FairJudge: Calibrated Multimodal Judges for Fairness and Alignment Evaluation in Text-to-Image Models

Anonymous ACL submission

Abstract

Text-to-image (T2I) systems lack simple, reproducible ways to evaluate prompt–image alignment and social attributes. Conventional pipelines rely on face-centric recognizers or contrastive similarity scores as proxies for these judgments. Such proxies can fail when target attributes are only weakly encoded in pixels, when confounding context is present, or when uncertainty should be expressed through abstention. As a result, current fairness evaluations can overstate model capability or miss context-dependent failure modes.

We introduce FairJudge, a lightweight protocol that treats instruction-following multimodal LLMs as *fair judges* for T2I evaluation. FairJudge constrains outputs to closed label sets, requires explicit visual evidence, allows abstention when cues are insufficient, and scores prompt–image alignment with an explanation-oriented rubric mapped to $[-1, 1]$. Unlike CLIP-only pipelines, it produces accountable, evidence-aware judgments and targets *evaluation fairness* rather than generation-time mitigation.

We evaluate demographic attributes on standard benchmarks, assess religion and disability beyond strictly face-visible cues, and measure profession correctness and prompt–image alignment on both existing and newly introduced datasets.Across datasets, judge models consistently outperform contrastive and face-centric baselines on demographic prediction and improve alignment assessment while maintaining high profession accuracy. We also introduce DIVERSIFY and DIVERSIFY-PROFESSIONS, two resources that broaden diversity-oriented evaluation of gender, race, age, religion, disability, and profession grounding in non-iconic, context-rich settings. Upon acceptance, we will release code, prompts, datasets, parser logs, and per-image judge outputs to support reproducible audits.

1 Introduction

Text-to-image (T2I) systems are increasingly used to generate and retrieve people-centric imagery in creative, assistive, and analytical settings. Yet these models inherit and can amplify social biases from their data and training objectives, yielding skewed portrayals across gender, skin tone, religion, disability, and professions. Conventional evaluation pipelines rely heavily on face-centric attribute recognizers or contrastive similarity scores to stand in for human judgment. Such proxies can fail when the target attribute is only weakly encoded in pixels, when confounding context is present (e.g., clothing, scene props), or when the appropriate decision is to abstain. As a result, fairness assessments may overstate model capability or obscure failure modes that would be obvious to a careful human reviewer.

Large-scale studies have documented representation disparities in T2I and vision–language models and have introduced diagnostic tests and scorecards for auditing them (Luccioni et al., 2023; Wan et al., 2024). Remedial approaches range from lightweight sampling control to small learned adapters. *Fair Diffusion* steers the sampler toward user-specified demographic proportions without retraining the backbone (Friedrich et al., 2023); *Fair Mapping* inserts a linear adapter to systematically reduce demographic skew with minimal compute (Li et al., 2025). Generation-time prompting with chain-of-thought (CoT), as in *FairCoT*, leverages reasoning to mitigate bias during image synthesis (Al Sahili et al., 2025b). While effective for *mitigation*, these approaches operate on the *generator*; they do not provide an evaluation protocol that is calibrated, evidence-grounded, and appropriately abstaining. On the evaluation side, contrastive similarity (e.g., CLIP) remains a dominant proxy for text–image faithfulness

(Radford et al., 2021; Hessel et al., 2021), and VQA-style checklists probe finer-grained grounding (Hu et al., 2023), but both can reward surface correlation and penalize uncertainty. Preference-trained rankers (IMAGEREWARD, PICKSCORE) improve agreement with human choices (Xu et al., 2023; Kirstain et al., 2023) yet remain scalar scorers with limited transparency. Recent "LLM-as-a-judge" work reports high agreement with human raters but also reveals position and verbosity biases, motivating careful protocol design and meta-evaluation (Zheng et al., 2023; Chen et al., 2024; Wang et al., 2025). In short, we lack a lightweight, *evaluation*-oriented procedure that (i) reasons explicitly about visual evidence, (ii) adheres to label constraints, (iii) abstains when cues are insufficient, and (iv) applies beyond face-visible demographics to more contextual attributes such as religion, disability, and profession.

