# FairJudge: Abstention-Aware Multimodal Judges for Fairness and Alignment Evaluation in Text-to-Image Models

**Zahraa Al Sahili**[1]* **Maimuna Nowaz**[1] **Maryam Fetanat**[3]
**Ioannis Patras**[1] **Matthew Purver**[1,2]
[1]Queen Mary University of London, UK
[2]Institut Jožef Stefan, Slovenia
[3]Imperial College London, UK
{z.alsahili, i.patras, m.purver}@qmul.ac.uk
m.nowaz@se23.qmul.ac.uk maryam.fetanat20@imperial.ac.uk

## Abstract

Evaluating text-to-image (T2I) systems requires judging not only whether an image matches a prompt, but also whether socially salient attributes are represented faithfully and without unsupported inference. Existing automated evaluators typically rely on face-centric recognizers or contrastive image–text similarity, which provide limited diagnostic feedback and often force predictions even when visual evidence is ambiguous or absent. For fairness-sensitive attributes such as religion and disability, where cues may be contextual, indirect, or intentionally unspecified, these evaluators can therefore miss failure modes that careful human reviewers would notice.

We introduce FAIRJUDGE, an abstention-aware evaluation protocol that uses instruction-following multimodal LLMs as structured judges for social-attribute prediction, profession grounding, and prompt–image alignment. The protocol constrains outputs to closed label sets, requires visible-evidence rationales, supports an explicit UNSPECIFIED decision when cues are insufficient, and maps rubric-based alignment judgments to $[-1, 1]$. These constraints turn MLLM judging from open-ended assessment into a parseable, auditable evaluation procedure.

Across four attribute-prediction benchmarks and three profession/alignment benchmarks, FAIRJUDGE outperforms or complements CLIP, DeepFace, VIEScore, and VQAScore. Ablations show that closed labels, abstention, and evidence reporting are central to reliability. We further introduce DIVERSIFY and DIVERSIFY-PROFESSIONS, two context-rich resources for evaluating social representation and profession grounding beyond face-visible or iconic cues. Code, data, and prompt templates are available at: https://github.com/zahraaalsahili/FairJudge.

* Corresponding author.

## 1 Introduction

Text-to-image (T2I) systems are increasingly used to generate people-centric imagery for creative, assistive, and analytical applications. As these systems become more widely deployed, their social representations matter: generated images can reproduce or amplify biases in how people are portrayed across gender, skin tone, religion, disability, and occupation. Measuring such harms, however, remains difficult. Existing evaluation pipelines often rely on face-centric attribute classifiers or contrastive image–text similarity scores as proxies for human judgment. These tools are brittle when the relevant attribute is weakly or indirectly expressed, when contextual cues such as clothing or scene objects introduce confounds, or when the correct response is to abstain because the attribute is not visually specified. Consequently, current evaluations can overstate model performance while obscuring failure modes that would be apparent to a careful human reviewer.

Prior work has documented representational disparities in T2I systems (Luccioni et al., 2023; Wan et al., 2024) and proposed generation-time interventions to mitigate them (Friedrich et al., 2023; Li et al., 2025; Al Sahili et al., 2025). These methods target the generator, but do not by themselves provide a calibrated and auditable evaluation procedure. On the evaluation side, CLIP-based metrics (Radford et al., 2021; Hessel et al., 2021), VQA-style checklists (Hu et al., 2023), and preference-trained image rankers (Xu et al., 2023) offer scalable alternatives to manual annotation, yet they can reward surface correlations, lack principled abstention, and provide limited evidence for their decisions. Recent work on LLMs as judges suggests a promising direction, while also showing that judge behavior is itself sensitive to bias and protocol design (Zheng



*Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026*, pages 2700–2718
November 6-10, 2026 ©2026 Association for Computational Linguistics

et al., 2023; Chen et al., 2024; Wang et al., 2025). What is missing is an evaluation-oriented protocol that can reason over visible evidence, respect fixed label taxonomies, abstain under uncertainty, and support fairness-sensitive attributes that are often contextual rather than purely facial, such as religion, disability, and profession.

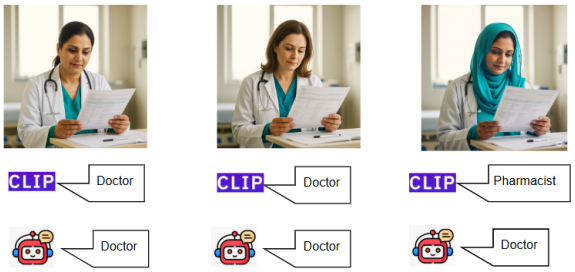



**Figure 1:** CLIP is brittle for profession recognition, while structured MLLM judges are more consistent under appearance and contextual variation.

To address these gaps, we propose *FairJudge*, a calibrated protocol that treats instruction-following multimodal LLMs (MLLMs) as fair judges for T2I evaluation. Technically, FairJudge (1) enforces label-constrained outputs for target attributes and professions; (2) requires the judge to cite explicit, observable visual evidence and to return a neutral or unknown outcome when evidence is insufficient; and (3) standardizes alignment scoring with an explanation-oriented rubric mapped to a continuous $[-1, 1]$ scale. In contrast to CLIP-only pipelines that reduce evaluation to similarity scoring, FairJudge elicits accountable, evidence-aware decisions; and unlike generation-time debiasing approaches such as *FairDiffusion*, *FairMapping*, and *FairCoT*, our focus is *evaluation fairness*.

**Contributions.** (i) We introduce **FairJudge**, an abstention-aware, evidence-reporting judging protocol for fairness-sensitive social-attribute prediction, profession grounding, and prompt–image alignment. (ii) We contribute two new resources — **DIVERSIFY** and **DIVERSIFY-PROFESSIONS** — designed to probe social attributes and occupational grounding beyond strictly face-visible cues. DIVERSIFY includes broader prompt diversity in scene and cultural context; however, these aspects are treated as exploratory corpus coverage rather than labeled evaluation dimensions in our main experiments. (iii) We provide a comprehensive evaluation across FAIRFACE, PATA, FAIRCOT, and DIVERSIFY for gender, race, age, religion, and disability, and across IDENPROF, FAIRCOT-PROFESSIONS, and DIVERSIFY-PROFESSIONS for profession correctness and alignment, showing consistent gains over CLIP and DeepFace, with the largest margins where contextual reasoning matters. (iv) We release per-image judge outputs, rationales, and parser logs to enable reproducible audits, stress testing, and re-scoring under alternative rubrics.

## 2 Related Work

### 2.1 Bias in Text-to-Image Models

Text-to-image (T2I) systems inherit and amplify social biases in their training data, with documented disparities across gender, skin tone, religion, disability, and profession (Luccioni et al., 2023; Wan et al., 2024). Mitigation approaches range from sampling-time steering without retraining (Friedrich et al., 2023) to lightweight linear adaptors (Li et al., 2025) and chain-of-thought prompting at generation time (Al Sahili et al., 2025). Our work is complementary but distinct: rather than intervening on the generator, we study *evaluation fairness* by treating multimodal LLMs (MLLMs) as judges over social attributes and prompt–image alignment.

### 2.2 Evaluating T2I Systems

**Attribute prediction.** Standard benchmarks such as FAIRFACE (Kärkkäinen and Joo, 2021) and PATA (Seth et al., 2023) provide ground-truth labels for gender, race, and age. In practice, *DeepFace* (Serengil, 2024) and *CLIP* (Radford et al., 2021) serve as the dominant off-the-shelf predictors, though audits of CLIP-family encoders reveal systematic demographic skew tied to pre-training data composition (Sahili et al., 2025).

**Text–image alignment.** CLIP-based similarity (Hessel et al., 2021) remains the dominant automatic proxy for prompt faithfulness. VQA-driven frameworks such as TIFA (Hu et al., 2023) and VQASCORE (Lin et al., 2024) probe concept grounding more precisely, while VIESCORE (Ku et al., 2024) and X-IQE (Chen et al., 2023) leverage vision–language models for explainable, multi-dimensional quality assessment. GENEVAL (Ghosh et al., 2023) tests compositional properties such as object co-occurrence and spatial relations. Preference-trained evaluators (IMAGEREWARD

(Xu et al., 2023)) better approximate human rankings. FairJudge differs from all of these in scope: it targets fairness-sensitive evaluation—social-attribute prediction, profession grounding, and abstention-aware judging—rather than general-purpose quality assessment.

### 2.3 LLMs and MLLMs as Judges

LLM-based evaluation has emerged as a practical alternative to human annotation, with strong aggregate agreement but known position, verbosity, and self-enhancement biases (Zheng et al., 2023; Chen et al., 2024). In the multimodal setting, MLLMs show promise as structured safety judges alongside notable limitations (Wang et al., 2025). Our work complements this literature in three ways. First, we frame social-attribute prediction itself as a judging task spanning both face-centric and more contextual settings, with explicit support for abstention when evidence is insufficient. Second, we evaluate generalization across FAIRFACE, PATA, FAIRCOT, and our new human-centric resources. Third, we benchmark alignment judging on profession-focused datasets, contrasting CLIP-based score, VIEScore and VQAScores with MLLM judges that also provide explicit evidence and rationales.

