# Security and Privacy of Machine Learning, 2025 Critique 11/12: Fairness – (1) FairNet: Dynamic Fairness Correction without Performance Loss via Contrastive Conditional LoRA (2) Guiding LLM Decision-Making with Fairness Reward Models (3) On Fairness of Unified Multimodal Large Language Model for Image Generation

Shih-Yu Lai
*National Taiwan University*
Taipei, Taiwan
akinesia112@gmail.com

## Critique: (1) FairNet: Dynamic Fairness Correction without Performance Loss via Contrastive Conditional LoRA

### Summary

The paper proposes **FairNet**, a dynamic, instance-conditioned debiasing framework that inserts (i) lightweight *bias detectors* at intermediate layers and (ii) *conditionally-activated LoRA adapters* to adjust internal representations only for samples flagged as vulnerable. The LoRA modules are trained with a *contrastive* (triplet-style) loss that explicitly minimizes *intra-class, inter-group* distances, aiming to uplift minority-group representations without perturbing others. A key theoretical claim is a sufficient condition—expressed via a ratio of the detector's TPR/FPR and group prevalences—under which worst-group performance improves *without* degrading overall accuracy, thereby challenging the conventional fairness–accuracy trade-off. Empirically, on CelebA, MultiNLI, and HateXplain, FairNet variants (full/partial/unlabeled group labels) improve worst-group accuracy (WGA) and reduce Equalized Odds Difference (EOD) while matching or slightly exceeding ERM accuracy.

### Strengths

- **Selective, representation-level correction.** The conditional LoRA mechanism is a principled form of *conditional computation* that targets only suspected failure modes, avoiding global shifts that often harm accuracy. This is a fresh angle relative to standard pre-/in-/post-processing families [1], [2].
- **Contrastive alignment objective.** Training LoRA with an intra-class, inter-group objective seeks label-conditional invariance; this aligns with the spirit of individual/group fairness while maintaining class separability. It also operationalizes a concrete recipe for leveraging majority-group signal to uplift minority subgroups.
- **Coverage of label regimes.** The unified treatment for full, partial, and unlabeled sensitive attributes—including an unsupervised detector—increases deployability in realistic settings where collecting group labels is costly or restricted [3].
- **Theoretical sufficiency condition.** The condition ties detector quality (TPR/FPR) to performance preservation in a transparent way, giving practitioners a knob (the activation threshold) to navigate fairness–accuracy trade-offs.
- **Compelling empirical ablations.** Detector ablations isolate accuracy preservation; contrastive-loss ablations isolate fairness gains; threshold sweeps expose a pragmatic Pareto frontier. The intersectional experiment on HateXplain is a valuable step beyond single-axis bias.

### Weaknesses / Concerns

- **Detector learning target and identifiability.** The detector is trained to predict minority status (or an unsupervised proxy), not *error propensity*. If minority membership is an imperfect surrogate for failure risk, conditional activation can misfire, creating *disparate mistreatment* even if group metrics improve [1]. A direct *error-risk* detector might be more aligned with the objective.

- **Triplet construction and class imbalance.** The contrastive objective hinges on high-quality anchor/positive/negative selection. In sparse minorities, hard positive mining may be unstable; averaging majority prototypes risks *representation collapse* that erases informative within-group variation (useful subpopulation structure).
- **Theoretical assumption gap.** The sufficiency condition requires the unobservable quantities $P(M_{\text{LoRA}}, G)$ (counterfactual accuracies if LoRA were always on). The paper assumes they are favorable for minorities and benign for majorities, but offers limited diagnostics to validate these assumptions per task.
- **Metric scope and calibration.** Results emphasize WGA/EOD; impacts on *calibration*, *PPV parity*, and decision-threshold fairness are not reported. Conditional representation alignment can improve EOD while worsening calibration disparities, which matter in high-stakes domains.
- **OOD and drift robustness.** Detectors keyed to intermediate features may be brittle under covariate shift; conditional adapters could then activate on the wrong regimes. There is no stress-test on distribution shift or temporal drift, which are common in deployment.