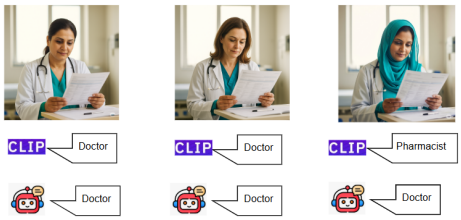


Figure 1: CLIP is brittle for profession recognition; MLLM judges are more consistent. Each panel shows the same profession under appearance and contextual variation.

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**, a lightweight, label-constrained, abstention-capable judging protocol for social-attribute prediction 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.

Taken together, our results indicate that calibrated MLLM judges narrow long-standing gaps in T2I evaluation by improving assessment of non-facial and context-dependent attributes while producing interpretable alignment scores. The resulting protocol remains lightweight, auditable, and easy to reproduce across rapidly evolving models. We release code, prompts, and parsers for replication upon acceptance.

2 Related Work

2.1 Bias in Text-to-Image Models

Text-to-image (T2I) systems inherit and can amplify social biases present in their training data. Large-scale analyses have documented systematic disparities across gender, skin tone, religion, disability, and professions, and have proposed diagnostic protocols and scorecards for auditing representation (Luccioni et al., 2023; Wan et al., 2024). Several debiasing methods have been explored either post hoc or with lightweight adaptation. *Fair Diffusion* steers sampling toward user-specified demographic proportions without retraining the backbone (Friedrich et al., 2023), while *Fair Mapping* introduces a small linear adaptor to reduce demographic skew with minimal compute (Li et al., 2025). More recently, *FairCoT* leverages chain-of-thought (CoT) reasoning in multimodal LLMs to mitigate bias during generation-time prompting (Al Sahili et al., 2025b). Our work differs in objective:

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: Attribute Prediction vs. Alignment

Attribute prediction. Face-centric attribute benchmarks remain widely used to quantify gender, race, and age performance. FAIRFACE provides a balanced dataset across race groups together with age and gender labels (Kärkkäinen and Joo, 2021). For context-embedded protected-attribute cues beyond faces, PATA was introduced alongside DeAR for debiasing vision–language models (Seth et al., 2023). In practice, many works rely on the open-source *DeepFace* toolkit for age, gender, and race prediction (Serengil, 2024), and on *CLIP* for zero-shot attribute attribution through prompted label matching (Radford et al., 2021). At the same time, audits of CLIP-family encoders show systematic gender and racial skew whose magnitude depends on model scale and, importantly, on pre-training data composition (Al Sahili et al., 2025a). We extend this line of work by prompting MLLMs (GPT-4.1, Gemini, and Llama-4) to infer social attributes directly under label constraints with explicit evidence requirements and abstention, enabling evaluation beyond strictly face-visible categories, including religion and disability.

Text–image alignment. CLIP-based similarity remains a dominant automatic proxy for prompt faithfulness (Hessel et al., 2021). Beyond cosine-based scores, VQA-driven frameworks such as TIFA generate question–answer checklists to probe fine-grained concept grounding (Hu et al., 2023). Preference-trained evaluators, including IMAGEREWARD and PICKSCORE, use large-scale human comparisons to better match human judgments in T2I ranking and evaluation (Xu et al., 2023; Kirstain et al., 2023). More recently, X-IQE proposes an explanation-oriented evaluator based on visual LLMs and reports improved agreement with human judgments while producing interpretable rationales (Chen et al., 2023). Our alignment setup is closest in spirit to this line: we compare CLIP similarity against MLLM-based judging under a standardized rubric that maps explanation-grounded assessments to the continuous range $[-1, 1]$.

2.3 LLMs/MLLMs as Judges

Using LLMs as evaluators ("LLM-as-a-judge") has become a practical alternative to costly human evaluation. Foundational work reports strong agreement with human preferences but also identifies position, verbosity, and self-enhancement biases, motivating careful protocol design and meta-evaluation (Zheng et al., 2023). Subsequent studies further examine reliability, consistency, and judge bias across open-ended tasks (Chen et al., 2024). In the multimodal setting, recent work shows that MLLMs can act as structured safety and policy judges for images, highlighting both the promise and the limitations of multimodal evaluation pipelines (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 scores with MLLM judges that also provide explicit evidence and rationales.