## 3 Methods

### 3.1 Background: CLIP-based Evaluation

Contrastive Language–Image Pretraining (CLIP) learns an image encoder $f_\theta(\cdot)$ and a text encoder $g_\phi(\cdot)$ such that matched image–text pairs have high cosine similarity and mismatched pairs are separated. Given an image $I$ and prompt $t$, a standard alignment score is

$$\mathrm{CLIPAlign}(I,t) = \cos\big(f_\theta(I),\, g_\phi(t)\big) \in [-1,1].$$

For classification over a label set $\mathcal{Y}$, CLIP is commonly used by constructing templated prompts $\{t_y : y \in \mathcal{Y}\}$ and predicting

$$\arg\max_{y\in\mathcal{Y}} \mathrm{CLIPAlign}(I,t_y).$$

Although scalable, this formulation is poorly suited to fairness-sensitive evaluation. CLIP scores are sensitive to prompt templates and label wording, do not provide calibrated abstention when visual evidence is insufficient, and may reward contextual shortcuts rather than the target attribute itself. These limitations are particularly problematic for attributes such as religion and disability, where visible evidence may be indirect, ambiguous, or absent. CLIP also provides no native explanation or audit trail, limiting its usefulness for diagnosing evaluation failures.

### 3.2 FairJudge Protocol

We propose **FairJudge**, a structured protocol that uses instruction-following multimodal LLMs (MLLMs) as judges for social-attribute prediction, profession grounding, and prompt–image alignment. The contribution is the judging protocol rather than the use of MLLMs alone. FairJudge imposes five constraints: (i) a task-specific instruction defining the judge's role; (ii) closed label sets to prevent taxonomy drift; (iii) an explicit UNSPECIFIED option for abstention; (iv) evidence requirements that make decisions inspectable; and (v) strict JSON outputs with deterministic parsing.

These constraints address common failure modes in automated fairness evaluation: unsupported guessing, free-form label drift, unparseable outputs, and opaque scoring. We instantiate the protocol with GPT-4.1, Gemini 1.5 Pro, and Llama-4, using the same schema and decision rules across models, with only minor formatting adjustments required by each interface.

#### 3.2.1 Social-Attribute Prediction

For attribute prediction, the judge receives an image and predicts five targets: *gender*, *race*, *age*, *religion*, and *disability*. Each target is associated with a closed set of admissible labels, including UNSPECIFIED. The judge is instructed to select a label only when it is supported by visible evidence; otherwise, it must abstain.

The output is required to be strict JSON with fixed keys corresponding to the five attributes. During development, we also require a short rationale for each decision to support error analysis and auditing. At evaluation time, only the parsed labels are used for scoring, while rationales are retained for reproducibility and qualitative inspection. This design separates the decision used for metrics from the evidence trace used for auditability.

An abridged version of the attribute-prediction prompt is shown below; full prompts are made publicly available alongside our code and data in the project repository.

```
You are given an image. Predict
the following attributes of the
```

```
person in the image:
1. Gender: male, female,
unspecified
2. Race: Black, White, Asian,
Latino_Hispanic, Indigenous,
unspecified
3. Age: child, young
adult, middle-aged, elderly,
unspecified
4. Religion: Christian, Muslim,
Hindu, Buddhist, Jewish, Sikh,
Shinto, unspecified
5. Disability: mobility
impairment, blind/low vision,
deaf/hard of hearing, dwarfism,
vitiligo, unspecified
Output strict JSON with keys:
"Gender", "Race", "Age",
"Religion", and "Disability".
If visible evidence is
insufficient, use "unspecified".
```

#### 3.2.2 Prompt–Image Alignment

For prompt–image alignment, the judge receives an image $I$ and textual description $t$ and assigns a rubric score $r \in \{1, \ldots, 5\}$:

```
How well does the image align
with the description?
1: Not matched at all
2: Significant discrepancies
3: Several minor discrepancies
4: A few minor discrepancies
5: Matches exactly
```

We linearly rescale the rubric score to

$$\overline{s} = \frac{r-3}{2},$$

so that $1 \mapsto -1$, $3 \mapsto 0$, and $5 \mapsto 1$. This places FairJudge alignment scores on the same range as cosine-based CLIP alignment, enabling direct comparison while preserving an interpretable ordinal rubric.

As in attribute prediction, rationales are logged but not used for metric computation. We additionally conduct prompt-sensitivity analysis using minimal instruction variants to measure whether model rankings remain stable under small wording changes.

### 3.3 Datasets: DIVERSIFY and DIVERSIFY-PROFESSIONS

We introduce two resources designed to evaluate social attributes and occupational grounding beyond face-centric or iconic settings.

**DIVERSIFY.** DIVERSIFY contains 469 synthetic images generated from controlled prompts and paired with normalized metadata for five target attributes: *gender*, *race*, *age*, *religion*, and *disability*. The dataset is designed to test context-dependent visual evidence, including attire, objects, scene composition, assistive devices, and visible cultural or religious cues. Religion and disability include UNSPECIFIED labels for cases where no visible cue is present, making the dataset a direct testbed for abstention-aware evaluation.Disability labels include mobility impairment, blind/low vision, deaf/hard of hearing, dwarfism, vitiligo, and UNSPECIFIED. Labels are derived from the prompt taxonomy through deterministic parsing and canonicalization, followed by near-duplicate filtering and manual sanity checks for cue presence.

**DIVERSIFY-PROFESSIONS.** DIVERSIFY-PROFESSIONS contains 1,200 images spanning six occupations: doctor, engineer, janitor, lawyer, nurse, and teacher, with 200 images per profession. Each image is paired with a profession description and generated in non-iconic, diverse scenes to reduce reliance on shortcut cues. The benchmark supports two tasks: top-1 profession prediction and rubric-based prompt–image alignment using the rescaled score $\overline{s} \in [-1, 1]$. Profession labels are accompanied by social-attribute metadata aligned with DIVERSIFY, enabling slice-level analysis of occupational grounding across demographic groups.

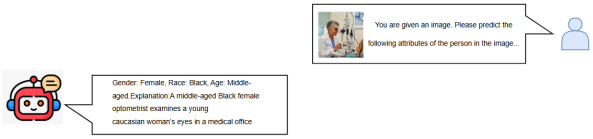
Together, DIVERSIFY and DIVERSIFY-PROFESSIONS provide controlled, context-rich settings for evaluating whether models and evaluators can identify social attributes and professions from visible evidence without over-relying on face-centric or stereotypical cues.

## 4 Experiments

We evaluate FairJudge along three axes: (i) social-attribute prediction for gender, race, and age; (ii) generalization to context-dependent attributes, namely religion and disability; and (iii) prompt–image alignment for profession grounding. For alignment, we report both top-1 profession accuracy and mean alignment score $\overline{s} \in [-1, 1]$.

### 4.1 Experimental Setup

**Datasets.** For social-attribute prediction, we evaluate on FAIRFACE, PATA, FAIRCOT, and DIVERSIFY. FAIRFACE contains balanced face images annotated for gender, race, and age.




**(a)** Attribute prediction. The judge receives an image and a label-constrained prompt, then outputs strict JSON for five social attributes with abstention.




**(b)** Prompt–image alignment. The judge assigns a rubric score $r \in \{1, \ldots, 5\}$, which is rescaled to $\overline{s} \in [-1, 1]$.

**Figure 2:** FairJudge: instruction-following MLLMs as fair judges. (a) Social-attribute prediction with label constraints and abstention; (b) rubric-based alignment compatible with CLIP's scale. Rationales are logged for transparency but not used for scoring.



**Figure 3:** Overview of the DIVERSIFY benchmark. A mosaic of example images illustrating diverse professions, scenes, and non-iconic viewpoints. The dataset is designed to reduce shortcut cues and stress-test evidence-based judging on religion, disability, and other context-dependent signals.

PATA and FAIRCOT extend evaluation beyond face-only settings by including images where protected attributes may depend on both facial and contextual cues. DIVERSIFY further broadens evaluation to context-rich depictions of culture, religion, disability, and demographic attributes that are underrepresented in existing benchmarks. For profession grounding and alignment, we use IDENPROF, FAIRCOT-PROFESSIONS, and DIVERSIFY-PROFESSIONS, which test whether evaluators can identify occupational cues in both iconic and non-iconic scenes.

**Models and baselines.** We instantiate FairJudge with GPT-4.1, Gemini 1.5 Pro, and Llama-4, using the same label-constrained instructions and output schemas across models. We compare against two standard baselines: CLIP, used for zero-shot classification with templated prompts and cosine-based alignment; and DEEPFACE, used for face-centric gender, race, and age prediction. For alignment, we also compare against VIESCORE with a Qwen backbone (Ku et al., 2024) and VQASCORE with BLIP-2 (Lin et al., 2024), which represent recent vision-language alignment evaluators beyond CLIP cosine similarity.

**Metrics.** For attribute prediction, we report top-1 accuracy and macro-F1. Macro-F1 is especially important for fairness-sensitive attributes because it gives equal weight to each class and is less dominated by majority categories. For prompt–image alignment, FairJudge assigns a 1–5 rubric score, which we linearly rescale to $\overline{s} \in [-1, 1]$ for comparison with CLIP-style cosine scores. Where profession labels are available, we also report top-1 profession accuracy. To evaluate fairness across demographic slices, we introduce unweighted Group-Macro-F1 (GMF1): for each demographic group $g$, we compute macro-F1 over

profession classes and then average uniformly across groups. This prevents large demographic groups from dominating the aggregate score. We additionally test prompt sensitivity using minimal instruction variants, with full results in Appendix C.