### Potential Improvements / Extensions

- **Risk-aligned detectors.** Train detectors to predict *error likelihood* (or margin-based uncertainty) directly, optionally multi-tasking with group membership when available. This better targets instances needing correction and decouples protected-attribute prediction from intervention triggering.
- **Counterfactual and causal regularization.** Incorporate counterfactual representation constraints or causal invariance penalties to prevent over-alignment that removes task-relevant group structure and to address fairness impossibility frontiers [4].
- **MoE-style routing and budgeted activation.** View conditional LoRA as sparse experts with learned gates and an activation budget. Optimize the gate with a constrained objective (e.g., FLOPs, fairness budget) to guarantee predictable overhead and to shape whom the model helps.
- **Hard-negative curricula and proxy-free anchors.** Replace triplets built from potentially biased proxies with semi-supervised or self-supervised positives (e.g., label-preserving augmentations) and adversarial hard negatives to reduce reliance on noisy group labels.
- **Expanded evaluation.** Report calibration (ECE), cost-sensitive metrics, threshold-free curves, and false-positive/false-negative parity. Add OOD (subpopulation/temporal) stress-tests and label-noise robustness. Compare with last-layer debiasing and plug-in thresholding baselines that are surprisingly strong [2], [3].

### Questions for the Authors

1) **Detector alignment:** Have you tried training detectors on *error* (or high-loss) indicators rather than group identity? How do WGA/EOD and ACC change?
2) **LoRA scope:** Which layers benefit most from conditional LoRA? Does concentrating adapters at earlier vs. later blocks alter the fairness–accuracy Pareto?
3) **Calibration:** What is the effect on group-wise calibration and decision thresholds? Can the conditional adapters be combined with group-aware thresholding post-processing [1] without double-counting fairness?
4) **Unlabeled regime:** For the unsupervised detector, how sensitive are results to the outlier-detection method (LOF vs. Isolation Forest) and to the chosen representation layer?
5) **Shift robustness:** Under subpopulation shift or domain shift, does the detector over- or under-fire? Can you add a small OOD detector to gate LoRA conservatively out-of-support?

## Critique: (2) Guiding LLM Decision-Making with Fairness Reward Models

### Summary

This paper proposes a **Fairness Reward Model (FRM)** that scores the step-wise fairness of chain-of-thought (CoT) reasoning and re-weights multiple sampled reasoning chains during inference to down-weight biased trajectories and favor equitable ones. Training uses weak supervision from an LLM judge to label biased/unbiased steps on BBQ [5]. At test time, the FRM is applied post hoc to decisions in three high-stakes domains (COMPAS, CivilComments, Bias-in-Bios), optimizing a temperature-controlled weighted vote. Results show consistent reductions in equalized opportunity/odds gaps [**?**] with little or no accuracy loss, sometimes improving accuracy. The work positions itself amid CoT/ToT scaling [6] and evidence that CoT can amplify bias [7], arguing for process-level fairness verification as a model-agnostic, inference-time control.

### Strengths

- **Clear problem framing:** Recognizes that CoT scaling improves accuracy but can entrench disparities [7], and targets *reasoning-process* fairness instead of only outcomes.
- **Simple, deployable mechanism:** Post-hoc re-weighting preserves base-model performance and auditability; no fine-tuning of task models required.
- **Transfer beyond training domain:** A single FRM trained on BBQ generalizes across tasks, attributes, and even unseen model families.
- **Tunable trade-off:** Temperature parameter offers explicit control between self-consistency and fairness—useful for practitioners balancing risk and performance.
- **Interpretability:** Step-level scores expose *where* bias occurs in a rationale, supporting traceable, process-based reviews (useful for policy and governance).

### Weaknesses

- **Weak-label reliance & alignment risk:** Fairness supervision comes from another LLM. This may propagate the annotator's latent biases and norms, yielding a brittle standard that drifts with foundation-model updates; human agreement rates suggest nontrivial noise.
- **Uniform step weighting:** Averaging step scores ignores causal contribution to the final decision; a briefly biased aside and a bias-driving step count equally.
- **Metric narrowness:** Focus on equalized odds/opportunity [**?**] omits calibration within groups, individual fairness, and counterfactual/casual notions—important in hiring, credit, and justice.
- **US/English-centric scope:** Training/evaluation reflects US sociopolitical categories; fairness judgments may not transfer to multilingual or culturally distinct settings.
- **Adversarial dynamics:** The approach assumes truthful rationales; strategic models or jailbreak prompts could sidestep detection (e.g., burying bias in implicit features) as noted in prior CoT safety work [7].