3 Methods

3.1 Background: CLIP and CLIP-based Alignment

CLIP. Contrastive Language–Image Pretraining (CLIP) learns an image encoder $f_\theta(\cdot)$ and a text encoder $g_\phi(\cdot)$ such that paired image–text embeddings have high cosine similarity while mismatched pairs are pushed apart. Given an image I and a textual prompt t, the canonical alignment score is the cosine

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

For multi-class recognition with a label set $\mathcal{Y}$, one typically forms templated prompts $\{t_y : y \in \mathcal{Y}\}$ and predicts $\arg\max_{y\in\mathcal{Y}} \mathrm{CLIPAlign}(I, t_y)$, optionally temperature-scaled into a softmax over classes.

Limitations for fairness evaluation. CLIP-based scores are sensitive to prompt wording and class-name choices; they provide *no calibrated abstention* when evidence is weak; and they reward surface correlations, especially for attributes only partially encoded in pixels (e.g., religion and

disability). In addition, CLIP cannot natively justify decisions or enforce label constraints beyond the provided prompts, which limits its utility as an *auditable* evaluator.

3.2 MLLM-as-Judge Protocol

We treat instruction-following multimodal LLMs (MLLMs) as *judges*. The protocol is lightweight and calibrated: (i) *label constraints* restrict outputs to a closed set (with UNKNOWN/UNSPECIFIED as a neutral fallback); (ii) the model must cite *explicit visual evidence* for each decision; and (iii) the judge *abstains* when evidence is insufficient. We instantiate judges with GPT-4.1, Gemini 1.5 Pro, and Llama-4; prompts are identical across models up to minor formatting adjustments.

3.2.1 Social-Attribute Prediction

We frame attribute recognition as a single structured judgment over five targets: *gender*, *race*, *age*, *religion*, and *disability*. The judge receives the image and an instruction that (a) enumerates admissible labels, including UNKNOWN/UNSPECIFIED, (b) requests a *strict JSON* output with fixed fields, and (c) requires a short, image-grounded evidence rationale for each prediction. For each model, we instantiate the *same* schema and decision rules with only minor phrasing adjustments. An abridged template appears below; complete prompts will be released upon acceptance.

```
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, Neutral,
unspecified)
5. Disability (mobility impairment,
blind/low vision, deaf/hard of hearing,
dwarfism, vitiligo, unspecified)
Output strict JSON with keys: "Gender",
"Race", "Age", "Religion", "Disability".
If evidence is insufficient, use
"unspecified".
```

Calibration and neutrality. We adopt two safeguards. First, we explicitly instruct the judge to select a label *only* when it is supported by visible cues; otherwise it must output UNSPECIFIED. Second, during development we request a terse rationale per field to audit evidence use and failure modes; at evaluation time, scoring uses the parsed JSON labels only, while rationales are retained for reproducibility and spot checks. This design reduces overconfident guessing and makes abstention systematic rather than ad hoc.

3.2.2 Prompt–Image Alignment Judging

We score prompt–image alignment using a rubric-based judgment inspired by explanation-oriented evaluators. Given an image I and description t, the judge assigns a discrete rating $r \in \{1, \ldots, 5\}$ using the following anchors:

```
According to the image and your previous
description, how well does the image
align with the following description?
(1) Not match at all; (2) Has
significant discrepancies; (3) Several
minor discrepancies; (4) A few minor
discrepancies; (5) Matches exactly.
```

We map r to a continuous alignment score $\overline{s} \in [-1, 1]$ via linear rescaling,

