International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 10 | Oct 2025 www.irjet.net p-ISSN: 2395-0072

Reinforcement Learning from Human Feedback for Trustworthy Vision–Language Models

Laraib Ahmad Siddiqui[1], Mohd Shahzad[2]

[1]Program Control Services Analyst, Accenture, India
[2]AWS and DevOps Consultant, Deloitte, India

Abstract - *Multimodal foundation models that jointly process vision and language, such as CLIP, BLIP-2, and GPT-4V, demonstrate impressive perceptual and reasoning abilities but remain prone to hallucination, bias, and misalignment with human intent. This paper introduces an RLHF-based alignment framework for multimodal models, designed to teach them what humans actually value in visual understanding tasks. We construct a preference dataset of human-rated visual explanations across the domains of image captioning, visual question answering, and video narration. A reward model jointly optimizes for factual consistency, contextual grounding, and bias penalties, and a scalable evaluation harness built on top of Kubernetes enables the automated comparison of pre- and post-alignment model outputs. Empirical results show measurable improvements in factual accuracy (↑ 18%), bias reduction (↓ 22%), and overall human preference alignment (↑ 25%) on multimodal benchmarks. Our findings offer a reproducible path toward trustworthy vision–language alignment, laying the groundwork for safer multimodal agents in deployment contexts.*

Key Words: Reinforcement Learning from Human Feedback (RLHF), Vision–Language Models (VLMs), Human Preference Alignment, Factual Consistency, Bias Mitigation, Multimodal Model Evaluation, Visual Grounding, Trustworthy AI

1. INTRODUCTION

1.1 Motivation

The next generation of AI assistants must *see, talk, and act* safely. Multimodal models such as GPT-4V, Gemini 1.5 Pro, and Flamingo have bridged language and perception, yet their reasoning often drifts from visual evidence, producing hallucinated captions, biased predictions, or contextually implausible narratives.

As these systems increasingly drive applications in autonomous robotics, medical imaging, and content moderation, alignment with human intent becomes essential not only for user trust but also for regulatory compliance under frameworks like the EU AI Act.

1.2 Problem Statement

Traditional supervised fine-tuning optimizes likelihood on human-authored text but fails to encode the nuanced preferences underlying human visual understanding of truthfulness, contextual sensitivity, and social fairness.

While RLHF has proven transformative for large language models, extending it to vision–language domains presents unique challenges:

- Multi-modal reward modeling requires joint visual and textual grounding.
- Human preferences depend on both *factual alignment* and *perceptual saliency*.
- Evaluation must be scalable and reproducible across large, heterogeneous datasets.

1.3 Proposed Approach

We propose a Reinforcement Learning from Human Feedback (RLHF) framework tailored to multimodal models.

The approach consists of:

1. **Human-Preference Dataset Creation:** Collect pairwise preferences on visual outputs (captions, VQA answers, rationales).
2. **Reward Modeling:** Train a composite reward model combining (a) relevance to visual content, (b) factual faithfulness, and (c) bias-sensitive regularization.
3. **Policy Optimization:** Fine-tune a CLIP- or BLIP-based encoder–decoder via proximal policy optimization (PPO) using the learned reward.
4. **Evaluation Harness:** Implement scalable comparison using a Python/Kubernetes pipeline with automatic metric logging (accuracy, bias, human preference scores).

1.4 Contributions

Our contributions are threefold:

1. A **novel multimodal RLHF pipeline** integrating human preferences directly into vision–language reasoning.

2. A **composite reward model** that operationalizes factual and ethical alignment.
3. A **scalable evaluation suite** enabling reproducible benchmarking of alignment gains across vision–language datasets.

2. RELATED WORK

2.1 Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) has become a cornerstone for aligning large language models (LLMs) with human values and intent. Initial work by Christiano et al. (2017) demonstrated that preference-based reinforcement learning could effectively replace explicit reward engineering in text generation.

Subsequent advances, such as InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2023), established scalable human feedback pipelines using pairwise preference comparisons and proximal policy optimization (PPO) to align LLM behavior with subjective quality metrics (helpfulness, honesty, harmlessness). Recent extensions to multimodal models (e.g., GPT-4V, Gemini, Kosmos-2, LLaVA-RLHF) indicate the growing feasibility of applying human feedback to visual–language tasks.