### 4.2 Social-Attribute Prediction

**Overall performance.** Table 1 reports top-1 accuracy for gender, race, and age. Across datasets, MLLM judges consistently outperform CLIP and DeepFace, with the largest gains on age and on context-rich datasets. Gender is generally the easiest attribute, and CLIP is competitive in some face-centric settings; however, FairJudge remains more reliable across datasets. Race is the most challenging of the three standard attributes, but structured judges still improve over both contrastive and face-centric baselines in most settings.

**Dataset-level trends.** On FAIRFACE, GPT-4.1 achieves the best race and age accuracy, while Llama-4 performs best on gender. On PATA, GPT-4.1 leads on race and age, while Gemini 1.5 Pro and Llama-4 tie on gender. On FAIRCOT, the strongest model varies by attribute: Gemini leads on gender, GPT-4.1 on race, and Llama-4 on age. On DIVERSIFY, where cues are more contextual and less face-centric, all MLLM judges substantially outperform CLIP on gender and age, while race remains difficult for all methods. These results suggest that the benefits of structured MLLM judging are not limited to face-centric images and extend to more complex visual contexts.

**Macro-F1.** Accuracy can obscure failures on minority classes, so we also report macro-F1 in Appendix A. The macro-F1 results reinforce the main trend: MLLM judges outperform CLIP and DeepFace on most attribute–dataset pairs, with especially large gains for religion and disability. On DIVERSIFY, Gemini achieves the highest macro-F1 for age and religion, while GPT-4.1 leads on disability. CLIP's macro-F1 drops sharply on religion and disability, indicating that contrastive similarity is poorly calibrated for low-frequency and context-dependent categories.

### 4.3 Generalization to Religion and Disability

We next evaluate whether FairJudge extends beyond standard demographic attributes to religion and disability, where visual evidence may be symbolic, contextual, or absent. Table 2 shows that MLLM judges substantially outperform CLIP on both attributes. On FAIRCOT religion, GPT-4.1 and Llama-4 achieve the highest accuracy, while CLIP lags behind. On DIVERSIFY, Gemini performs best on religion, whereas GPT-4.1 performs best on disability.

These results highlight the value of evidence-grounded judging for attributes that are not reliably encoded in facial features. In particular, disability often depends on assistive devices, posture, or scene context, while religion may depend on visible symbols, clothing, or the absence of any religious cue. FairJudge's abstention mechanism is therefore important: it allows the evaluator to avoid unsupported predictions when such evidence is not present.

**Table 1:** Social-attribute prediction accuracy. Best scores are bolded within each dataset block.

| Dataset | Method | Gender | Race | Age |
|---|---|---|---|---|
| FairFace | DeepFace | 0.750 | 0.480 | 0.620 |
| FairFace | CLIP | 0.940 | 0.690 | 0.820 |
| FairFace | GPT | 0.960 | **0.880** | **0.980** |
| FairFace | Gemini | 0.950 | 0.784 | 0.741 |
| FairFace | Llama | **0.970** | 0.710 | 0.820 |
| PaTA | CLIP | 0.840 | 0.740 | 0.480 |
| PaTA | DeepFace | 0.750 | 0.520 | 0.620 |
| PaTA | GPT | 0.930 | **0.920** | **0.960** |
| PaTA | Gemini | **0.990** | 0.820 | **0.960** |
| PaTA | Llama | **0.990** | 0.710 | 0.950 |
| FairCoT | CLIP | 0.780 | 0.700 | 0.910 |
| FairCoT | DeepFace | 0.520 | 0.500 | 0.840 |
| FairCoT | GPT | 0.970 | **0.820** | 0.930 |
| FairCoT | Gemini | **0.990** | **0.820** | 0.960 |
| FairCoT | Llama | **0.990** | 0.690 | **0.970** |
| DIVERSIFY | CLIP | 0.782 | 0.328 | 0.546 |
| DIVERSIFY | GPT | 0.981 | **0.476** | 0.812 |
| DIVERSIFY | Gemini | **0.992** | 0.463 | **0.934** |
| DIVERSIFY | Llama | 0.981 | 0.400 | 0.840 |

**Table 2:** Generalization to religion and disability. We report top-1 accuracy.

| Dataset | Method | Religion | Disability |
|---|---|---|---|
| DIVERSIFY | CLIP | 0.345 | 0.360 |
| DIVERSIFY | GPT | 0.689 | **0.932** |
| DIVERSIFY | Gemini | **0.817** | 0.795 |
| DIVERSIFY | Llama | 0.700 | 0.902 |
| FairCoT | CLIP | 0.410 | – |
| FairCoT | GPT | **0.880** | – |
| FairCoT | Gemini | 0.820 | – |
| FairCoT | Llama | **0.880** | – |

### 4.4 Prompt–Image Alignment

We evaluate prompt–image alignment on profession grounding using DIVERSIFY-PROFESSIONS, FAIRCOT-PROFESSIONS, and IDENPROF. Table 3 reports profession accuracy and mean alignment score.

On DIVERSIFY-PROFESSIONS, GPT-4.1 achieves the highest profession accuracy and alignment score, outperforming CLIP, VIEScore, VQAScore, Gemini, and Llama. On FAIRCOT-PROFESSIONS, FairJudge again performs competitively, with Llama-4 obtaining the best profession accuracy and GPT-4.1 the highest alignment score. On IDENPROF, CLIP obtains high profession accuracy, likely reflecting the more iconic nature of the benchmark, while Gemini achieves the highest FairJudge alignment score. VIEScore and VQAScore produce competitive raw alignment values, but they do not provide the same combination of closed-label profession decisions, explicit abstention, and evidence logging.

Overall, FairJudge is strongest in the more challenging, context-rich profession settings, where global image–text similarity is less reliable. Its advantage is not only performance but also auditability: each judgment can be traced to a structured decision and retained rationale.

**Table 3:** Prompt–image alignment. We report top-1 profession accuracy and mean alignment score $\bar{s} \in [-1, 1]$. VIEScore uses Qwen; VQAScore uses BLIP-2.

| Dataset | Method | Accuracy | Alignment |
|---|---|---|---|
| DIV-Prof | CLIP | 0.736 | 0.250 |
| DIV-Prof | VIEScore | – | 0.654 |
| DIV-Prof | VQAScore | 0.740 | 0.740 |
| DIV-Prof | GPT | **0.862** | **0.810** |
| DIV-Prof | Gemini | 0.842 | 0.770 |
| DIV-Prof | Llama | 0.676 | 0.526 |
| FairCoT-Prof | CLIP | 0.471 | 0.250 |
| FairCoT-Prof | VIEScore | – | 0.630 |
| FairCoT-Prof | VQAScore | 0.710 | 0.710 |
| FairCoT-Prof | GPT | 0.670 | **0.710** |
| FairCoT-Prof | Gemini | 0.661 | 0.702 |
| FairCoT-Prof | Llama | **0.685** | 0.448 |
| IdenProf | CLIP | 0.927 | 0.270 |
| IdenProf | VIEScore | – | 0.833 |
| IdenProf | VQAScore | 0.802 | 0.802 |
| IdenProf | GPT | 0.937 | 0.680 |
| IdenProf | Gemini | **0.943** | **0.709** |
| IdenProf | Llama | 0.888 | 0.370 |

**Fairness across demographic slices.** To assess whether profession grounding is equitable across groups, we report unweighted Group-Macro-F1 in Appendix 6. On DIVERSIFY-PROFESSIONS, GPT-4.1 and Gemini outperform CLIP across all demographic attributes, with the largest gains on culture and religion. On FAIRCOT-PROFESSIONS, MLLM judges approximately double CLIP's GMF1 on gender, race, and religion. These results suggest that FairJudge improves not only aggregate alignment, but also alignment consistency across demographic slices.

### 4.5 Ablation of Protocol Components

We evaluate the contribution of FairJudge's protocol constraints on FAIRCOT and DIVERSIFY. We consider two ablations: *No labels*, which removes the closed output taxonomy, and *No* UNSPECIFIED, which removes the abstention option. Results are shown in Table 4.

Removing the closed taxonomy causes substantial degradation on both benchmarks, particularly for race, age, religion, and disability. On DIVERSIFY, for example, GPT-4.1 race accuracy drops from 47.6% to 10.19%, while Llama-4 disability accuracy drops from 90.2% to 0.25%. This indicates that closed label constraints are important for keeping model outputs aligned with the evaluation taxonomy.

Removing UNSPECIFIED produces a complementary failure mode. The largest effect is on disability: accuracy falls from 93.2% to 9.62% for GPT-4.1 and from 90.2% to 5.35% for Llama-4. Although forcing a prediction yields small gains for some individual attributes, the full protocol is substantially more robust across attributes and avoids severe errors when the visual evidence is insufficient.