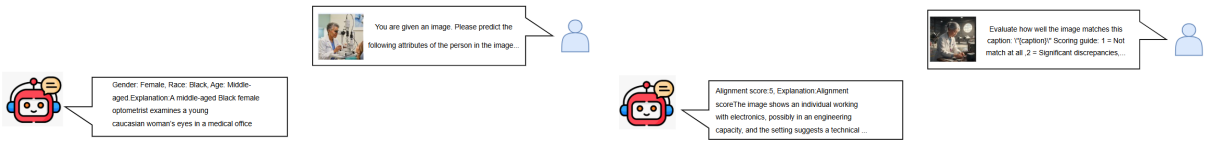
$$\overline{s} = \frac{r-3}{2}, \qquad \text{so that } 1 \mapsto -1,\ 3 \mapsto 0,\ 5 \mapsto 1,$$

which makes it directly comparable to cosine-based CLIP alignment. As with attribute prediction, the prompt requests evidence-backed justifications during development; at evaluation time we use the scalar $\overline{s}$ while retaining rationales for audit. To assess instruction sensitivity, we also evaluate minimal template perturbations (e.g., "a photo of" $\leftrightarrow$ "an image of") and verify that model ordering remains stable even when absolute scores shift.

3.3 Dataset Contributions: DIVERSIFY and DIVERSIFY-PROFESSIONS

We introduce two complementary resources that extend fairness evaluation *beyond* strictly face-visible cues and iconic settings.

DIVERSIFY. DIVERSIFY comprises 469 synthetic images rendered from controlled prompts and paired with normalized metadata for five labeled targets: *gender*, *race*, *age*, *religion*, and *disability*. The corpus is designed to surface context-dependent visual evidence—including attire, objects, scene composition, and assistive devices—for cases where simple face heuristics are insufficient. Demographic coverage is balanced by design (gender approximately 50/50), with race and age distributed across meaningful strata. Beyond demographics, DIVERSIFY explicitly



(a) Attribute prediction (MLLM as judge). The image and label-constrained prompt are fed to the judge, which outputs strict JSON for *Gender*, *Race*, *Age*, *Religion*, and *Disability*, with calibrated abstention.

(b) Prompt–image alignment (MLLM as judge). Given an image and caption, the judge assigns a rubric score $r \in \{1, \ldots, 5\}$ with a brief justification; we rescale to $\overline{s} \in [-1, 1]$ for direct comparison with CLIP cosine.

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.

varies religion (including *Neutral* where no visible symbol appears) and disability-related cues (e.g., mobility aids, blind/low vision, deaf/hard of hearing, dwarfism, vitiligo, and *unspecified* when absent). Labels are derived from the prompt taxonomy using deterministic parsing and canonicalization, followed by near-duplicate filtering and sanity checks for cue presence. In addition to these labeled dimensions, the corpus includes broader variation in scene and cultural context as exploratory coverage; these aspects are not used as labeled evaluation targets in the main experiments.

DIVERSIFY-PROFESSIONS. To evaluate occupational grounding and alignment under the same diversity principles, DIVERSIFY-PROFESSIONS contains 1,200 images organized as *6 professions* $\times$ *200* (doctor, engineer, janitor, lawyer, nurse, teacher). Each example pairs an image with a profession description and is constructed in non-iconic, diverse scenes to reduce shortcut cues. We report (i) top-1 profession *accuracy* and (ii) a rubric-based *alignment* score $\overline{s} \in [-1, 1]$ derived from a 1–5 judge rating. Distractor professions are selected to be visually plausible, increasing the need for explicit, evidence-backed decisions. Metadata mirror DIVERSIFY on labeled social attributes, enabling slice-wise fairness analysis conditioned on profession.

Summary. Together, DIVERSIFY (469 images; five labeled target attributes including religion and disability) and DIVERSIFY-PROFESSIONS (1,200 images; six professions) furnish controlled, diverse evaluation settings that complement FAIRFACE, PATA, and FAIRCOT by exposing context-driven failure modes and enabling counterfactual, slice-specific audits.

4 Experimental Results

We evaluate three questions: (i) *social-attribute prediction* for gender, race, and age on FAIRFACE, PATA, FAIRCOT, and DIVERSIFY; (ii) *generalization* to religion and disability beyond strictly face-visible cues; and (iii) *prompt–image alignment* on IDENPROF, FAIRCOT-PROFESSIONS, and DIVERSIFY-PROFESSIONS. For alignment, we report top-1 profession accuracy and mean alignment $\overline{s} \in [-1, 1]$.