However, most existing RLHF implementations remain text-centric, rewarding linguistic fluency rather than visual faithfulness. Our work addresses this gap by incorporating *visual grounding* and *factual alignment* directly into the reward structure, enabling end-to-end multimodal alignment.

2.2 Multimodal Vision–Language Alignment

The emergence of contrastive pretraining frameworks such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) enabled scalable representation learning across image–text pairs, forming the foundation for modern multimodal systems. BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) extended these ideas to generative settings, introducing visual-language encoders that can produce captions, answer questions, and perform zero-shot reasoning.

Yet, these models are optimized for *correspondence*, not *alignment*: they learn statistical associations between vision and language without incorporating human preference or ethical judgment.

Recent efforts such as RLHF-V (Zhou et al., 2024) and Visual Alignment Tuning (Liang et al., 2024) experiment with applying RLHF to visual domains, but evaluation remains fragmented across datasets and lacks standardized bias or factuality metrics.

Our framework builds upon these precedents, offering a composite multimodal reward and scalable evaluation harness that unifies preference learning with real-world reliability objectives.

2.3 Explainable and Faithful Visual Reasoning

Explainable AI (XAI) methods for vision models have evolved from post-hoc saliency visualizations (Grad-CAM, Integrated Gradients) to language-based rationalization that describes *why* a model made a certain decision. Works such as VQA-X (Park et al., 2018) and e-SNLI-VE (Camburu et al., 2020) introduced datasets pairing images with natural-language explanations.

However, most explanation models optimize for *linguistic plausibility* rather than *factual grounding*. Recent multimodal transformers (Flamingo, LLaVA, MiniGPT-4) show promise in generating coherent rationales but still exhibit hallucinations and context drift.

By embedding human preference signals into both the captioning and reasoning stages, our approach explicitly trains models to prefer visually faithful rationales bridging the divide between interpretability and factual alignment.

2.4 Robustness, Fairness, and Safety in Vision Models

Large-scale vision transformers are vulnerable to dataset bias, domain shift, and adversarial perturbations. Studies such as Taori et al. (2020) and Geirhos et al. (2021) reveal sharp accuracy drops under distributional shifts (ImageNet-A, ImageNet-R), while Buolamwini & Gebru (2018) highlighted demographic biases in facial recognition systems. Efforts like FairFace, Balanced Faces, and RAI benchmarks aim to mitigate bias through data curation, but they rarely integrate feedback-driven adaptation during deployment.

Our proposed RLHF framework introduces bias penalties and reward shaping within the training objective, allowing human evaluators to directly influence fairness criteria rather than rely solely on post-hoc audits. This links *alignment learning* with *robustness assurance,* a step toward unified multimodal safety evaluation.

2.5 Summary of Gaps

Gap in Literature	Our Contribution
RLHF is limited to language-only tasks	Extend RLHF to multimodal (vision–language) alignment
Visual models lack human preference grounding	Introduce reward models based on human-rated visual faithfulness
Evaluation lacks reproducibility and scale	Build a Kubernetes-based evaluation harness
Bias and factuality are treated as separate problems	Combine both via a composite reward function

3. METHODOLOGY

Our proposed framework integrates Reinforcement Learning from Human Feedback (RLHF) into vision–language models (VLMs) to improve factuality, faithfulness, and fairness. It consists of four components:

1. **Base Multimodal Model Initialization**
2. **Human Preference Dataset Construction**
3. **Reward Modeling with Composite Objectives**
4. **Policy Optimization via PPO and Scalable Evaluation Harness**

3.1 Base Model

We begin with a pretrained encoder–decoder VLM, denoted $\pi_\theta(y \mid x)$, where:

- x= visual input (image or video frames + optional text prompt),
- y= generated text (caption, answer, rationale).

In experiments, BLIP-2 and LLaVA-1.5 are used as base architectures due to their strong visual-language grounding. The encoder $f_v$produces a latent representation $v = f_v(x)$, fused with a language embedding $e = f_l(t)$via cross-attention layers before decoding.

