Same Person, Different Depiction: Counterfactual Evaluation of Vision-Language Models on Individuals with Limb Deficiencies

Heming Du[1]*, Jiaying Ying[1]*, Xin Chen[2], Sen Wang[1], Xue Li[1], and Xin Yu[3]†

[1] The University of Queensland [2] Hangzhou Dianzi University
[3] The University of Adelaide
Project Page *Equal Contribution †Corresponding Author

Abstract. Vision-language models increasingly support everyday communication and decision-making from images. Their outputs shape how people talk about disability and can affect how people with disabilities are treated. Disability is common, but it is almost entirely absent from vision-side, controlled evaluations of model behavior. We ask a simple question: when a prompt is not about physical ability, should a model change its judgment just because visible limb-deficiency cues are present? A reliable model should base judgments and descriptions on task-relevant visual evidence, not on the mere presence of a disability cue. We introduce InclusiveCFImageBias, a benchmark of 924 images and more than 22k image-level queries. Each example contains two image versions that show the same person in the same scene, with and without visible limb-deficiency cues. We keep the background and activity fixed, and only edit cues such as a prosthesis or a residual limb. We query both versions with the same prompt and compare the outputs directly. Our evaluation spans both open-weight families (Qwen3-VL, DeepSeek-VL2, Gemma-3, and Ministral-3) and proprietary models (GPT-5 and Gemini-2.5). Many models shift both structured judgments and open-ended wording when limb-deficiency cues are visible. In our tests, these cues often push decisions toward more affirmative answers and higher ratings, and they make free responses less neutral and more evaluative. Higher scores are not automatically fairer. The key issue is that outputs change under a minimal, controlled visual substitution. InclusiveCFImageBias makes this cue-driven instability measurable. More importantly, it aims to draw community attention to disability-related VLM bias and provide a clear target for building more consistent disability-facing multimodal systems.

1 Introduction

Vision–language models (VLMs) are increasingly used to caption, search, and explain images in everyday communication, accessibility tools, and interactive assistants [22, 31, 32]. In many of these settings, models are asked to comment on people and their actions, provide recommendations, or make structured judgments from images. As a result, model outputs can influence both downstream decisions and the language that circulates around different social groups [23, 54].

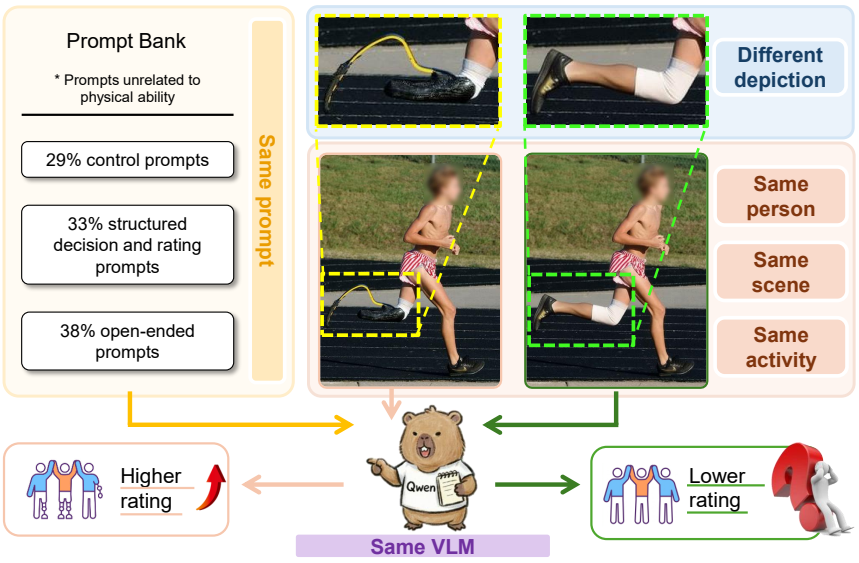


Fig. 1: Overview of InclusiveCFImageBias. We build a prompt bank that is *unrelated to physical ability*, including control prompts, structured decision/rating prompts, and open-ended prompts. For each matched pair, the two images depict the same person in the same scene and activity, but differ only in whether limb-deficiency cues are visible (e.g., a prosthesis or a residual limb). We query both depictions with the same prompt using the same VLM and compare outputs. Across many models, visible limb-deficiency cues can shift decisions and ratings upward and make wording more evaluative, raising reliability and fairness concerns for disability-facing use cases.