4.1 Experimental Setup

Datasets. For attribute prediction, we evaluate on FAIRFACE, PATA, FAIRCOT, and DIVERSIFY. FAIRFACE provides balanced face images with human annotations for gender, race, and age. PATA and FAIRCOT broaden evaluation to settings where protected attributes must be inferred from faces together with surrounding context, and both are human-labeled. DIVERSIFY further expands coverage to culture, religion, disability, and demographics that remain underrepresented in existing benchmarks. For alignment, we use IDENPROF, FAIRCOT-PROFESSIONS, and DIVERSIFY-PROFESSIONS. These benchmarks move beyond iconic profession depictions by testing whether models can recognize profession cues in more contextual and culturally diverse settings, thereby improving benchmark quality.

Models and baselines. Judges are GPT-4.1, Gemini 1.5 Pro, and Llama-4, prompted with identical, label-constrained instructions (Sec. 3.2). Baselines are *CLIP* (zero-shot classification via templated prompts; cosine similarity for alignment) and *DeepFace* (face-centric gender/race/age prediction).

Metrics. For attributes, we report top-1 accuracy per attribute and dataset, together with macro

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.

means across populated cells only. For alignment, judges assign a 1–5 rubric score that we linearly rescale to $\overline{s} \in [-1, 1]$; where annotated, we also report profession accuracy. We additionally include minimal template perturbations ("a photo of"/"an image of") to assess instruction sensitivity.

4.2 Social-Attribute Prediction

Overall trends. Across all datasets, judge models deliver consistent, often double-digit improvements over CLIP and clearly exceed DeepFace on gender and age. Race remains the most challenging attribute, but judges still narrow the gap relative to both contrastive and face-centric baselines. This pattern holds across both face-centric settings (FAIRFACE, FAIRCOT) and more context-rich settings (PATA, DIVERSIFY), indicating that evidence-seeking instructions are beneficial even when cues are subtle or distributed across the scene.

By dataset. On FAIRFACE, GPT-4.1 leads on race and age, while Llama-4 slightly surpasses the others on gender; CLIP is competitive on gender but lags notably on age. On PATA, GPT-4.1 attains the strongest race and age performance, while Gemini 1.5 Pro and Llama-4 tie for the top gender score. On FAIRCOT, Gemini 1.5 Pro tops gender, GPT-4.1 tops race, and Llama-4 tops age, reflecting complementary strengths across judges. On DIVERSIFY, where demographic cues are less iconic and often embedded in richer scenes, judge models continue to outperform CLIP by large margins on gender and age and provide the strongest overall coverage. Detailed results appear in Table 1; best scores per column are bolded within each dataset block.

Table 1: Social-attribute prediction (top-1 accuracy). Columns: Gender, Race, Age.

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

Judge instructions that enforce label constraints, require visible evidence, and permit abstention translate into higher accuracy on attributes anchored more directly in appearance (gender, age) and competitive gains on harder, context-affected targets such as race. The fact that leadership swaps across datasets suggests complementary strengths

across MLLMs rather than a single uniformly dominant judge.

4.3 Generalization to Religion and Disability

Judge models also extend beyond standard demographic attributes. On FAIRCOT–*religion*, GPT-4.1 and Llama-4 achieve high accuracy while CLIP trails substantially, reflecting the value of evidence-based reasoning for symbol- and context-dependent judgments. On DIVERSIFY, judge models outperform CLIP by large margins on both *religion* and *disability*, with especially strong gains for disability where visible assistive cues and embodied context play an important role(Table 2).

Table 2: Generalization beyond demographics (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**	–

Where attributes require contextual reasoning or symbol recognition, judges offer a clear advantage over contrastive similarity. These results support the broader claim that evidence-grounded MLLM judging can extend fairness evaluation beyond strictly face-visible categories while still preserving transparent, abstention-capable decision making.

4.4 Prompt–Image Alignment

Judges also improve semantic grounding between descriptions and images. On IDENPROF, judge models match or exceed CLIP on profession accuracy and substantially increase the mean alignment score $\overline{s}$. On FAIRCOT-PROFESSIONS, GPT-4.1 yields the highest alignment with competitive accuracy, while on DIVERSIFY-PROFESSIONS all judges surpass CLIP by a wide margin on alignment while maintaining strong profession accuracy. Minimal template changes do not alter model ordering, indicating that the judging rubric is reasonably robust to light prompt variation(Table 3).