### Potential Improvements or Extensions

- **Causal contribution scoring:** Replace uniform step averaging with causal credit assignment (e.g., counterfactual masking or step-wise Shapley over reasoning tokens) to penalize *decision-causal* biased steps most.
- **Multi-objective decoding:** Integrate FRM into search-time guidance (e.g., ToT or lookahead) and RL finetuning as a secondary reward beside correctness verifiers [6], enabling proactive rather than post-hoc fairness.
- **Expanded fairness portfolio:** Add calibration-within-groups, individual similarity metrics, and counterfactual fairness checks; report Pareto fronts to transparently show trade-offs practitioners must accept.
- **Cross-cultural & multilingual validation:** Train lightweight adapters on non-English fairness corpora and sensitive attributes appropriate to local law/regulation; measure transfer and failure modes.
- **Robustness to prompt attacks:** Stress-test with red-teaming focused on implicit stereotyping and proxy features; combine FRM with toxicity/bias verifiers and reasoning-verifier ensembles [7].
- **Human-in-the-loop calibration:** Use active learning to collect targeted human labels on *disagreement* slices between FRM and outcome verifiers; continually recalibrate thresholds/temperatures per application.
- **Procedural guarantees:** Explore certified upper bounds on parity gaps under re-weighting assumptions; add uncertainty quantification to fairness scores and decisions for risk-aware deployment.

### Questions for Authors

1) **Label governance:** How would you mitigate FRM drift if the supervising LLM is updated? Can you provide a protocol for periodic calibration with small human audits?
2) **Causal sensitivity:** Have you tried importance-weighting steps by estimated influence on the final answer (e.g., deletion diagnostics or counterfactual rerolls)? What trade-offs in compute/fairness arise?
3) **Metric choice:** In settings with base-rate imbalance, how does FRM behave for within-group calibration and predictive parity? Any evidence of overcorrection (e.g., fairness gerrymandering)?
4) **Implicit bias detection:** Can FRM catch proxy use (e.g., ZIP code, career breaks) that correlate with protected attributes when they are never named explicitly?
5) **Operationalization:** For real deployments, how should practitioners pick temperature and thresholds given legal constraints (e.g., 80% rule) and business KPIs? Any guidance or search procedure?

## Critique: (3) On Fairness of Unified Multimodal Large Language Model for Image Generation

### Summary

The paper investigates demographic bias (gender, race) in unified multimodal LLMs (U-MLLMs) that both understand and generate images via a single autoregressive pipeline. It benchmarks several recent U-MLLMs (e.g., VILA-U) and reports substantial disparities in generated outputs for neutral prompts across occupations. Through component-level auditing, the authors argue bias primarily stems from the LM that autoregressively produces image tokens, rather than the vision tokenizer. To mitigate bias, they synthesize a balanced training set using a diffusion model and propose a *Balanced Preference Optimization* loss derived from reference-free preference objectives (*e.g.*, ORPO) to equalize generation odds across demographics, showing reductions in representation disparity while maintaining image fidelity [8]–[12].

### Strengths

- **Clear localization of bias.** The paper moves beyond aggregate metrics to inspect where bias arises in a unified pipeline, providing evidence that token-level generation in the LM is the dominant contributor. This shifts the fairness discussion from vision towers to sequence modeling—an important reframing [8].
- **Methodological simplicity with practical utility.** The balanced preference loss is conceptually simple, compatible with existing preference-optimization frameworks [10], [11], and does not require a separate reward model. That makes adoption realistic for practitioners.
- **Balanced synthetic data as a lever.** Using a diffusion model to curate demographic balance is a pragmatic solution when web-scale data are skewed. The pairing of neutral prompts with multi-demographic image sets is neat and reduces confounding in downstream alignment.
- **Comprehensive evaluation protocol.** The study uses an occupation benchmark with a consistent sampling regime and reports multiple quality metrics alongside fairness, helping assess trade-offs [9].