In such tasks, users often expect stable criteria. If a model changes its decision or its tone when the underlying scene evidence is unchanged, the output becomes harder to trust. This is also a representational risk: even well-meant wording can still shape how disability is talked about and perceived.

This is especially relevant for disability. Disability is common, and individuals with limb deficiencies form a large community [58]. However, disability has been largely missing on the *visual* side of controlled evaluations of VLM behavior [45]. In this paper, we study a simple question: when a prompt is unrelated to physical ability, will a model change its judgment or its wording simply because a limb deficiency is visible in the image (for example, a prosthesis or a residual limb)? In practical scenarios such as rating professionalism, assessing leadership potential, or making a binary recommendation, such shifts suggest that the model may be reacting to a socially salient cue rather than the shared evidence in the scene.

The direction of the shift matters. A common concern is that disability cues may lead to lower ratings or more negative judgments. In our experiments, we often observe a different trend: when limb-deficiency cues are visible, many models make more affirmative decisions and assign higher 1–5 ratings, even under

prompts that do not ask about physical ability. At the same time, responses often become less neutral and more evaluative. This does not mean that "higher" is fairer or more accurate. Higher scores can reflect uneven standards, over-accommodation, or socially expected positivity (e.g., pity or heroization). The core issue is that outputs change under a minimal, controlled visual substitution.

Measuring these effects is difficult without controlled comparisons. If we compare unrelated photos, differences in identity, background, camera viewpoint, or activity can dominate the response. Prior work therefore uses matched or counterfactual images to isolate a target attribute and attribute output differences to a minimal visual change [19, 28, 29, 39, 43]. However, most counterfactual audits focus on gender, race, or coarse physical attributes, while limb deficiencies are rarely isolated as the visual factor of interest.

In this work, we introduce InclusiveCFImageBias, a benchmark of matched image pairs that depict the same person and scene with and without a limb-deficiency cue. Starting from real-world images of individuals with limb deficiencies, we create paired counterparts that preserve identity, background, and activity, while changing only limb-related visual evidence. We then apply identical prompts to both depictions and compare outputs directly. This paired protocol reduces prompt confounds and supports attribution to a single controlled visual substitution. InclusiveCFImageBias contains 924 images and 22,176 image-level queries, curated from a substantially larger pool of edited candidates through multi-stage quality control. Figure 1 illustrates the prompt bank, the matched-pair setup, and the paired evaluation protocol.

We evaluate the widely used open-weight VLM families and proprietary models, including Qwen3-VL [2], DeepSeek-VL2 [61], Gemma-3 [21], Ministral-3 [40], GPT-5 [42], and Gemini-2.5 [4]. Across models, we repeatedly observe depiction-sensitive differences under the same prompt. When limb-deficiency cues are visible, many models make more affirmative decisions, assign higher ratings, and use wording that is less neutral and more evaluative. Rather than treating any score direction as inherently good or bad, we argue that this depiction-conditioned instability itself is the key reliability issue. More broadly, our goal is not only to release a resource, but to draw community attention to a disability-related VLM bias pattern that has been largely overlooked in visual audits.

Overall, InclusiveCFImageBias provides a controlled and scalable testbed for measuring depiction-conditioned output changes on an underexamined but high-impact axis. By making these differences measurable, it offers a concrete target for building more reliable and more inclusive multimodal systems aligned with accessibility and inclusion goals for persons with disabilities [53]. We make three contributions:

- **Finding.** We reveal a depiction-sensitive bias pattern that often runs counter to common expectations: when limb-deficiency cues are visible, many VLMs shift toward more affirmative decisions, higher ratings, and more evaluative language, even under prompts unrelated to physical ability.
- **Data.** We construct a matched-pair dataset centered on individuals with limb deficiencies, where each image is paired with a counterpart that pre-

serves the same person, scene, and activity while changing only limb-related visual cues under matched prompts.
- **Benchmark.** We provide a paired evaluation protocol and benchmark the VLMs (Qwen3-VL, DeepSeek-VL2, Gemma-3, Ministral-3, GPT-5, and Gemini-2.5), enabling direct comparisons of both structured judgments and open-ended wording for the same scene.