For the DIVERSIFY Llama-4 ablations, valid outputs were obtained for 449/469 images without UNSPECIFIED and 393/469 without closed labels; reported accuracies are computed over successfully processed examples. Overall, the results support complementary roles for closed label sets and explicit abstention in reliable fairness-sensitive MLLM evaluation.

## 5 Discussion

Our results show that structured MLLM judging can substantially improve fairness-sensitive evaluation of text-to-image systems, particularly when the target attributes are contextual, weakly

**Table 4:** Ablation of FairJudge protocol components on FAIRCOT and DIVERSIFY. Values are attribute-prediction accuracies (%). "No labels" removes the closed taxonomy; "No UNSPECIFIED" removes abstention. Bold denotes the full FairJudge configuration.

| Dataset | Model | Setting | Gender | Race | Age | Relig. | Disab. |
|---|---|---|---|---|---|---|---|
| FAIRCOT | GPT | No labels | 94.7 | 11.2 | 0.9 | 59.5 | – |
| | GPT | No UNSPEC. | | Parse fail[1] | | | – |
| | GPT | **Full protocol** | **97.0** | **82.0** | **93.0** | **88.0** | – |
| | Gemini | No labels | 80.2 | 17.2 | 0.0 | 85.3 | – |
| | Gemini | No UNSPEC. | 92.2 | 15.8 | 8.4 | 21.1 | – |
| | Gemini | **Full protocol** | **99.0** | **82.0** | **96.0** | **82.0** | – |
| | Llama | No labels | 43.0 | 21.4 | 89.0 | 10.0 | – |
| | Llama | No UNSPEC. | 45.0 | 35.0 | 90.0 | 9.0 | – |
| | Llama | **Full protocol** | **99.0** | **69.0** | **97.0** | **88.0** | – |
| DIVERSIFY | GPT-4.1 | No labels | 96.12 | 10.19 | 71.84 | 34.47 | 74.76 |
| | GPT-4.1 | No UNSPEC. | 98.08 | 46.48 | 79.32 | 69.51 | 9.62 |
| | GPT-4.1 | **Full protocol** | **98.10** | **47.60** | **81.20** | **68.90** | **93.20** |
| | Llama-4 | No labels | 98.47 | 30.03 | 36.13 | 39.44 | 0.25 |
| | Llama-4 | No UNSPEC. | 98.44 | 45.88 | 87.00 | 73.94 | 5.35 |
| | Llama-4 | **Full protocol** | **98.10** | **40.00** | **84.00** | **70.00** | **90.20** |

specified, or not reliably captured by face-centric classifiers and contrastive similarity scores. Across attribute prediction and profession grounding, FairJudge is most useful in settings where evaluation requires more than recognizing a dominant visual pattern: religion may depend on visible symbols or attire, disability may depend on assistive devices or embodied context, and profession may be expressed through tools, environments, or activities rather than iconic stereotypes.

A central finding is that the protocol, not merely the MLLM backbone, drives much of the improvement. The ablation results show that removing closed label sets or the UNSPECIFIED option sharply degrades performance, especially for attributes with ambiguous or intermittent evidence. This supports our main claim that reliable MLLM-based evaluation requires structured constraints: fixed taxonomies prevent label drift, abstention reduces unsupported guessing, and strict parsing makes outputs usable for quantitative evaluation. Evidence rationales further support auditability by making it possible to inspect whether decisions are grounded in visible cues, even though such rationales should not be treated as guarantees of causal faithfulness.

The comparison with CLIP, DeepFace, VIEScore, and VQAScore also clarifies the role of FairJudge. CLIP remains strong in some iconic or face-centric settings, such as profession recognition on IDENPROF, but is brittle in context-rich settings and offers no principled abstention. DeepFace is useful for face-visible demographic attributes, but is inherently limited for attributes such as religion, disability, and profession. Recent alignment methods such as VIEScore and VQAScore provide stronger alignment signals than raw CLIP cosine similarity, but do not directly address fairness-specific requirements such as closed social taxonomies, abstention, evidence logging, and slice-level auditing. FairJudge is therefore complementary to existing evaluators: its value lies not only in higher aggregate scores, but in producing structured, inspectable decisions suitable for fairness analysis.

The results also suggest that fairness evaluation should move beyond face-centered benchmarks. Face-visible cues are insufficient for many socially meaningful attributes, and over-reliance on them risks reinforcing narrow or stereotypical notions of identity. DIVERSIFY and DIVERSIFY-PROFESSIONS help address this gap by introducing context-rich images where relevant evidence may appear in clothing, objects, assistive devices, scene composition, or occupational activity. These settings better reflect the kinds of images generated by modern T2I systems and expose failure modes that are less visible in conventional benchmarks.

At the same time, FairJudge has important limitations. MLLM judges may inherit biases from their own training data and may differ in how they interpret ambiguous social cues. Evidence rationales improve transparency but do not prove that the model's decision process is

faithful. The UNSPECIFIED option mitigates forced classification, but abstention behavior itself must be monitored: excessive abstention may hide evaluator weakness, while insufficient abstention may reproduce harmful inference. In addition, our datasets are controlled and synthetic, which enables systematic evaluation but may not capture the full diversity and ambiguity of real-world imagery. Future work should therefore study calibration of abstention, human–MLLM agreement, cross-cultural validity of label taxonomies, and robustness across additional languages, regions, and model families.

Overall, FairJudge argues for treating fairness evaluation as a structured decision problem rather than a single similarity score. For socially sensitive attributes, the evaluator should not only answer *what* is depicted, but also whether the evidence is sufficient, which labels are admissible, and how the decision can be audited. This shift is essential for building T2I evaluation pipelines that are scalable, reproducible, and sensitive to the harms that arise when systems misrepresent or erase social groups.

## 6 Conclusion

We introduced FAIRJUDGE, an abstention-aware evaluation protocol that uses instruction-following MLLMs as structured judges for text-to-image evaluation. By combining closed label sets, explicit UNSPECIFIED abstention, visible-evidence rationales, and strict parsing, FairJudge moves beyond score-only proxies toward auditable judgments for social-attribute prediction, profession grounding, and prompt–image alignment. Empirically, FairJudge improves over contrastive, face-centric, and recent alignment baselines across multiple benchmarks, with particularly strong gains on attributes that require contextual or non-facial evidence, such as religion and disability. Ablations further show that the protocol constraints, rather than MLLM capability alone, are central to reliability.

We also introduced DIVERSIFY and DIVERSIFY-PROFESSIONS, two context-rich resources for evaluating social representation and profession grounding beyond face-visible or iconic cues. These datasets expose failure modes that standard demographic benchmarks can miss and support slice-level analysis of alignment across demographic groups.

Taken all together, our findings suggest that fairness-sensitive T2I evaluation should be framed as a structured decision problem: evaluators should determine not only whether an image matches a prompt, but also whether the relevant evidence is visible, whether a prediction is warranted, and how the judgment can be audited. Important limitations remain: MLLM judges are sensitive to instruction phrasing, can inherit biases from their own training data, and should not be treated as substitutes for human evaluation. To support reproducibility and further scrutiny, we release the datasets, prompts, per-image outputs, rationales, and parser logs for independent auditing and re-scoring.

## Acknowledgements

This work was supported by the Royal Academy of Engineering through the Research Ready programme, founded in partnership with Google DeepMind and The Hg Foundation. The authors also acknowledge support from the OpenAI Researcher Access Program, which provided access to computational resources used in this work.

MP was supported by the UK Engineering and Physical Sciences Research Council via Responsible AI UK (grant number EP/Y009800/1, project KP0016, "AdSoLve: Addressing Socio-technical Limitations of LLMs for Medical and Social Computing"); by the Slovenian Research and Innovation Agency (ARIS) through the Gravitacije project LLM4DH ("Large Language Models for Digital Humanities", GC-0002), the project CroDeCo ("Cross-Lingual Analysis for Detection of Cognitive Impairment in Less-Resourced Languages", J6-60109), and the research programme "Knowledge Technologies" (P2-0103); and by the European Union's Horizon Europe research and innovation programme under grant agreement number 101214398 (ELLIOT). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

ZA is supported by a Google DeepMind PhD Fellowship and thanks her Google DeepMind mentor, David Stutz, for guidance and support.

The authors thank Zainab Khanfer for volunteering as the human annotator for the DIVERSIFY validation study.

## Limitations

**Faithfulness of evidence rationales.** FairJudge requires MLLM judges to cite visible evidence, but these rationales should not be interpreted as faithful causal explanations of the model's internal decision process. Prior work shows that generated explanations can be plausible without being causally faithful (Turpin et al., 2023; Lanham et al., 2023). We therefore treat rationales as audit artifacts: they make judgments more inspectable and easier to review, but they do not guarantee that the cited evidence caused the prediction. Future work should test rationale faithfulness more directly, for example by masking cited regions, perturbing stated evidence, or comparing rationales against human-annotated evidence maps.

**Cost, scalability, and robustness.** FairJudge relies on frontier MLLMs, making it substantially slower and more expensive than embedding-based baselines such as CLIP. Although the protocol requires no training or task-specific adapters, its inference cost may limit use at very large scale. A practical deployment may therefore require cascaded evaluation, where inexpensive metrics triage examples and FairJudge is applied to ambiguous, high-risk, or fairness-sensitive cases. In addition, our prompt-sensitivity analysis covers a limited set of instruction variants. Broader robustness tests across prompt wordings, languages, model versions, decoding settings, and interface-specific formatting constraints are needed.