Rubric-calibrated judging produces alignment scores that better reflect profession correctness and scene semantics than cosine similarity, while avoiding over-penalization from minor prompt phrasing differences.

Table 3: Prompt–image alignment. We report top-1 Accuracy and mean Alignment ($[-1, 1]$).

Dataset	Method	Accuracy	Alignment
DIVERSIFY-Prof	CLIP	0.736	0.250
DIVERSIFY-Prof	GPT	**0.862**	**0.810**
DIVERSIFY-Prof	Gemini	0.842	0.770
DIVERSIFY-Prof	Llama	0.676	0.526
FairCoT-Prof	CLIP	0.471	0.250
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	GPT	0.937	0.680
IdenProf	Gemini	**0.943**	**0.709**
IdenProf	Llama	0.888	0.370

5 Discussion

Our results show that *judge-style* multimodal LLMs provide a more faithful and conservative read of sensitive attributes than contrastive or face-centric baselines. The strongest gains appear where visual evidence is diffuse or context-dependent. On gender and age, judges reliably outperform CLIP and DeepFace across datasets, and on religion and disability they close large gaps that face-only tools fail to address (Tables 1, 2). In prompt–image alignment, judges also raise mean alignment substantially over CLIP while maintaining high profession accuracy (Table 3). Taken together, these findings suggest that judge models capture something closer to *task intent supported by visible evidence*, rather than surface similarity alone.

A key reason for this improvement is the combination of label-constrained prompting and neutral fallbacks. By requiring the model to ground its decisions in observable cues and to abstain otherwise, we reduce the overconfident guesses that often plague open-ended evaluators. This effect is especially visible in domains such as religion and disability, where naive inference from facial appearance alone is inappropriate. In this sense, the judges behave more like cautious annotators: they succeed when the image contains verifiable evidence and resist filling gaps with speculation. By contrast, contrastive models remain competitive on attributes that are strongly anchored in pixel-level regularities and annotation conventions (for example, gender on FAIRFACE), but they do not provide the same calibrated abstention or explicit evidence structure.

The approach is complementary to representation learning rather than a replacement

for it. In practical pipelines, contrastive similarity can still serve as a high-recall retrieval signal, while the judge acts as a verifier that enforces label constraints and filters spurious matches. This division of labor is especially useful for professions and other multi-attribute settings, where retrieving plausible categories is relatively easy but verifying the precise category with defensible evidence is not. Our template-perturbation check further indicates that although absolute scores move with phrasing, the judge-based ordering remains stable, reinforcing the view that judge outputs reflect semantically grounded correspondences rather than brittle lexical artifacts.

At the same time, the results also highlight important limits. Performance varies across datasets, and the leading judge changes by attribute and benchmark. This variability suggests that apparent difficulty depends not only on the underlying phenomenon but also on dataset construction, visual salience, and label conventions. It therefore remains important to report results disaggregated by both attribute and dataset rather than collapsing them into a single headline number. More broadly, not every socially salient attribute is equally well suited to image-only evaluation, and the strongest use case for FairJudge is where the target label can in principle be supported by visible evidence and where abstention is treated as a valid outcome rather than an error.

Finally, there are important safety considerations. Judging sensitive attributes can cause harm when labels are ill-defined or spuriously inferred. Our protocol mitigates some of this risk by constraining outputs, privileging explicit evidence, and encouraging abstention, but these safeguards do not eliminate the underlying ethical concerns. In practice, such audits should be used to characterize models and datasets rather than people, and decisions affecting individuals should never be automated on the basis of these inferences. Within that scope, judge-style evaluation offers a clearer and more accountable lens on fairness than score-only pipelines.

6 Conclusion

We introduced a lightweight, calibrated protocol that treats instruction-following MLLMs as *fair judges* for T2I evaluation. By constraining outputs to label sets, requiring explicit visual evidence, and encouraging principled abstention, the protocol shifts evaluation away from score-only heuristics toward accountable, evidence-aware judgments. Empirically, judge models deliver consistent gains over contrastive and face-centric baselines on gender and age, and close large gaps on attributes that require contextual, non-facial reasoning, notably religion and disability. They also improve prompt–image alignment while retaining high profession accuracy, indicating that their judgments better reflect task-relevant evidence than surface correlations alone.

