33 | 82.73 | 83.33 | 1.82 | 8.48 | 4.85 |
| InternVL3-8B | Festival | 90.45 | 90.30 | 92.88 | 1.21 | 5.45 | 2.42 |
| InternVL3-8B | Clothes | 95.45 | 96.36 | 95.45 | 0.61 | 3.03 | 3.03 |
| InternVL3-38B | Food | 89.70 | 91.82 | 92.27 | 1.21 | 3.64 | 3.03 |
| InternVL3-38B | Festival | 92.73 | 93.18 | 94.39 | 0.00 | 3.64 | 3.03 |
| InternVL3-38B | Clothes | 94.24 | 95.76 | 95.00 | 2.42 | 3.03 | 2.42 |
| Gemma4-12B | Food | 65.45 | 63.03 | 68.03 | 11.52 | 21.21 | 13.33 |
| Gemma4-12B | Festival | 84.85 | 89.39 | 91.36 | 1.82 | 4.85 | 2.42 |
| Gemma4-12B | Clothes | 81.21 | 83.03 | 85.91 | 0.61 | 7.27 | 6.06 |
| Gemma4-31B | Food | 90.61 | 90.15 | 92.42 | 1.21 | 3.03 | 1.82 |
| Gemma4-31B | Festival | 91.82 | 94.09 | 95.91 | 1.82 | 1.82 | 1.82 |

Table 18 (continued)

| Model | Category | Acc. O | Acc. S | Acc. T | Wrong O | Wrong S | Wrong T |
|---|---|---|---|---|---|---|---|
| Gemma4-31B | Clothes | 94.70 | 94.09 | 95.00 | 0.61 | 3.03 | 1.21 |

Table 19: MixCuBe group states (%). C: consistently correct; V: varying predictions. O/S/T denote Original/Suppression/TVMCC. Together with consistently-wrong rates in Table 18, the three states sum to 100% before rounding.

| Model | Category | C: O | C: S | C: T | V: O | V: S | V: T |
|---|---|---|---|---|---|---|---|
| Qwen3.5-9B | Food | 70.30 | 84.24 | 86.06 | 29.09 | 10.91 | 10.91 |
| Qwen3.5-9B | Festival | 80.61 | 88.48 | 87.88 | 18.18 | 6.06 | 7.27 |
| Qwen3.5-9B | Clothes | 75.76 | 91.52 | 90.30 | 21.21 | 3.64 | 7.27 |
| Qwen3.5-27B | Food | 67.88 | 81.21 | 83.64 | 30.91 | 11.52 | 12.12 |
| Qwen3.5-27B | Festival | 70.30 | 87.88 | 87.88 | 28.48 | 8.48 | 10.30 |
| Qwen3.5-27B | Clothes | 72.73 | 89.09 | 86.06 | 26.67 | 6.67 | 11.52 |
| InternVL3-8B | Food | 69.09 | 78.79 | 75.76 | 29.09 | 12.73 | 19.39 |
| InternVL3-8B | Festival | 80.61 | 86.67 | 87.27 | 18.18 | 7.88 | 10.30 |
| InternVL3-8B | Clothes | 90.91 | 95.76 | 93.94 | 8.48 | 1.21 | 3.03 |
| InternVL3-38B | Food | 77.58 | 89.70 | 88.48 | 21.21 | 6.67 | 8.48 |
| InternVL3-38B | Festival | 84.85 | 89.09 | 90.91 | 15.15 | 7.27 | 6.06 |
| InternVL3-38B | Clothes | 89.09 | 94.55 | 93.33 | 8.48 | 2.42 | 4.24 |
| Gemma4-12B | Food | 46.06 | 50.91 | 55.76 | 42.42 | 27.88 | 30.91 |
| Gemma4-12B | Festival | 66.67 | 83.64 | 83.03 | 31.52 | 11.52 | 14.55 |
| Gemma4-12B | Clothes | 58.79 | 74.55 | 78.18 | 40.61 | 18.18 | 15.76 |
| Gemma4-31B | Food | 80.61 | 84.85 | 87.27 | 18.18 | 12.12 | 10.91 |
| Gemma4-31B | Festival | 80.61 | 89.09 | 93.33 | 17.58 | 9.09 | 4.85 |
| Gemma4-31B | Clothes | 86.06 | 90.91 | 90.91 | 13.33 | 6.06 | 7.88 |

**Corrections and newly introduced errors.** A fix is an image answered incorrectly by suppression and correctly by TVMCC; a break is the reverse transition. Their difference divided by 1,980 gives each model's net accuracy gain over suppression. Table 20 reports both counts and the paired effect. Across the six models, there are 315 fixes and 145 breaks, corresponding to a mean gain of 1.43 percentage points. These are 11,880 model–image evaluations of the same 1,980 images, not 11,880 distinct images. TVMCC has higher accuracy than same-mask suppression in 15 of the 18 model–category cells, while the consistently-wrong rate decreases in 16 cells and is unchanged in two. The remaining accuracy decreases occur in clothing for Qwen3.5-27B and both InternVL3 models.