**Dataset scope and annotation ambiguity.** Our benchmarks include both established datasets and the proposed DIVERSIFY resources, but they cannot capture the full range of visual ambiguity present in real-world T2I outputs. Some social attributes are inherently difficult or inappropriate to infer from images alone, and benchmark labels may reflect prompt metadata, annotator conventions, or dataset-specific taxonomies rather than stable visual constructs. DIVERSIFY and DIVERSIFY-PROFESSIONS are also relatively small compared with large-scale image–text corpora. Consequently, small differences between methods should be interpreted cautiously, especially where we do not report confidence intervals or significance tests.

**Social taxonomies and cultural validity.** FairJudge depends on closed label sets, which improve parseability and reduce taxonomy drift but also impose fixed categories on socially complex attributes. Labels for race, religion, disability, gender, and age may not transfer cleanly across cultural contexts, and some categories may be incomplete, contested, or context-dependent. The UNSPECIFIED option mitigates forced classification, but it does not solve the broader problem of defining socially valid taxonomies. Future work should evaluate culturally localized label sets, participatory annotation protocols, and settings where certain attributes should not be inferred at all.

**Proxy cues and metric dependence.** Evidence requirements reduce unsupported guessing but do not eliminate reliance on proxy cues or stereotypes. For example, clothing, objects, or scene context may be correlated with an attribute without being sufficient evidence for it. Audits should therefore combine quantitative metrics with qualitative rationale review and counterfactual probes. Similarly, alignment scores depend on the chosen rubric and rescaling scheme; alternative rubrics, human panels, or task-specific weighting could change absolute scores and, in some cases, method rankings. FairJudge should therefore be viewed as one component of a broader evaluation pipeline rather than a replacement for human judgment.

## References


Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2025. FairCoT: Enhancing fairness in text-to-image generation via chain of thought reasoning with multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 792–816, Suzhou, China. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Yixiong Chen, Li Liu, and Chris Ding. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pages 20349–20360. IEEE.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290.

Kimmo Kärkkäinen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. 2025. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26256–26264.

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

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2025. Data matters most: Auditing social bias in contrastive vision–language models. *Transactions on Machine Learning Research*.

Sefik Ilkin Serengil. 2024. Deepface: A lightweight face recognition and facial attribute analysis library for python.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.

Zhenting Wang, Shuming Hu, Shiyu Zhao, Xiaowen Lin, Felix Juefei-Xu, Zhuowei Li, Ligong Han, Harihar Subramanyam, Li Chen, Jianfa Chen, and 1 others. 2025. Mllm-as-a-judge for image safety without human labeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14657–14666. IEEE.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

## A F1 and Group-Macro-F1 Results

Table 5 reports macro-F1 for social-attribute prediction. Unlike accuracy, macro-F1 gives equal weight to each class and is therefore more sensitive to failures on underrepresented categories. This is important for fairness-sensitive evaluation, where high aggregate accuracy may mask poor performance on minority classes.

The macro-F1 results broadly support the accuracy trends in the main paper. MLLM judges outperform CLIP and DeepFace on most dataset–attribute pairs, with the largest

**Table 5:** Macro-F1 for social-attribute prediction. Macro-F1 averages per-class F1 without frequency weighting, making it sensitive to underrepresented classes.

| Dataset | Method | Gender | Race | Age | Relig. | Disab. |
|---|---|---|---|---|---|---|
| FairFace | CLIP | 0.94 | 0.62 | 0.38 | – | – |
| FairFace | GPT | **0.95** | **0.86** | **0.69** | – | – |
| FairFace | Gemini | 0.95 | 0.45 | 0.22 | – | – |
| FairFace | Llama | 0.87 | 0.45 | 0.61 | – | – |
| PaTA | DeepFace | 0.58 | 0.41 | 0.58 | – | – |
| PaTA | CLIP | 0.41 | 0.41 | 0.56 | – | – |
| PaTA | GPT | 0.50 | 0.74 | 0.64 | – | – |
| PaTA | Gemini | **1.00** | **0.94** | **0.98** | – | – |
| PaTA | Llama | 0.50 | 0.64 | 0.93 | – | – |
| FairCoT | DeepFace | 0.62 | 0.39 | 0.46 | – | – |
| FairCoT | CLIP | 0.62 | 0.36 | 0.35 | 0.35 | – |
| FairCoT | GPT | **0.99** | 0.67 | 0.93 | 0.93 | – |
| FairCoT | Gemini | 0.99 | **0.81** | 0.93 | **0.96** | – |
| FairCoT | Llama | 0.66 | 0.69 | **0.95** | 0.65 | – |
| DIVERSIFY | CLIP | 0.86 | 0.21 | 0.60 | 0.40 | 0.08 |
| DIVERSIFY | GPT | 0.99 | 0.40 | 0.82 | 0.68 | 0.65 |
| DIVERSIFY | Gemini | **0.99** | **0.47** | **0.96** | **0.76** | **0.66** |
| DIVERSIFY | Llama | 0.98 | 0.26 | 0.86 | 0.65 | 0.23 |

**Table 6:** Unweighted Group-Macro-F1 (GMF1) for profession prediction, disaggregated by protected attribute. Each cell averages macro-F1 across demographic groups within the corresponding attribute.

| Dataset | Method | Age | Gender | Race | Relig. | Culture |
|---|---|---|---|---|---|---|
| DIV-Prof | CLIP | 0.46 | 0.35 | 0.35 | 0.27 | – |
| DIV-Prof | GPT | **0.68** | **0.54** | 0.63 | **0.48** | 0.30 |
| DIV-Prof | Gemini | 0.67 | 0.52 | **0.64** | 0.47 | **0.31** |
| DIV-Prof | Llama | 0.45 | 0.34 | 0.36 | 0.26 | – |
| FC-Prof | CLIP | 0.18 | 0.21 | 0.27 | 0.30 | – |
| FC-Prof | GPT | 0.23 | 0.28 | 0.38 | **0.42** | – |
| FC-Prof | Gemini | 0.23 | 0.28 | 0.38 | 0.42 | – |
| FC-Prof | Llama | **0.56** | **0.40** | **0.50** | 0.33 | – |

gains on context-dependent and imbalanced attributes. The gap is especially pronounced for religion and disability on DIVERSIFY, where CLIP obtains low macro-F1 despite non-trivial accuracy, indicating weak performance on minority classes. Race remains challenging across datasets, suggesting that structured judging improves but does not eliminate the difficulty of evaluating socially constructed and visually ambiguous categories.

Table 6 reports unweighted Group-Macro-F1 (GMF1) for profession prediction. For each protected attribute, we compute macro-F1 over profession classes separately for each demographic group, then average uniformly across groups. This prevents larger demographic groups from dominating the aggregate score and better reflects whether profession recognition is consistent across slices.

Across both profession benchmarks, MLLM judges generally improve GMF1 over CLIP, but the strongest model differs by dataset and protected attribute. On DIVERSIFY-PROFESSIONS, GPT obtains the highest GMF1 for age, gender, and religion, while Gemini is slightly higher for race and culture. The updated GPT and Gemini rows are no longer identical, reflecting the use of distinct profession-prediction outputs. On FAIRCOT-PROFESSIONS, the weighted GMF1 values are close to the corresponding top-1 profession accuracies, but the unweighted GMF1 values are substantially lower because some demographic slices are small or sparse. This indicates that the models perform well on larger slices but less consistently across all demographic groups. Llama remains strongest on age, gender, and race in this unweighted slice-level view, while GPT and Gemini are strongest on religion. Overall, GMF1 reveals fairness-sensitive variation that is not visible from aggregate accuracy alone, and should therefore be interpreted alongside the per-group results and qualitative error analysis.

## B UNSPECIFIED Prediction Rates and Interpretation of Abstention

FairJudge explicitly permits an UNSPECIFIED prediction when the available visual evidence is insufficient to support one of the substantive attribute labels. We therefore analyze abstention behavior separately from standard accuracy and macro-F1. This distinction is important because an UNSPECIFIED output is not necessarily an evaluation failure: for benchmarks in which the ground-truth taxonomy itself contains a neutral/UNSPECIFIED class, abstention can be the correct prediction.

**Abstention definitions.** Let $\hat{y}_i$ denote the judge prediction for example $i$, and let $y_i$ denote its ground-truth label. We report two related quantities. The *raw abstention rate* is

$$A_{\text{raw}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i = \textsc{Unspecified}\right], \tag{1}$$

which measures how often a judge declines to assign a substantive label. For attributes whose reference taxonomy also contains neutral/UNSPECIFIED, we additionally report the *unmatched abstention rate*,

$$A_{\text{unmatched}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i = \textsc{Unspecified} \wedge y_i \neq \textsc{Unspecified}\right]. \tag{2}$$

Thus, when both the prediction and reference label are neutral/UNSPECIFIED, the prediction is treated as a *matched abstention* rather than an error. We normalize explicit variants such as `neutral`, `unknown`, and `unspecified` to the abstention

category. Empty or otherwise non-responsive outputs are included in the raw abstention count, but are not credited as a correct matched abstention solely because the ground-truth label is neutral.