A second contribution is the release of DIVERSIFY and DIVERSIFY-PROFESSIONS, which probe fairness and grounding beyond strictly face-visible cues. These resources expose failure modes that standard demographic benchmarks can miss and provide more diverse, non-iconic settings for evaluating evidence-based judgments. In addition to the labeled dimensions used in our main experiments, DIVERSIFY includes broader variation in scene and cultural context as exploratory coverage for future work.

Taken together, our results support a practical design pattern: pair contrastive retrieval for high-recall candidate generation with judge-style verification for precision, justification, and calibrated abstention. This cascaded approach preserves the efficiency of representation learning while inserting an accountable reasoning step where it matters most. To foster rigor and transparency, we release per-image judge outputs, rationales, and parser logs alongside the new datasets, enabling independent analysis, stress testing, and re-scoring under alternative rubrics.

Important limitations remain. Instruction phrasing can shift absolute scores, domain shift can reorder leaders across datasets, and not every socially salient attribute is equally well suited to image-only evaluation. We view these limitations as motivation for broader stress tests, stronger cross-dataset calibration, and closer triangulation with human evaluation. Nevertheless, the central message remains: evidence-grounded judging and evaluation beyond face-only proxies offer a concrete path toward more reliable, transparent, and interpretable assessment of multimodal systems.

7 Limitations

Despite the gains, several limitations deserve emphasis. First, ambiguity in annotation constrains

ceiling performance, especially for socially sensitive attributes whose operationalization may vary across datasets and labeling schemes. Even when judges abstain appropriately, measured accuracy can reflect the benchmark's definitions as much as any underlying construct. Second, some attributes may not be reliably inferable from a single image. High scores may therefore arise from dataset-specific regularities rather than genuine visual grounding; conversely, responsible abstention may be penalized when labels require extra-visual context. Third, judge outputs remain sensitive to instruction framing. We mitigate this with label constraints, neutral defaults, and template checks, but broader stress testing across prompts and task wordings is still needed to quantify instruction variance.

Fourth, dataset shift remains substantial. The identity of the top-performing method changes across datasets and attributes, indicating residual coupling to collection practices, visual conventions, and label taxonomies. Robust conclusions therefore require leave-one-dataset-out analyses, calibrated thresholds, and explicit reporting of uncertainty (e.g., bootstrap intervals), which we encourage but do not exhaustively explore here. Fifth, judges can still exploit ethically fraught correlates. Evidence grounding reduces spurious inference but does not eliminate the possibility that a model relies on proxy cues. Audits should therefore include qualitative review of rationales and, where possible, counterfactual probes designed to break such shortcuts.

Sixth, the alignment rubric introduces metric dependence. Although our rubric tracks profession accuracy and yields more semantically grounded judgments than cosine similarity, different rubrics or panel-based judgments could shift absolute values. Future work should triangulate these scores with human evaluation and alternative scoring schemes. Finally, computational cost and latency are higher for judge models than for single-pass contrastive scoring, which may limit their use at scale. We view the additional cost as justified for fairness auditing, but practical deployments will benefit from cascaded designs, caching, and selective judging only where contrastive filters indicate uncertainty or sensitivity.

References

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

Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2025b. Faircot: Enhancing fairness in text-to-image generation via chain of thought reasoning with multimodal large language models. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

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, Miami, Florida, USA. Association for Computational Linguistics.

Yixiong Chen, Li Liu, and C. Ding. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *ArXiv*, abs/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*.

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. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20349–20360.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663.

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.

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.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. In *Neural Information Processing Systems*.

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.

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.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14657–14666.