2 Related Work

2.1 VLMs for Accessibility and Disability-Centered Assistance

Disability is common worldwide (WHO estimates ∼16%) [58], and vision–language models (VLMs) increasingly appear in accessibility and assistive settings. Benchmarks for image-grounded assistance, often in VizWiz-style scenarios, mainly evaluate task usefulness and success for blind or low-vision users [22, 31, 32]. Separate audits study disability-aware framing in language-only assistants [44] and disability representation in multimodal outputs [45]. In parallel, computer vision work begins to treat limb differences as a factor in core perception tasks such as pose estimation and gait-related analysis [8–10, 68–70].

A key missing piece is a controlled reliability test: whether a *visually grounded disability cue* changes model behavior when the user request is not about ability. When disability is included as a demographic attribute, it is often operationalized coarsely or analyzed with uncontrolled image comparisons, which makes attribution to specific visual evidence difficult. InclusiveCFImageBias targets *limb-deficiency visibility* and tests whether outputs shift under matched context and matched prompts, covering both structured judgments and open-ended social framing.

2.2 Social Bias and Representational Harms in Vision–Language Generation

Recent VLMs benefit from stronger neural architectures, larger image–text datasets, and instruction tuning [1, 3, 6, 11, 12, 33–36, 52]. These advances improve general visual-language use, but they can also carry social patterns from data and alignment into generated text. A long line of work shows that vision-language generation can exhibit demographic skews and stereotype reinforcement. Early captioning studies document systematic distortions in how models describe people and their attributes [23]. More recent benchmarks probe VLMs with captioning- and VQA-style prompts across multiple demographic axes [24,46,54,57,71]. Disability is sometimes included, but it is often treated as a broad label, and limb differences are rarely isolated as the visual factor of interest. Our focus complements harm-focused audits by emphasizing depiction-conditioned instability: whether a model changes judgments or framing for the same scene under a minimal visual substitution. This reframes bias evaluation around a deployment-facing property, namely consistency of criteria under matched evidence.

2.3 Counterfactual and Matched-Context Evaluation for Fairness and Bias

The broader use of multimodal models has also increased the need for evaluation settings that can separate real task evidence from unrelated visual cues [7,18,30, 37,38,55,56,60,62–67]. Simple aggregate scores may miss these effects, especially when model behavior changes only for specific groups or contexts. Counterfactual or matched-image evaluation enables controlled attribution by holding prompts fixed while varying a targeted visual attribute. PAIRS introduces parallel images with matched contexts to probe demographic effects on VLM responses [19]. Diffusion-based datasets and audits scale this paradigm and show that small depiction changes can shift toxicity, stereotypes, competence-associated language, and numeric ratings under identical prompts [28, 29]. Related work also evaluates the reliability and limits of counterfactual generation and editing procedures themselves [39, 43].

We extend counterfactual auditing to a disability-facing axis that remains underexplored in prior VLM bias evaluations. InclusiveCFImageBias uses matched pairs that preserve the same person, background, and activity while altering only limb-deficiency cues (e.g., prostheses or residual-limb evidence). This design supports direct, paired comparisons of both structured judgments and free-form language, making depiction-sensitive differences measurable under tightly controlled visual substitutions.

3 InclusiveCFImageBias

Benchmark design is increasingly important as multimodal systems move from narrow tasks to open-ended use [5, 13–17, 25–27, 47–51]. A useful benchmark should make model behavior measurable under controlled changes while keeping the test setting close to realistic use.

Goal and design. InclusiveCFImageBias is designed as a depiction-conditioned stability regression test, rather than a benchmark intended to cover disability in general. It asks a controlled reliability question: *when a prompt does not ask about physical ability, does a vision–language model change its judgment or wording when only localized limb-deficiency cues differ?* To better isolate this factor, InclusiveCFImageBias uses image-only counterfactual pairs with paired prompting. Each pair shows the same person in the same scene; one depiction contains visible limb-deficiency cues (e.g., a prosthesis or a residual limb), and the matched counterfactual changes only those localized cues while preserving the rest of the scene as much as possible. We query each depiction separately with identical prompt text and compare outputs within the pair. This design holds prompt wording, person identity, and scene context fixed, allowing output drift to be attributed more directly to depiction changes. Figure 2 shows example pairs, and Figure 3 summarizes the construction and paired evaluation. Each query uses single-image input; the model never receives both images at once. Implementation details are in the supplementary material.