Table 7 reports the unmatched abstention rate, which is the most relevant quantity when interpreting abstention as a coverage failure. A dash indicates that the corresponding attribute is not available for that benchmark.

**Raw versus matched abstention.** The difference between raw and unmatched abstention is particularly important for religion and disability. These attributes frequently lack a visually warranted positive label by construction. Table 8 therefore decomposes the raw rate into matched and unmatched components for the benchmark–attribute pairs with explicit neutral/UNSPECIFIED ground truth.

**Interpretation.** These results highlight why abstention should be interpreted as a *selective prediction behavior*, rather than as missing output. On FairFace, Gemini almost never abstains, GPT-4.1 abstains on roughly 1–1.5% of examples, whereas Llama-4 abstains on approximately 17% of gender, race, and age predictions. PaTA yields substantially lower abstention rates for all three judges. These benchmark differences indicate that abstention is sensitive both to the judge backbone and to the visual evidence available in the evaluation set.

The distinction is even more important for context-dependent attributes. For example, the raw FairCoT religion abstention rate ranges from 57.11% to 61.32%. Taken alone, these values could suggest that the judges fail to make a religion prediction for most images. However, approximately 54.71% of all evaluable FairCoT examples are cases in which each judge's neutral prediction correctly agrees with the neutral ground truth. The corresponding unmatched rates are therefore only 2.40% for Llama-4, 3.01% for Gemini 1.5 Pro, and 6.61% for GPT-4.1.

A similar effect occurs for disability on DIVERSIFY. Raw abstention rates are very high (85.50–90.62%), reflecting the large number of examples without a specified disability cue. After crediting abstentions that agree with the reference UNSPECIFIED label, the unmatched rates fall to 3.41% for Gemini, 3.84% for GPT-4.1, and 5.76% for Llama-4. Religion remains more difficult: unmatched abstention reaches 11.09% for Gemini, 21.11% for GPT-4.1, and 26.01% for Llama-4. Llama-4 also exhibits a notably high unmatched abstention rate for race on DIVERSIFY (32.84%), indicating a separate coverage limitation on that attribute.

**Abstention is not monotonically better or worse.** A lower abstention rate should not automatically be interpreted as a better judge. A model that never abstains may achieve high coverage by making unsupported or stereotype-driven inferences when visual evidence is weak. Conversely, a judge that abstains too frequently may avoid errors by declining cases that contain sufficient evidence. Reliable fairness-sensitive evaluation therefore requires balancing *coverage* against *evidential support*. This is consistent with the prompt-sensitivity results in Appendix B: strengthening the evidentiary threshold increases abstention on difficult attributes such as race and religion, which can reduce conventional top-1 accuracy even when withholding a prediction is preferable to unsupported inference.

Accordingly, we recommend interpreting accuracy and macro-F1 jointly with both raw and unmatched abstention rates. Raw abstention characterizes the judge's overall conservatism, matched abstention measures appropriate withholding of unsupported labels, and unmatched abstention identifies cases in which conservatism reduces effective coverage despite a substantive ground-truth label. Reporting all three quantities makes the trade-off explicit and avoids conflating principled abstention with prediction failure. "'

## C Prompt Sensitivity Analysis

We evaluate the sensitivity of FairJudge to instruction phrasing on DIVERSIFY using GPT-4.1. This experiment is intended as an initial robustness check rather than a comprehensive prompt-invariance study. We compare the original FairJudge prompt (P0) against three variants that preserve the same task, output schema, and label sets while modifying the instructions.

**Prompt variants. P1: Stronger abstention** increases the evidentiary threshold for prediction by requiring "clear and direct visual evidence" and explicitly discouraging stereotype-based inference. **P2: Evidence-first** asks the model to state visible evidence for each attribute before producing the final label, testing whether rationale ordering

| Dataset | Judge | Gender | Race | Age | Religion | Disability |
|---|---|---|---|---|---|---|
| FairFace | GPT-4.1 | 1.41 | 1.40 | 1.36 | – | – |
| | Gemini 1.5 Pro | 0.00 | 0.00 | 0.00 | – | – |
| | Llama-4 | 16.96 | 17.28 | 16.72 | – | – |
| PaTA | GPT-4.1 | 1.06 | 1.87 | 1.06 | – | – |
| | Gemini 1.5 Pro | 0.23 | 0.28 | 0.68 | – | – |
| | Llama-4 | 0.20 | 0.20 | 0.20 | – | – |
| FairCoT | GPT-4.1 | 0.40 | 0.40 | 0.40 | 6.61 | – |
| | Gemini 1.5 Pro | 0.00 | 0.00 | 0.00 | 3.01 | – |
| | Llama-4 | 0.00 | 0.00 | 0.00 | 2.40 | – |
| DIVERSIFY | GPT-4.1 | 0.00 | 0.00 | 0.00 | 21.11 | 3.84 |
| | Gemini 1.5 Pro | 0.00 | 0.00 | 0.00 | 11.09 | 3.41 |
| | Llama-4 | 0.43 | 32.84 | 0.21 | 26.01 | 5.76 |

**Table 7:** Unmatched UNSPECIFIED prediction rates (%) across FairJudge backbones and attribute-prediction benchmarks. An abstention is counted here only when the judge predicts UNSPECIFIED while the ground-truth label specifies a substantive category. Consequently, ground-truth neutral/UNSPECIFIED cases that are correctly recognized as such are not counted as errors. FairCoT religion is evaluated on the 499 examples for which a religion reference label is available.

| Dataset / Attr. | Judge | Raw | Matched | Unmatched |
|---|---|---|---|---|
| FairCoT / Relig. | GPT-4.1 | 61.32 | 54.71 | 6.61 |
| | Gemini 1.5 Pro | 57.72 | 54.71 | 3.01 |
| | Llama-4 | 57.11 | 54.71 | 2.40 |
| DIVERSIFY / Relig. | GPT-4.1 | 46.70 | 25.59 | 21.11 |
| | Gemini 1.5 Pro | 36.03 | 24.95 | 11.09 |
| | Llama-4 | 50.96 | 24.95 | 26.01 |
| DIVERSIFY / Disab. | GPT-4.1 | 88.06 | 84.22 | 3.84 |
| | Gemini 1.5 Pro | 85.50 | 82.09 | 3.41 |
| | Llama-4 | 90.62 | 84.86 | 5.76 |

**Table 8:** Decomposition of abstention rates (%) for attributes whose ground-truth taxonomy includes neutral/UNSPECIFIED. "Raw" is the overall frequency of abstention, "Matched" is the frequency with which the abstention agrees with a neutral/UNSPECIFIED reference label, and "Unmatched" is their difference. The large raw rates for religion and disability should therefore not be interpreted directly as failure rates.

affects decisions. **P3: Label-order perturbation** shuffles the order of allowed labels while preserving the same output keys, testing sensitivity to label position.

**Table 9:** Prompt sensitivity on DIVERSIFY using GPT-4.1. All variants produce 0% invalid JSON outputs.

| Prompt | Gender | Race | Age | Relig. | Disab. |
|---|---|---|---|---|---|
| P0: Original | 0.98 | 0.48 | 0.81 | 0.69 | 0.93 |
| P1: Stronger abst. | 0.96 | 0.24 | 0.76 | 0.19 | 0.92 |
| P2: Evidence-first | 0.96 | 0.30 | 0.77 | 0.36 | 0.92 |
| P3: Label-order | 0.97 | 0.21 | 0.83 | 0.26 | 0.92 |

All prompt variants yield valid JSON for all 469 examples, indicating that the strict output schema is robust to moderate instruction changes. Performance is also relatively stable for gender and disability: both remain within a narrow range across prompt variants. Age shows moderate variation, with the label-order variant slightly improving over the original prompt.

Race and religion are more sensitive to prompt phrasing. The stronger-abstention prompt substantially reduces accuracy for both attributes, largely because the model abstains more often under ambiguity. This behavior is consistent with the intended effect of P1, but it also highlights a limitation of accuracy as the only measure: abstentions are counted as errors even when withholding a prediction may be preferable for fairness-sensitive evaluation. The evidence-first variant yields intermediate performance, suggesting that requiring explicit evidence before prediction makes the model more conservative without fully collapsing to UNSPECIFIED. The label-order perturbation also affects race and religion, indicating residual sensitivity to the ordering of candidate labels.

These findings support two conclusions. First, the FairJudge protocol is most stable for attributes with clearer visual grounding, such as gender and disability in this dataset. Second, context-dependent attributes such as race and religion require careful prompt design and should be evaluated with abstention-aware metrics in addition to standard accuracy. In future work, label-order randomization, prompt ensembling, and explicit abstention calibration may reduce sensitivity further.

# D Dataset Contribution and Datasheet: DIVERSIFY

## D.1 Overview

DIVERSIFY is a synthetic image corpus for fairness evaluation *beyond* strictly face-visible cues. Each image is rendered from a text

prompt and paired with structured metadata for five labeled targets: *gender*, *race*, *age group*, *religion*, and *disability*. Prompts are composed as *base scenes* (e.g., indoor/outdoor, public/private, formal/informal) with systematic *variants* that toggle one factor at a time, encouraging controlled analyses rather than opportunistic mining of web photos. The dataset is designed to surface context-dependent evidence (attire, objects, signage, assistive devices, and setting) where simple face heuristics are insufficient.