Table 20: MixCuBe paired transitions and effects. Fixes/breaks compare TVMCC with same-mask suppression over 1,980 images per model. Differences are in pp with marginal 95% group-bootstrap intervals.

| Model | Fixes | Breaks | vs. Suppression | vs. Original |
|---|---|---|---|---|
| Qwen3.5-9B | 52 | 21 | +1.57 [+0.45, +2.73] | +4.39 [+2.73, +6.11] |
| Qwen3.5-27B | 40 | 20 | +1.01 [+0.15, +1.97] | +4.04 [+2.47, +5.61] |
| InternVL3-8B | 48 | 33 | +0.76 [-0.40, +1.97] | +0.81 [-0.61, +2.17] |
| InternVL3-38B | 29 | 23 | +0.30 [-0.61, +1.21] | +1.67 [+0.40, +2.32] |
| Gemma4-12B | 103 | 38 | +3.28 [+1.87, +4.75] | +4.60 [+2.58, +6.67] |
| Gemma4-31B | 43 | 10 | +1.67 [+0.86, +2.58] | +2.07 [+0.81, +3.28] |

These statistics distinguish correct agreement from shared errors. In particular, a lower consistently-wrong rate need not imply a higher image accuracy in every category: a method can break a previously shared wrong answer while also changing correct answers. The paired image transitions and group-state rates describe these different outcomes.

# F CC GROUP-LEVEL RESULTS

We use the six race annotations supplied by CC for descriptive subgroup evaluation. No demographic attributes are inferred from the images. For each scoring rule, worst-group accuracy is the minimum accuracy over these annotations, and the group gap is the maximum minus the minimum. Tables 21 and 22 report Original, Suppression, and TVMCC separately, so changes from person suppression are distinguishable from subsequent correction.

Table 21: CC nationality: overall accuracy, worst-group accuracy, and maximum–minimum group gap (%).

| Model | Rule | Accuracy | Worst group | Gap |
|---|---|---|---|---|
| Qwen3.5-9B | Original | 64.47 | 42.63 | 33.29 |
| Qwen3.5-9B | Suppression | 77.62 | 75.35 | 3.77 |
| Qwen3.5-9B | TVMCC | 79.06 | 77.21 | 3.97 |
| Qwen3.5-27B | Original | 62.30 | 43.24 | 32.60 |
| Qwen3.5-27B | Suppression | 78.87 | 76.19 | 3.66 |
| Qwen3.5-27B | TVMCC | 79.11 | 76.92 | 3.72 |
| InternVL3-8B | Original | 61.75 | 47.13 | 24.90 |
| InternVL3-8B | Suppression | 73.59 | 70.73 | 4.22 |
| InternVL3-8B | TVMCC | 76.74 | 75.64 | 2.57 |
| InternVL3-38B | Original | 65.72 | 46.62 | 33.19 |
| InternVL3-38B | Suppression | 81.67 | 79.65 | 2.83 |
| InternVL3-38B | TVMCC | 82.63 | 81.10 | 2.43 |
| Gemma4-12B | Original | 53.12 | 42.83 | 16.51 |
| Gemma4-12B | Suppression | 63.31 | 60.81 | 3.64 |
| Gemma4-12B | TVMCC | 64.46 | 63.02 | 3.28 |
| Gemma4-31B | Original | 71.91 | 57.70 | 22.46 |
| Gemma4-31B | Suppression | 80.10 | 77.38 | 3.85 |
| Gemma4-31B | TVMCC | 82.27 | 80.17 | 4.09 |

Table 22: CC religion: overall accuracy, worst-group accuracy, and maximum–minimum group gap (%).

| Model | Rule | Accuracy | Worst group | Gap |
|---|---|---|---|---|
| Qwen3.5-9B | Original | 78.57 | 69.09 | 14.20 |
| Qwen3.5-9B | Suppression | 82.31 | 80.83 | 2.19 |
| Qwen3.5-9B | TVMCC | 83.08 | 82.52 | 1.12 |
| Qwen3.5-27B | Original | 78.02 | 72.13 | 11.07 |
| Qwen3.5-27B | Suppression | 81.26 | 79.29 | 3.32 |
| Qwen3.5-27B | TVMCC | 83.10 | 81.59 | 3.00 |
| InternVL3-8B | Original | 78.41 | 74.16 | 8.66 |
| InternVL3-8B | Suppression | 79.98 | 78.76 | 2.49 |
| InternVL3-8B | TVMCC | 80.81 | 79.96 | 2.05 |
| InternVL3-38B | Original | 80.49 | 75.98 | 7.78 |
| InternVL3-38B | Suppression | 81.66 | 80.13 | 2.60 |
| InternVL3-38B | TVMCC | 83.18 | 82.04 | 2.41 |
| Gemma4-12B | Original | 69.71 | 63.26 | 10.60 |
| Gemma4-12B | Suppression | 73.53 | 71.81 | 3.38 |
| Gemma4-12B | TVMCC | 75.13 | 73.64 | 2.72 |
| Gemma4-31B | Original | 78.52 | 74.12 | 7.78 |
| Gemma4-31B | Suppression | 82.41 | 80.85 | 2.89 |
| Gemma4-31B | TVMCC | 83.65 | 82.18 | 2.16 |

Suppression accounts for a large reduction in the nationality gap. TVMCC raises worst-group accuracy relative to suppression in all six nationality models, but does not reduce the gap in every

model. On religion, it raises worst-group accuracy and reduces the gap for all six. These summaries describe performance across the dataset annotations; they do not replace attribution accuracy or establish fairness across unmeasured populations.