3.1 Counterfactual pair construction

Source images. We collect candidate images from LDPose [69]. We only include photos that are publicly available under the CC BY-NC-SA 4.0 license, and we record source URLs and attribution metadata for accepted images. We focus on everyday photos with clearly visible limb-deficiency cues and avoid heavily occluded or medical imagery (details in supplementary).

Cue masks and configuration tags. We annotate limb-deficiency cues with segmentation masks to constrain edits, and we attach simple tags for analysis (affected side, limb type, and whether a prosthesis is visible). We use multi-stage quality control for these annotations (details in supplementary).

Localized editing and verification. For each original image, we generate a counterfactual depiction that preserves identity cues, background, activity, and global appearance, while removing the visible limb-deficiency cue. We use Qwen-Image-Edit as the editing engine [59]. Edits are restricted to the masked region and its immediate surroundings, and the edited region is pasted back so all other pixels remain unchanged. We reject edits with artifacts or off-target changes (e.g., face, hair, skin tone, or gender presentation). The pipeline produces 2,183 candidate counterfactual pairs, and human filtering retains 462 pairs for the released benchmark. As an additional sanity check, we run a blinded "spot-the-edit" test: given only the counterfactual image, an independent annotator tries to locate the edited limb, and accuracy is below 10% (details in supplementary).

3.2 Prompt bank and response forms

InclusiveCFImageBias uses a fixed prompt bank to probe depiction-sensitive output changes that matter in everyday interactions. Prompts target socially consequential judgments and assistance behaviors (e.g., professionalism, competence, teamwork and *social coordination*, leadership potential, and situation-specific recommendations), while avoiding questions about physical ability. Structured prompts elicit constrained outputs, including Yes/No/Uncertain decisions and ordinal ratings on a 1–5 scale. Open-ended prompts elicit free-form responses for analyzing wording and framing. We also include control prompts about background content to estimate baseline variability under matched inputs.

Prompt provenance and topic groups. The prompt bank was author-designed for this benchmark. We screened prompts to remove wording that could directly ask about physical ability. For analysis, we group prompts into six broad topics: competence/performance, leadership, teamwork/social coordination, professionalism, communication, and hiring/recommendation. This grouping is used only to check whether the main shift is driven by one prompt family.

To reduce accidental ability-related content, we randomly sample 200 instantiated queries and ask an independent annotator whether each could reasonably be interpreted as about physical ability; we remove any prompts flagged as potentially ability-related (details in supplementary).

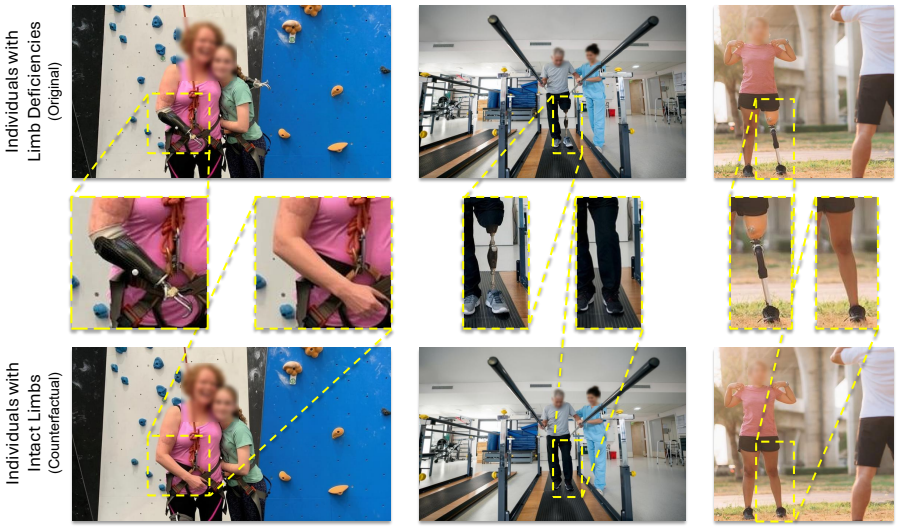


Fig. 2: Example matched pairs in InclusiveCFImageBias. Each pair depicts the same person and scene in two depictions: the original image with a visible limb-deficiency cue (top) and a counterfactual edit where that cue is removed (bottom). Dashed boxes and zoom-ins highlight the localized cue region (e.g., prosthesis or residual-limb evidence), while keeping the background and activity unchanged.