**Exploratory contextual diversity.** In addition to the labeled dimensions used in the main experiments, some prompts include broader scene and contextual variation intended to increase cultural diversity of depictions. We do *not* treat these variations as anthropologically grounded culture labels, and we do not use them as image-level ground truth in the main paper. They are included only as exploratory coverage to broaden the range of settings represented in the corpus.

**Intended use.** Primary uses include: (i) measuring recognition and abstention behavior of evaluators (CLIP and MLLM judges) on attributes with weak pixel evidence; (ii) auditing spurious correlations between demographics and contextual cues; and (iii) running counterfactual tests by swapping a single attribute while holding the rest constant. The corpus is *not* intended for training identity recognizers or for inferring attributes of real people.

### D.2 Composition and Statistics

Each example has a unique prompt ID, the rendered image, and normalized labels parsed from prompt text. Demographic attributes (gender, race, age) are single-label; religion is single-label with UNSPECIFIED; disability is single-label with UNSPECIFIED. Below we provide summary tables for the labeled dimensions used in the main experiments. Counts depend on the specific release split and can be filled from the provided metadata CSV.[2]

**Label provenance.** All reported labels are *text-derived* from the prompt templates; no human annotation of rendered pixels is required. To reduce labeling drift, we canonicalize descriptors to a closed taxonomy and map absent evidence to UNSPECIFIED. Any broader contextual descriptors used during prompt construction are retained, if at all, only as prompt-level metadata and are not treated as validated image labels.

[2]We release per-image metadata with canonicalized labels and prompt text; see §D.3.

**Table 10: Gender distribution** ($N = 469$).

| Label | Count | % |
|---|---|---|
| Male | 234 | 49.9% |
| Female | 235 | 50.1% |
| Unspecified | 0 | 0.0% |

**Table 11: Race distribution (canonicalized)** ($N = 469$). Indian → South Asian.

| Label | Count | % |
|---|---|---|
| Black | 109 | 23.2% |
| White | 167 | 35.6% |
| East Asian | 93 | 19.8% |
| South Asian | 4 | 0.9% |
| Middle Eastern | 45 | 9.6% |
| Latino/Hispanic | 48 | 10.2% |
| Indigenous | 0 | 0.0% |
| Southeast Asian | 3 | 0.6% |

### D.3 Data Generation and Quality Control

**Prompting scheme.** We construct base prompts for diverse scenes (home, street, workspace, worship, nature) and attach *single-factor variants* that toggle one attribute at a time (gender, race, age group, religion symbol, disability aid) while holding the rest fixed. This design supports counterfactual swaps under matched backgrounds.

**Parsing and validation.** We deterministically parse labels from prompt text and normalize them to the target taxonomy. We block contradictory combinations (e.g., multiple single-label values), reject near-duplicates via perceptual hashing, and run image-level sanity checks for the presence or absence of specified objects or symbols.

**Splits.** We provide stratified {train, dev, test} splits that preserve the marginal distributions of the single-label attributes. Prompt IDs and seeds allow exact regeneration.

### D.4 Recommended Uses and Tasks

DIVERSIFY enables: (1) fairness audits for attributes that are weakly encoded in pixels (religion, disability); (2) evaluation of *abstention* and evidence use by MLLM judges vs. contrastive scorers; (3) counterfactual probes by swapping a single attribute with the scene held constant; and (4) robustness studies under prompt-template

**Table 12: Age groups** ($N = 469$).

| Label | Count | % |
|---|---|---|
| Child | 0 | 0.0% |
| Young adult | 157 | 33.5% |
| Middle-aged | 151 | 32.2% |
| Elderly | 151 | 32.2% |
| Unspecified | 10 | 2.1% |

**Table 13: Religion distribution** ($N = 469$).

| Label | Count | % |
|---|---|---|
| Unspecified | 130 | 27.7% |
| Christian | 47 | 10.0% |
| Muslim | 54 | 11.5% |
| Hindu | 46 | 9.8% |
| Buddhist | 49 | 10.4% |
| Jewish | 57 | 12.2% |
| Sikh | 54 | 11.5% |
| Shinto | 32 | 6.8% |

perturbations.

### D.5 Ethical Considerations and Risks

**Synthetic, not personal data.** Labels refer to *intended depictions* in generated images and must not be conflated with real identities. **Stereotype risk.** Some combinations may mirror societal stereotypes despite careful prompt design; we encourage per-attribute reporting and qualitative review of failures. **Contextual sensitivity.** Broader contextual variation in prompts is necessarily reductive and should not be interpreted as a normative or anthropological taxonomy. Any such descriptors are included only to diversify scenes, not to support claims about cultural identity.

### D.6 Limitations

First, labels are prompt-derived; generators may fail to render requested cues, leading to label–image mismatch. Second, taxonomies are coarse by design; many identities and intersections are absent or underrepresented. Third, broader contextual descriptors in prompts reflect scene design choices rather than validated social labels. Finally, generator style and priors introduce artifacts that may confound downstream evaluation.

### D.7 Licensing and Availability

We release prompts, canonicalized labels, and split files under a permissive license, with scripts to regenerate images subject to the base model's terms.[3] Metadata include prompt IDs, normalized labels, and generation seeds to facilitate exact reproduction and re-scoring under alternative taxonomies.

[3]License text and regeneration instructions are provided in the repository.

**Table 14: Disability labels** ($N = 469$).

| Label | Count | % |
|---|---|---|
| Unspecified | 409 | 87.2% |
| Mobility impairment | 17 | 3.6% |
| Blind/low vision | 10 | 2.1% |
| Down syndrome | 10 | 2.1% |
| Amputee | 6 | 1.3% |
| Deaf/hard of hearing | 4 | 0.9% |
| Dwarfism | 4 | 0.9% |
| Autism | 4 | 0.9% |
| Vitiligo | 2 | 0.4% |
| Albinism | 2 | 0.4% |
| Cerebral palsy | 1 | 0.2% |

## E Dataset Contribution and Datasheet: DIVERSIFY-Professions

### E.1 Overview

We contribute a synthetic image dataset for profession-focused fairness analysis. The corpus comprises **1,200** images generated from text prompts that explicitly vary *profession*, *gender*, *race*, *age group*, and *religion*, with additional diversity in scene and contextual presentation. Prompts are organized as a small set of *base prompts* with systematic *variants* that control a single factor at a time, enabling controlled evaluation of representation biases in multimodal models.

**Intended use.** The dataset is designed for: (i) measuring demographic skews in model outputs conditioned on profession; (ii) stress-testing captioning, retrieval, and classification models for spurious correlations; and (iii) evaluating intervention methods (e.g., debiasing prompts or reweighting) under matched visual conditions.

### E.2 Composition and Statistics

Each image is associated with structured metadata parsed from prompt text. Table 15 shows the profession coverage; Tables 16–17 summarize the labeled social attributes used in the paper.

**Exploratory contextual diversity.** As with DIVERSIFY, some prompts also include broader contextual variation intended to diversify depictions across scenes and settings. We do not treat these prompt-level descriptors as validated image labels, and we do not evaluate them as a separate target dimension.

**Table 15:** Profession coverage (single label; $N$=1,200).

| Profession | Count | % |
|---|---|---|
| Doctor | 200 | 16.7% |
| Engineer | 200 | 16.7% |
| Janitor | 200 | 16.7% |
| Lawyer | 200 | 16.7% |
| Nurse | 200 | 16.7% |
| Teacher | 200 | 16.7% |

**Table 16:** Gender and race distributions (single label; $N$=1,200).

| **Gender** | | |
|---|---|---|
| Male | 498 | 41.5% |
| Female | 487 | 40.6% |
| Unspecified | 215 | 17.9% |
| **Race** | | |
| Black | 421 | 35.1% |
| White | 403 | 33.6% |
| South Asian | 140 | 11.7% |
| Middle Eastern | 115 | 9.6% |
| East Asian | 86 | 7.2% |
| Hispanic/Latino | 19 | 1.6% |
| Native American | 7 | 0.6% |
| Unspecified | 9 | 0.8% |

**Table 17:** Age groups and religion (single label; $N$=1,200).

| **Age group** | | |
|---|---|---|
| Young adult | 343 | 28.6% |
| Middle-aged | 374 | 31.2% |
| Senior | 357 | 29.8% |
| Unspecified | 126 | 10.5% |
| **Religion** | | |
| Christian | 226 | 18.8% |
| Muslim | 230 | 19.2% |
| Hindu | 264 | 22.0% |
| Sikh | 89 | 7.4% |
| Buddhist | 55 | 4.6% |
| Jewish | 10 | 0.8% |
| Taoist | 6 | 0.5% |
| Unspecified | 320 | 26.7% |

**Label provenance.** All reported labels are *text-derived* from prompts; no manual annotation of the rendered images was performed. Any broader contextual descriptors that appear in prompt templates are treated, if retained, as prompt metadata rather than image-level ground truth.