## Weaknesses

- **Fairness target unclear (*parity vs. realism*).** The objective implicitly enforces demographic parity for *neutral* prompts, yet does not articulate whether parity is the desired target under realistic base-rate differences across geographies, eras, or subdomains. Without a principled notion of the "reference distribution," parity may create distribution shift or misrepresent domain specifics.
- **Classifier-as-oracle risk.** The fairness metric hinges on a pretrained attribute classifier. If that classifier exhibits its own racial/gender error asymmetries, measured gains could reflect *gaming the classifier* rather than genuine fairness. No calibration or subgroup error analysis of the classifier is reported [12].
- **Synthetic data bias importation.** The pipeline bootstraps fairness using diffusion-generated images; however, diffusion models can encode style and cultural priors. The paper does not quantify how the chosen generator (and prompt templates) might bias skin tone, facial morphology, attire, or lighting, potentially steering U-MLLMs toward the diffusion model's aesthetic.
- **Limited scope of attributes and prompts.** Only binary gender and a coarse 4-way race categorization are used; other salient axes (age, disability, attire, religion, intersectional identities) and multilingual prompts are absent. Occupational prompts are English-only and Western-centric, risking cultural narrowness.
- **Causality not fully established.** The LM is identified as the main source via token-distribution similarity (JSD), but a formal causal mediation analysis (quantifying indirect vs. direct effects through modules) is missing. This leaves room for alternative explanations (e.g., tokenizer discretization interacting with LM priors).

## Potential Improvements or Extensions

- **Define fairness targets by context.** Introduce configurable targets: (i) demographic parity for general-purpose creative use, (ii) base-rate matching for *domain-aware* use (via external stats or curated datasets), and (iii) *conditional parity* given prompt attributes (region, year, industry). Learn a controller that interpolates among targets.
- **Auditor robustness.** Validate the attribute classifier: report subgroup ROC/AUC, calibration, and error parity; triangulate with a second auditor and human judgments to bound measurement error.
- **Causal probing.** Use interventional or modular ablations (e.g., randomized token masking, counterfactual prompting) and causal mediation to quantify LM vs. vision-tokenizer contributions under interventions, not just associational distances.
- **Broader attribute coverage.** Extend to age, attire, head-coverings, disability aids, and multilingual prompts; evaluate intersectionality beyond (gender, race) with compositional prompts (*e.g.*, language + occupation + setting).
- **Sampling-time control.** Analyze how temperature, nucleus $p$, classifier-free guidance, and token truncation affect fairness. Provide a fairness knob at inference with guardrails (e.g., per-prompt parity constraints).
- **Human-centered outcomes.** Add human evaluation focusing on stereotype salience, offensiveness, and perceived authenticity—not just frequency counts. Include qualitative audits for edge prompts (e.g., "CEO" vs. "care worker").
- **Real-data finetuning with privacy.** Combine synthetic balancing with private, consented, demographically rich datasets; audit domain shift between synthetic and real distributions.

## Questions

1) How sensitive are fairness gains to the *specific* diffusion model and prompt templates used for synthetic data? Would switching the generator reduce or amplify artifacts?
2) Can the method target *conditional* fairness (e.g., parity within region/language strata) rather than global parity?
3) What prevents over-correction when users *explicitly* ask for a demographic attribute? Is there a policy layer that respects user intent while guarding against stereotyped defaults?
4) How robust are results across decoding parameters (temperature, top-$p$) and different LM backbones or tokenizers?
5) Could causal mediation quantify the LM's contribution vs. the tokenizer's under controlled interventions, strengthening the attribution claim?

## References

[1] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NeurIPS*, 2016.

[2] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *International Conference on Learning Representations (ICLR)*, 2020.

[3] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without group labels," in *International Conference on Machine Learning (ICML)*, 2021.

[4] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Innovations in Theoretical Computer Science (ITCS)*, 2017.

[5] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "BBQ: A hand-built bias benchmark for question answering," in *Findings of ACL*, 2022.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.

[7] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning," in *ACL*, 2023.

[8] Y. Wu, Z. Zhang, J. Chen, H. Tang, D. Li, Y. Fang, L. Zhu, E. Xie, H. Yin, L. Yi *et al.*, "Vila-u: A unified foundation model integrating visual understanding and generation," *arXiv preprint arXiv:2409.04429*, 2024.

[9] X. Shen, C. Du, T. Pang, M. Lin, Y. Wong, and M. Kankanhalli, "Finetuning text-to-image diffusion models for fairness," *arXiv preprint arXiv:2311.07604*, 2024.

[10] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *Advances in Neural Information Processing Systems*, 2024.

[11] J. Hong, N. Lee, and J. Thorne, “Orpo: Monolithic preference optimization without reference model,” *arXiv preprint arXiv:2403.07691*, 2024.

[12] K. Kärkkäinen and J. Joo, “Fairface: Face attribute dataset for balanced race, gender, and age,” in *arXiv preprint arXiv:1908.04913*, 2019.