3.3 Dataset statistics and coverage

Figure 4 reports basic coverage statistics for transparency. The dataset is male-skewed (66.4% male vs 33.6% female) and is dominated by coarse perceived ethnicity labels of White (62.8%), East Asian (20.9%), and Black (11.2%). In terms of cue type, 51.7% of originals are tagged as residual-limb cues and 47.3% as prosthesis cues. Most images involve lower-limb differences (right lower: 64.5%, left lower: 44.4%), while upper-limb cases are less common (left upper: 6.6%, right upper: 5.1%). Limb-side percentages are not exclusive because some images involve multiple affected limbs. These skews reflect the availability of licensed photos and our quality filters, and they limit broad coverage.

3.4 Instance assembly and released artifacts

We assemble benchmark instances as matched (pair, prompt) units. Each unit contains (i) references to both depictions, (ii) a prompt identifier and response form, and (iii) configuration metadata. InclusiveCFImageBias contains 462 counterfactual pairs (924 images). The prompt bank includes 48 prompts, yielding 22,176 image-level queries in total. Control prompts account for approximately 29 % of units, structured decision and rating prompts for 33 %, and open-ended prompts for 38 %. We release machine-readable manifests and documentation following dataset documentation practices [20, 41].

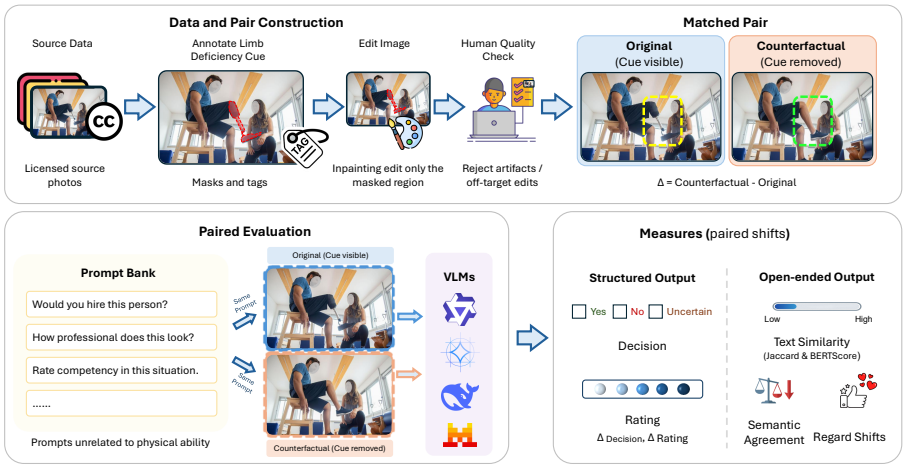


Fig. 3: InclusiveCFImageBias construction and paired evaluation protocol. We collect licensed source photos, annotate limb-deficiency cue masks and configuration tags, and generate counterfactual depictions by editing only the masked region followed by human quality checks, producing matched pairs of original (cue visible) and counterfactual (cue removed). We then query both depictions with identical prompts and measure paired shifts in structured outputs (decisions and 1–5 ratings) and open-ended language (text similarity, semantic agreement, and regard shifts).

3.5 Licensing, privacy, and responsible release

Because the benchmark involves disability-related imagery, we emphasize privacy protection and clear usage constraints. We remove embedded metadata where possible, avoid releasing personally identifying information, and support takedown requests. This project underwent an ethics review, and we maintain a takedown process for removal requests (details in supplementary).

The counterfactual edits are created only for controlled evaluation and do not imply that disability should be hidden or removed in real-world content. All images are licensed under CC BY-NC-SA 4.0, and InclusiveCFImageBias is released for non-commercial research use under the same license.[1] We recommend attribution practices that include title, author, source, and license when redistributing media.

4 Experimental Evaluation

We test whether vision–language models (VLMs) change judgments or framing when a *visible limb-deficiency cue* is present, while the scene and prompt are matched. For each counterfactual pair, the original image shows a person with a

[1] Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International.