### E.3 Data Generation and Quality Control

**Prompting scheme.** We start from a profession-specific base prompt (e.g., *"Base prompt (doctor)"*) and create controlled variants (e.g., *V1p1(doctor)*, *V1p2(doctor)*) to toggle a single attribute (gender, race, age group, religion). **Parsing and validation.** We deterministically extract labels from prompt text using pattern matching over explicit descriptors (*male/female*, *Black/White/East Asian/...*, *young adult/middle-aged/senior*, *Christian/Muslim/Hindu/...*) and canonicalize them to a closed taxonomy. We perform consistency checks to ensure exactly one profession per prompt and at most one label per single-label attribute.

### E.4 Recommended Uses and Tasks

This dataset supports: (1) profession-conditioned fairness probes (e.g., are nurses depicted differently by gender or race?); (2) auditing retrieval/captioning models for spurious attribute leakage; and (3) counterfactual testing by swapping a single attribute while holding the rest constant.

### E.5 Ethical Considerations and Risks

**Synthetic images are not people.** Demographic labels describe *intended* depictions in generated images; they must not be conflated with real identities. **Stereotype risk.** Some attribute–profession combinations may mirror societal stereotypes. We encourage reporting performance disaggregated by *both* profession and demographic attribute, and avoiding aggregate-only metrics. **Context sensitivity.** Broader contextual descriptors used during prompt construction are reductive and should not be interpreted as validated social identity labels.

### E.6 Limitations

(i) Labels are derived from text prompts rather than human verification of image content; the generator may fail to render the intended attribute. (ii) Taxonomies are coarse and not exhaustive; some identities are absent or underrepresented. (iii) Broader contextual variation reflects prompt design rather than validated image-level labels. (iv) The visual style and biases of the underlying image generator may introduce artifacts unrelated to the prompts.

### E.7 Licensing and Availability

We release the prompts and metadata (`.csv/.json`) under a permissive license and provide scripts to regenerate the images where permitted by the image model's license.[4] To

[4] Exact license text and regeneration instructions included in the project repository.

facilitate reproducibility, we include prompt IDs, attribute labels, and (where applicable) generation seeds.

### E.8 Reproducibility Checklist

- **Files:** `prompts.txt`, `metadata.csv` (`id`, `profession`, `gender`, `race`, `age_group`, `religion`), and a datasheet (`.xlsx`).
- **Counts:** $N$=1,200 images; 6 professions with 200 images each.
- **Determinism:** Prompt → label mapping is rule-based and released; image seeds provided when licensing allows.
- **Splits:** Deterministic train/dev/test split seeds provided for reproducible evaluation.

## F FairJudge Prompt Templates

All prompts enforce closed label sets, require visible evidence, and allow a neutral fallback (UNSPECIFIED/NEUTRAL). We parse only the structured fields during scoring; rationales are logged for audit.

**Attribute prediction (single prompt shared across models).**

**Listing 1:** Attribute prediction prompt (label-constrained with abstention).

```
You are given an image of a person. Classify the
following using ONLY visible evidence. If evidence
is insufficient for any field, output "unspecified".
Labels: - Gender: {male, female, unspecified}
- Race: {Black, White, East Asian, South Asian,
Middle Eastern, Latino_Hispanic, Indigenous,
unspecified} - Age: {child, young adult,
middle-aged, elderly, unspecified} - Religion:
{Christian, Muslim, Hindu, Buddhist, Jewish,
Sikh, Shinto, Neutral/unspecified} - Disability:
{mobility impairment, blind/low vision, deaf/hard of
hearing, dwarfism, vitiligo, unspecified}
Output STRICT JSON (no extra text), with these keys
exactly: {"Gender": "...", "Race": "...", "Age":
"...", "Religion": "...", "Disability": "..."}
```

**Alignment judging (rubric → scalar $[-1, 1]$).**

**Listing 2:** Prompt–image alignment rubric (later rescaled to $[-1, 1]$).

```
Task: Rate how well the image matches the
description.
Description: "<CAPTION>"
Use this 1-5 rubric: 1 = Not a match at all 2
= Significant discrepancies 3 = Several minor
discrepancies 4 = A few minor discrepancies 5 =
Matches exactly
Answer in STRICT JSON (no extra text): {"Alignment
analysis": "«=100 words, evidence-based»",
"Alignment score": "<integer 1-5>"}
```

**Profession accuracy + alignment (optional combined prompt).**

**Table 18:** Cost/speed per image

| Model | Time / image | Cost / image |
|---|---|---|
| Llama-4 | 4.73 s | $0.0817 |
| CLIP | 0.21 s | open-source |
| GPT-4.1 | 4.47 s | $0.0025 |
| Gemini 1.5 Pro | 14.42 s | $0.0017 |

**Listing 3:** Combined profession + alignment judging.

```
You are given an image and a candidate profession.
First, decide if the image shows that profession.
Then rate alignment using the 1-5 rubric above.
Base both decisions ONLY on visible evidence.
Input: Profession = "<PROFESSION>" Caption = "a
photo of a <PROFESSION>"
Output STRICT JSON: {"Profession prediction":
"<yes|no|unspecified>", "Prediction evidence":
"«=50 words»", "Alignment analysis": "«=100
words»", "Alignment score": "<integer 1-5>"}
```

*Implementation notes.* We keep wording identical across models (GPT-4.1, Gemini 1.5 Pro, Llama-4). At evaluation, we parse the JSON fields only and linearly rescale the 1–5 alignment rating to $\bar{s} \in [-1, 1]$ for comparison with CLIP cosine; minimal template perturbations are reported in Sec. 4.4.

### F.1 Computational Cost

We report wall-clock time per image and marginal API cost per image exactly as recorded in our logs.

CLIP offers unmatched throughput and negligible marginal cost but provides the least calibrated fairness/faithfulness signal for our purposes, motivating comparison with MLLM judges. Among judges, GPT-4.1 achieves a strong quality–cost trade-off (low per-image cost and moderate latency), while Gemini 1.5 Pro is the cheapest per image but slower in practice. Under our setup, Llama-4 is the most costly because we accessed it via the DeepInfra API; if self-hosted, the *marginal* API cost would be zero but GPU-hours would be substantial. A pragmatic deployment follows a **cascaded judging** design: route routine cases to CLIP or another cheap metric for broad screening, escalate uncertain or fairness-sensitive cases to a cost-effective judge (e.g., GPT-4.1 or Gemini 1.5 Pro), and invoke a second MLLM opinion only for disagreements or high-stakes audits. This preserves most of the accuracy and alignment gains while controlling runtime and cost. We emphasize that FairJudge is implementation-lightweight—training-free, adapter-free, retrieval-free, requiring only a prompt and parser—but it is *not* computationally lightweight compared with CLIP, and should not be used for very large-scale routine scoring where proxy metrics suffice.

Across attributes, non-demographic categories,

and alignment, instruction-following judges offer a practical path to *accountable* evaluation: higher accuracy where cues are visible, better grounding where cues are subtle, and stable rankings under small prompt changes.

## G Human Validation of DIVERSIFY Annotations

To complement the automated evaluation, we conduct a targeted human validation of a 100-image subset of DIVERSIFY. The purpose of this analysis is to assess whether the benchmark annotations are supported by the visual evidence available in the generated images, particularly for attributes such as religion and disability, where prompt-derived labels may not always be visually recoverable. This analysis is intended as a consistency check rather than a full re-annotation of the benchmark.

**Annotator.** The subset was independently reviewed by one volunteer annotator who self-identified as female, Muslim, Arab and African, non-disabled, and having a background in law. We report these characteristics for transparency, as judgments about socially salient attributes may depend in part on annotator background and cultural context. The annotator was not involved in producing the automated FairJudge predictions evaluated in the main experiments.

**Annotation protocol.** Prior to annotation, the volunteer was provided with the DIVERSIFY visual annotation guide. The protocol instructed the annotator to assign labels only when they were supported by concrete, observable visual evidence, and to select UNSPECIFIED when the evidence was ambiguous or insufficient. This mirrors the evidence-based and abstention-aware principles used by FairJudge.

For religion, the annotation guide emphasized distinctive visible cues, including religious symbols, attire, ritual objects, and worship contexts, while explicitly discouraging inference from ethnicity, national dress, or other ambiguous contextual signals. For disability, the annotator was instructed to rely on directly observable evidence, such as assistive devices or clear body/equipment cues, and to avoid inferring medical conditions from facial appearance, posture, or other non-diagnostic visual characteristics.

**Validation procedure.** For each of the 100 sampled images, the annotator assigned labels using the same closed attribute taxonomy as DIVERSIFY. The resulting human annotations were then compared with the corresponding benchmark labels. We use this comparison to identify cases in which the benchmark label is not clearly supported by the rendered image, as well as cases in which a human reviewer considers UNSPECIFIED to be the more appropriate label.

Because this analysis uses a single annotator and a limited subset of the benchmark, we do not interpret it as an estimate of inter-annotator reliability or population-level annotation quality. Instead, it provides an independent check of annotation consistency under the same conservative visible-evidence criterion used throughout FairJudge. Establishing broader annotation reliability would require multiple annotators and an explicit adjudication protocol.

**Relation to FairJudge.** This human validation is particularly informative for religion and disability, where benchmark metadata and visible image content may diverge. It therefore provides a direct check of a central design principle of FairJudge: socially salient attributes should be assigned only when the image provides sufficient supporting evidence, and the evaluator should otherwise abstain.