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 Ablation

We conduct an ablation study on the FAIRCOT dataset to probe the contribution of each protocol component—*label constraints*, *abstention via* UNSPECIFIED, and *evidence prompting*. For scoring, we apply an attribute-specific policy for UNSPECIFIED (abstentions): (i) for *gender*, *race*, and *age*, UNSPECIFIED is treated as a *misclassification* (i.e., counted in the denominator with zero credit); (ii) for *religion*, UNSPECIFIED is treated as *neutral* and *excluded* from accuracy (we also report coverage as the fraction of non-abstentions); and (iii) for *disability*, UNSPECIFIED is interpreted as predicting *no disability* (i.e., scored as correct when the ground truth is "no disability" and as an error otherwise). Table 4 reports per-attribute accuracy under three settings per model: (i) *Without labels* (free-form answers), (ii) *without unspecified* (forced prediction), and (iii) *our* full FAIRJUDGE protocol.

Table 4: Ablations on FAIRCOT (per-attribute accuracy). "Without labels" removes the closed taxonomy; "without unspecified" disables abstention. Llama age reflects the author's normalization; religion notes indicate frequent abstention/free text.

Model	Setting	Gender	Race	Age	Religion
GPT-4.1	Without labels	94.7%	11.2%	0.86%	59.5%
GPT-4.1	without unspecified	–	–	–	–
GPT-4.1	**our**	**97%**	**82%**	**93%**	**88%**
Gemini 2.5 Pro	Without labels	80.17%	17.24%	0%	85.34%
Gemini 2.5 Pro	without unspecified	92.20%	15.80%	8.40%	21.10%
Gemini 1.5 Pro	**our**	**99%**	**82%**	**96%**	**82%**
Llama-4	Without labels	43%	21.40%	89%†	10%‡
Llama-4	without unspecified	45%	35%	90%†	9%‡
Llama-4	**our**	**99%**	**69%**	**97%**	**88%**

†"(Normalised)" in source sheet. ‡"Most said cannot be determined" (frequent abstention/free-text).

Removing the *label constraints* causes steep drops on the hardest axes (race, age): free-form outputs drift off the taxonomy and cannot be reliably parsed back, even when the model's prose sounds reasonable. Disabling *abstention* forces guesses and disproportionately harms attributes with weak or intermittent evidence (religion, age), as seen most clearly for Gemini. By contrast, the full protocol—with a closed label set, an explicit UNSPECIFIED option, and an evidence-seeking instruction—recovers high gender/age accuracy and substantially improves race/religion, especially for Llama-4 where the gap between "without labels" and "our" is largest. In short: constraints prevent drift; abstention curbs overconfidence; evidence prompts focus the judge on verifiable cues. These effects are strongest exactly where unconstrained or forced-guess pipelines are most brittle.

B Dataset Contribution and Datasheet: DIVERSIFY

B.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.

B.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.[1]

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.

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

Table 5: Gender distribution ($N = 469$).

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

Table 6: 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%

B.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.

B.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 7: 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 8: Religion distribution ($N = 469$).

Label	Count	%
Neutral	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%
Unspecified	0	0.0%

perturbations.

B.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.

B.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.

B.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.[2] Metadata include prompt IDs, normalized labels, and generation seeds to facilitate exact reproduction and re-scoring under alternative taxonomies.

Table 9: 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%

C Dataset Contribution and Datasheet: DIVERSIFY-Professions

C.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.

C.2 Composition and Statistics

Each image is associated with structured metadata parsed from prompt text. Table 10 shows the profession coverage; Tables 11–12 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

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

Table 10: 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 11: 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%

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.

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.

C.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.

Table 12: 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%

C.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.

C.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.

C.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.

C.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.[3] To facilitate reproducibility, we include prompt IDs, attribute labels, and (where applicable) generation seeds.

C.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.

D 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).

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

Table 13: 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 $\overline{s} \in [-1, 1]$ for comparison with CLIP cosine; minimal template perturbations are reported in Sec. 4.4.

D.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, so we focus on 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 as open source, the *marginal* API cost would be zero but it would be the *most computationally expensive* to run (GPU-hours/energy/ops). A pragmatic deployment is a progressive judging cascade: route easy cases to a cheaper judge (e.g., Gemini 1.5 Pro), escalate only *abstentions* or *low-evidence* cases to GPT-4.1, and invoke Llama-4 *only* for disagreements between judges or fairness-critical audits requiring a second opinion. This preserves most of the accuracy/alignment gains while controlling runtime and spend, and it aligns with an auditor's workflow where explanations, abstentions, and cross-checks matter most.

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.