Formally, the model seeks to maximize expected reward:

$$\max_\theta \ \mathbb{E}_{x,y\sim\pi_\theta}[R(x,y)]$$

where $R(x,y)$is defined below.

3.2 Human Preference Dataset

Data Collection

To align perception with human judgment, we curate pairwise comparisons of multimodal outputs:

$$(x, y_A, y_B, h) \rightarrow p = \text{pref}(y_A > y_B)$$

where human annotators $h$select which response (A or B) better satisfies truthfulness, relevance, and ethical correctness.

Annotation Protocols

Annotators are trained on three dimensions:

1. **Factual Consistency** – Does the response reflect the actual visual evidence?
2. **Contextual Grounding** – Is the description coherent with the scene's semantic context?
3. **Bias Sensitivity** – Does it avoid demographic or moral stereotypes?

Preference votes are aggregated via Bradley-Terry normalization to produce probabilistic labels $p_{A,B}$.

3.3 Reward Modeling

A reward model $r_\phi(x,y)$is trained to approximate human preference probabilities. Given a pair (y_A, y_B), the objective follows Christiano et al. (2017):

$$\mathcal{L}_{RM}(\phi) = -\mathbb{E}[p_{A,B}\log\sigma(r_\phi(x,y_A) - r_\phi(x,y_B)) + (1 - p_{A,B})\log\sigma(r_\phi(x,y_B) - r_\phi(x,y_A))]$$

Composite Reward Function

The scalar reward Rcombines three components:

$$R(x,y) = \alpha_1 R_{\text{relevance}} + \alpha_2 R_{\text{faithfulness}} - \alpha_3 R_{\text{bias}}$$

where:

- $R_{\text{relevance}}$: cosine similarity between vision and text embeddings (semantic coherence),
- $R_{\text{faithfulness}}$: factual score from visual-grounding QA metrics (e.g., CIDEr, F1 vs reference answers),
- R_{bias}: fairness penalty estimated using group-sensitive classifiers (e.g., FairFace distribution shift score).

Weights $\alpha_1, \alpha_2, \alpha_3$are tuned via Pareto optimization to maintain a balance between accuracy and ethics.

3.4 Policy Optimization

The aligned policy $\pi_\theta$is fine-tuned using Proximal Policy Optimization (PPO) with the learned reward. Let π_{ref}denote the base frozen model and $A_t$the advantage estimate.

The PPO objective is:

$$\mathcal{L}_{PPO}(\theta) = \mathbb{E}_t[\min(\rho_t A_t,\ \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_t)]\text{where } \rho_t = \frac{\pi_\theta(y_t \mid x_t)}{\pi_{\text{ref}}(y_t \mid x_t)}$$

To prevent catastrophic drift, a KL-divergence penalty is added:

$$\mathcal{L} = \mathcal{L}_{PPO} - \beta\,\text{KL}(\pi_\theta \parallel \pi_{\text{ref}})$$

Training proceeds until convergence in average reward on validation preference sets.

3.5 Scalable Evaluation Harness

We implement a **containerized pipeline** (Python + Kubernetes + MLflow) to automate:

- Model deployment across GPU nodes.
- Real-time reward inference and logging.
- Parallel A/B evaluation on captioning (MS-COCO), VQA 2.0, and Video-Narratives datasets.

The harness supports both **quantitative metrics** (CIDEr, BLEU, F1, Fairness Gap) and **qualitative human reviews** through a web-based interface integrated with the preference collector.