## G Computation and Timing

**Measurement protocol.** The timing records use 200 CC nationality images on one A100. Scoring batch sizes are eight for Qwen3.5-9B and six for InternVL3-8B. Scoring uses two warm-up batches, followed by GPU synchronization before and after the timed loop. Segmentation uses three warm-up images and synchronizes each timed call. View compositing is performed on CPU. Grounding is measured separately at batch sizes one and 32, after one warm-up batch; the batch-one measurement uses 64 images, while batch 32 uses 200. Grounding uses 768-pixel inputs and a generation limit of 80 new tokens.

Table 23: Measured component times and composed estimates (ms/image). A dash denotes an unmeasured quantity. Composed estimates are not directly measured end-to-end latencies.

| Component / estimate | Qwen3.5-9B | InternVL3-8B |
|---|---|---|
| Single-view scoring | 217.1 | 256.4 |
| One grayscale-blur composite | 46.7 | 41.4 |
| DeepLab segmentation | 16.8 | 15.3 |
| Grounding, batch 1 | 5026.6 | – |
| Grounding, batch 32 | 663.1 | – |
| Suppression, DeepLab estimate | 280.6 | 313.1 |
| TVMCC, DeepLab estimate | 761.6 | 867.2 |
| TVMCC, grounded estimate (batch 32) | 1424.7 | – |

**Pipeline estimates.** Let $T_{\text{score}}$, $T_{\text{view}}$, and $T_{\text{seg}}$ denote the measured single-view scoring, compositing, and segmentation times. The composed estimates are

$$T_{\text{suppression}} = T_{\text{seg}} + T_{\text{view}} + T_{\text{score}},$$
$$T_{\text{TVMCC,seg}} = T_{\text{seg}} + 2T_{\text{view}} + 3T_{\text{score}},$$
$$T_{\text{TVMCC,grounded}} = T_{\text{ground}} + T_{\text{TVMCC,seg}}.$$

For Qwen3.5-9B, these give 761.6 ms per image for segmenter-only TVMCC and 1,424.7 ms with batched grounding, compared with 217.1 ms for original-image scoring. The corresponding ratios are 3.51 and 6.56. These are sums of measured components, not direct end-to-end latency measurements. The same measured scoring time is used for all three views; mask postprocessing, file I/O, model loading, and scheduling overhead are not exhaustively included. Grounding generates coordinates, whereas attribution uses closed-set scoring without a reasoning trace. The estimates do not amortize mask construction across models; cached masks can be reused when the image and intervention specification are unchanged.

## H Qualitative Example Records

The selected Gemma4-31B examples illustrate residual suppression errors and a remaining TVMCC failure. The CC nationality pair is group `1006_young_East_Asian_woman`, with Brazil and South Africa contexts. The religion pair is `1309_young_South_Asian_man`, with Synagogue and Shinto shrine contexts. The MixCuBe clothing example is item `UK_Clothes_10`; all four edited versions share the reference answer Britain. Table 24 records the predictions. The person-pair relationships follow the dataset grouping; image edits need not leave pose, lighting, or all non-person pixels unchanged.

Table 24: Gemma4-31B example predictions. N/R/M identify nationality, religion, and MixCuBe images. Religious-place labels are abbreviated as Christian (church), Buddhist (temple), and Shinto (shrine). DP denotes Debias Prompt.

| Image | Reference | Original | DP | Suppression | OA-style | TVMCC |
|---|---|---|---|---|---|---|
| N1 | Brazil | China | China | Brazil | China | Brazil |
| N2 | South Africa | United States | United States | Brazil | United States | South Africa |
| R1 | Shinto | Buddhist | Buddhist | Buddhist | Buddhist | Shinto |
| R2 | Synagogue | Christian | Christian | Synagogue | Christian | Synagogue |
| M1 | Britain | America | Britain | Azerbaijan | America | Britain |
| M2 | Britain | Britain | Britain | Britain | Britain | Britain |
| M3 | Britain | Britain | Britain | Britain | Britain | Britain |
| M4 | Britain | Myanmar | Azerbaijan | Azerbaijan | Azerbaijan | Azerbaijan |

The images were selected after inspecting predictions, and the complete four-version clothing example includes the failure. They explain particular answer changes rather than estimate a population success rate. The aggregate accuracy, paired effects, and group-state tables provide the quantitative evaluation.