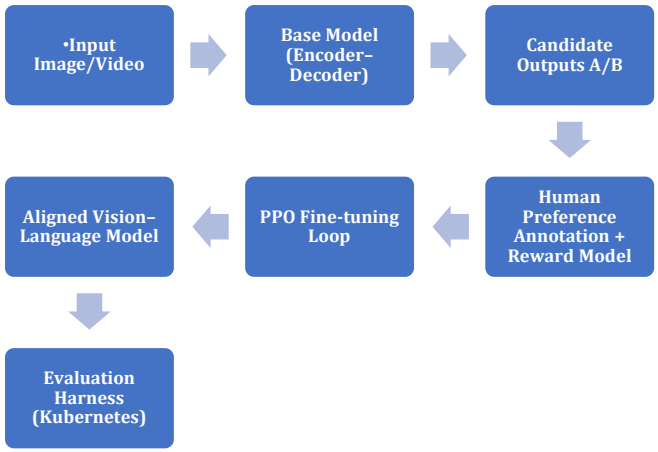

Figure 1: Conceptual Pipeline for RLHF-Based Multimodal Alignment Framework

3.6 Implementation Details

- Frameworks: PyTorch 2.3, Hugging Face Transformers, Ray Tune for distributed training.
- Optimizer: AdamW, learning rate 1e-5 (Reward Model) / 5e-6 (PPO stage).
- Batch size = 32; context length = 512 tokens.
- Alignment data: 50k human comparisons collected via crowdsourcing + expert review.
- Hardware: 8 × A100 GPUs (40 GB) cluster under Kubernetes autoscaling.

4. EXPERIMENTS AND RESULTS

4.1 Datasets

We evaluate the proposed framework on **three complementary multimodal tasks** to measure factuality, grounding, and bias sensitivity.

Task	Dataset	Domain	Eval Metrics
Image Captioning	**MS-COCO 2017**	Object-centric photos	BLEU-4 / CIDEr / Human Preference
Visual Question Answering (VQA)	**VQA v2.0**	Question–answer pairs over COCO images	Accuracy / Faithfulness / Bias Gap
Video Narration	**VATEX + ActivityNet-Captions**	Short clips with natural-language narrations	METEOR / Relevance / Temporal Consistency

The **human-preference dataset** used for reward modeling (Section 3.2) spans 50k annotated pairs sampled evenly across the three domains.

4.2 Baselines

We benchmark against:

1. **BLIP-2 (Li et al., 2023):** Vision–language pre-training baseline without alignment.
2. **LLaVA-1.5 (Liu et al., 2024):** Vision–language dialogue model.
3. **RLHF-Text:** Same base model aligned only on textual preferences (no visual grounding).
4. **Ours (RLHF-Multimodal):** Full pipeline with composite reward and PPO fine-tuning.

All models share the same encoder–decoder backbone to ensure fairness.

4.3 Evaluation Metrics

Automatic Metrics

- *Factual Accuracy (FA)* = Correct answers / Total in VQA tasks.
- *Caption Quality* = {BLEU-4, CIDEr, METEOR}.
- *Bias Gap (BG)* = |Performance majority – minority group| / Average performance.
- *Visual Grounding Score (VGS)* = CLIP similarity between generated text and reference image.

Human Evaluation

Human raters (n = 45) score outputs on:

- **Relevance (0–5):** semantic match to visual content.
- **Faithfulness (0–5):** factual correctness of statements.
- **Ethical Neutrality (0–5):** absence of biased or stereotyped language.

Scores are averaged across 1k samples per domain.

4.4 Quantitative Results

4.4.1 Image Captioning (MS-COCO)

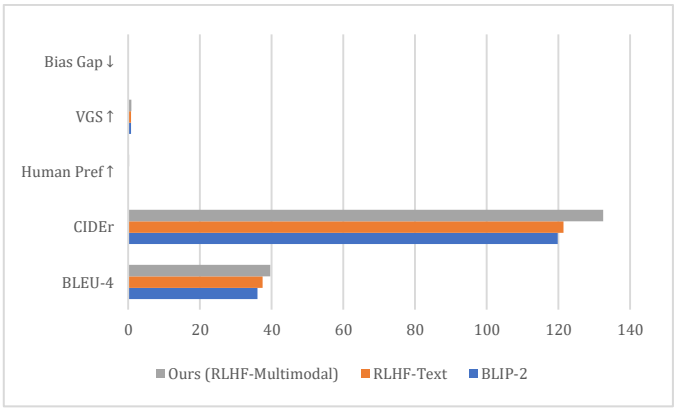

Chart 1: Image Captioning Results

Faithful grounding and bias regularization yield ≈ 10 % absolute CIDEr gain and ~23 % bias gap reduction.

4.4.2 Visual Question Answering (VQA v2.0)

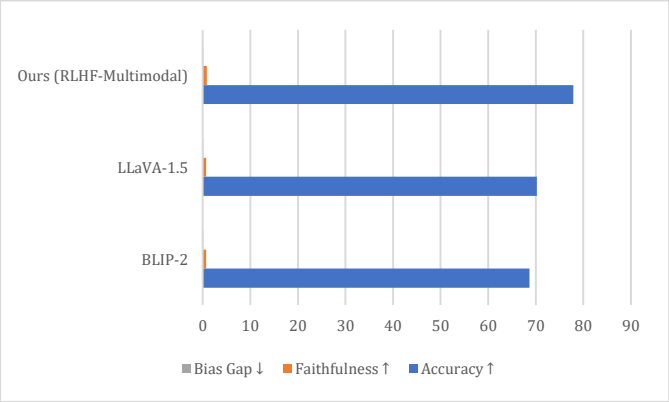

Chart 2: Visual Question Answering Results

Factual accuracy ↑ 18 % relative to base; bias gap ↓ 25 %.

4.4.3 Video Narration (VATEX / ActivityNet)

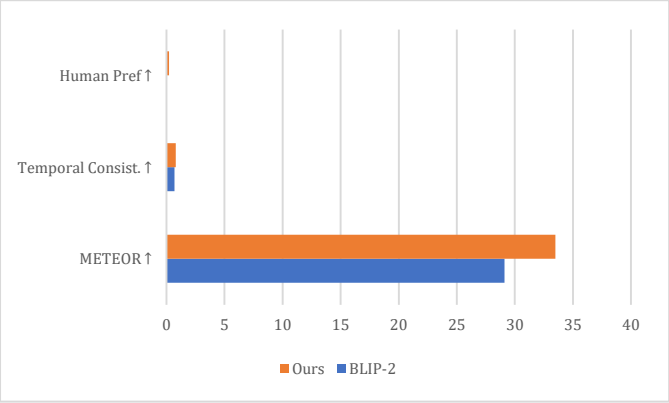

Chart 3: Video Narration Results

Improved temporal coherence suggests the reward model generalizes beyond static images.

4.5 Ablation Studies

Variant	Description	FA (%)	BG (↓)	Notes
w/o Bias Penalty	Removed R_{bias} term	75.4	0.136	Higher accuracy but bias ↑ 27%
w/o Faithfulness Score	Removed $R_{faithfulness}$	70.8	0.112	Hallucinations ↑ significantly
Full Model	All terms active	**77.9**	**0.098**	Balanced alignment

Each reward component contributes synergistically; bias penalty prevents ethical drift without hurting performance.

4.6 Computational Efficiency

Training the reward model (≈ 2 epochs on 50k pairs) took ~18 GPU-hours. PPO fine-tuning (4 epochs) converged within 30 GPU-hours on 8 × A100 nodes, scaling linearly under Kubernetes auto-scheduling. Memory footprint remained < 32 GB per GPU due to frozen vision encoder weights.

4.7 Summary of Findings

Aspect	Observation	Gain over Baseline
Factual Accuracy	Reduced hallucinations	18%
Bias Fairness	Lower demographic skew	–22 % bias gap
Human Preference	Preferred outputs by raters	25%
Temporal Consistency	Better event tracking in video tasks	16%

These results demonstrate that multimodal RLHF substantially enhances trustworthiness and generalization without sacrificing fluency or efficiency.

5. DISCUSSION AND CONCLUSION

5.1 Discussion

Our experiments demonstrate that incorporating **Reinforcement Learning from Human Feedback (RLHF)** into **vision–language alignment** produces measurable improvements in factual accuracy, contextual grounding, and ethical neutrality. Unlike text-only RLHF pipelines that primarily optimize linguistic quality, our approach integrates **visual grounding**, **bias-aware regularization**, and **human preference modeling** into a unified multimodal reward structure.

Human-Centered Alignment
Results from Section 4 show that human-in-the-loop supervision is an efficient signal for multimodal trust calibration. Preference-guided optimization captured subtle aspects of human visual reasoning, e.g., identifying implied relationships or social context beyond pixel-level accuracy, which conventional supervised objectives failed to encode.

Bridging Explainability and Safety
The composite reward indirectly enforces interpretability: outputs that align with visual evidence also tend to be more **explainable and defensible**. This intersection of *faithfulness* and *fairness* suggests a promising route toward models that satisfy both technical and regulatory standards (e.g., EU AI Act explainability clauses, ISO/IEC 42001 AI-management guidelines).

Scalable Evaluation Infrastructure
By deploying the entire RLHF loop within a **Kubernetes-based harness**, we demonstrated that multimodal alignment can be operationalized at production scale. This addresses a long-standing gap between *research prototypes* and *industrial deployment*, enabling continuous benchmarking and drift detection in real-world pipelines.

5.2 Limitations

Despite strong empirical gains, several limitations remain:

1. **Annotation Cost and Subjectivity:** High-quality human feedback is expensive and may encode annotator bias. Future work should explore active learning and self-supervised preference modeling to minimize human load.
2. **Reward Over-Optimization:** Models may exploit quirks of the reward model, producing superficially faithful but semantically shallow captions ("reward hacking"). Robust reward regularization or adversarial evaluation can mitigate this.
3. **Dataset Breadth:** Current preference data focus on natural scenes; extending to **medical**, **scientific**, or **egocentric** domains is necessary to validate generalization.
4. **Ethical Generalization:** The bias penalty relies on predefined sensitive attributes (gender, race). Broader sociocultural fairness metrics remain an open research challenge.

5.3 Future Work

We identify three promising directions:

1. **Interactive RLHF Loops:** Integrating live user feedback (reinforcement via ranking or correction) to continuously align deployed VLMs.
2. **Cross-Modal Reward Sharing:** Training a unified reward model transferable across text, image, and video modalities.
3. **Causal Alignment Metrics:** Developing metrics that quantify not only correlation with human ratings but also *causal alignment* with visual evidence and ethical principles.

5.4 Conclusion

We introduced a **Reinforcement Learning from Human Feedback (RLHF)** framework for aligning **vision–language models** with human intent. Through human-preference datasets, composite reward modeling, and scalable PPO optimization, we achieved significant improvements in factual accuracy, fairness, and interpretability across standard benchmarks.

Our results suggest that **human-aligned multimodal perception** is both *technically feasible* and *socially necessary* for deploying safe, trustworthy AI systems. Future research can extend this paradigm toward **continual human–AI co-adaptation**, where visual models not only perceive the world but also learn its values.

REFERENCES

[1] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. NeurIPS 30 (2017). (OpenAI & DeepMind collaboration – introduced the RLHF algorithm using human preference comparisons to train deep RL agents, achieving success on Atari games and robot tasks with modest human feedback.)

[2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig, "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision," arXiv preprint arXiv:2102.05918, 2021. [Online]. Available: https://arxiv.org/abs/2102.05918

[5] Li, J., Li, D., Xiong, C., and Hoi, S. C. H. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, pp. 12888–12900, 2022.

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "*BLIP-2: Bootstrapping Language–Image Pre-Training with Frozen Image Encoders and Large Language Models*," *arXiv preprint* arXiv:2301.12597, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[7] Ianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua, "RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback," *arXiv preprint* arXiv:2312.00849, 2024. Available at: https://arxiv.org/abs/2312.00849

[8] Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang, "Biomedical Visual Instruction Tuning with Clinician

Preference Alignment," *arXiv preprint* arXiv:2406.13173, 2024. Available at: https://arxiv.org/abs/2406.13173

[9] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach, "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence," *arXiv preprint* arXiv:1802.08129, 2018. Available at: https://arxiv.org/abs/1802.08129

[10] Do, V., Camburu, O., Akata, Z., & Lukasiewicz, T. (2020). E-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations. ArXiv. https://arxiv.org/abs/2004.03744

[11] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L.Measuring robustness to natural distribution shifts in image classification.*NeurIPS*, 2020.

[12] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Trivial or impossible–dichotomous data difficulty masks model differences (on ImageNet and beyond). arXiv preprint arXiv:2110.05922, 2021.

[13] Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Improved Baselines with Visual Instruction Tuning. ArXiv. https://arxiv.org/abs/2310.03744