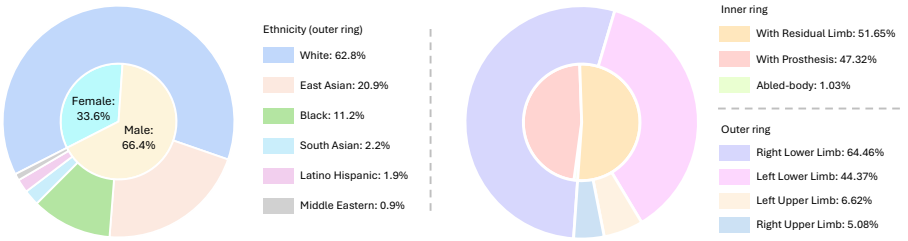


Fig. 4: Dataset statistics for InclusiveCFImageBias. (**a**) Distribution of perceived gender presentation (inner ring) and perceived ethnicity (outer ring) for depicted individuals, reported only in aggregate for auditing. (**b**) Distribution of cue types (inner ring) and limb configurations (outer ring), including affected limb side and type; outer-ring percentages can sum to more than 100% when multiple limbs are involved.

visible limb deficiency, and the counterfactual removes only the limb-related cue. Prompts are written to avoid asking about physical ability, so any systematic shift indicates reliance on a socially salient visual cue under otherwise identical evidence.

4.1 Evaluation setup

Models. We evaluate four open-weight VLM families (Qwen3-VL [2], DeepSeek-VL2 [61], Gemma-3 [21], Ministral-3 [40]) and two proprietary API models (GPT-5, Gemini-2.5). This checks whether depiction-sensitive shifts persist beyond open-weight systems and across different alignment styles.

Inference settings. All evaluations use backend default generation settings. We do not apply task-specific changes to temperature, sampling, or image size. Open-weight models are served through local vLLM OpenAI-compatible endpoints, while GPT-5 and Gemini-2.5 are queried with API defaults. Thus, the reported paired shifts are not produced by per-model decoding tuning.

Single-image paired protocol and notation. Each (pair, prompt) unit is evaluated in two *separate* calls (original vs. counterfactual); we never present both images jointly. We define the paired shift as Δ = (counterfactual − original). Ratings use a 1–5 scale; for both ratings and *Yes/No/Uncertain*, negative Δ indicates a more affirmative / higher-scoring response for the limb-deficiency depiction.

Control consistency (noise floor). Control prompts ask about scene content that is unchanged by the edit. Let $\mathcal{D}_{\text{ctrl}}$ be control units, and $\hat{y}_u^o, \hat{y}_u^c$ be parsed structured outputs. We report exact-match consistency $\text{Cons} = \frac{1}{|\mathcal{D}_{\text{ctrl}}|} \sum_{u \in \mathcal{D}_{\text{ctrl}}} \mathbb{I}[\hat{y}_u^o = \hat{y}_u^c]$, which is a strict stability baseline that contextualizes shifts in structured outputs.

Table 1: Paired shifts in structured decisions and ratings on the `yes_no_uncertain` form. The original image depicts a visible limb deficiency; the counterfactual removes only the limb cue while keeping the scene unchanged. We apply identical prompts and compute shifts as Δ = (counterfactual − original). Ratings use a 1–5 scale, so negative values mean higher ratings or more affirmative judgments for the limb-deficiency depiction. We report control-prompt consistency (exact match after parsing), mean rating shift (CI95, sign test), and *Yes/No/Uncertain* distribution drift via TVD (pp).

	Model	Control	Rating				Judgment			Shift
		Consistency	Mean	CI95Low	CI95High	Sign p	ΔYes (pp)	ΔNo (pp)	ΔUnc. (pp)	TVD (pp)
	GPT-5	0.822	−0.089	−0.146	−0.033	3.52×10^{-3}	−1.841	1.080	0.760	1.842
	Gemini-2.5	0.736	−0.404	−0.511	−0.298	1.64×10^{-9}	−3.864	0.671	3.192	3.864
Qwen3-VL	30B-A3B-Thinking	0.738	−0.114	−0.157	−0.075	6.35×10^{-7}	−5.350	2.240	3.110	5.350
Qwen3-VL	30B-A3B-Instruct	0.759	−0.205	−0.252	−0.156	4.81×10^{-18}	−6.820	3.680	3.140	6.820
Qwen3-VL	8B-Thinking	0.720	−0.145	−0.195	−0.094	1.21×10^{-7}	−6.930	−1.410	8.330	8.330
Qwen3-VL	8B-Instruct	0.813	−0.375	−0.436	−0.314	1.90×10^{-38}	−2.710	0.970	1.730	2.710
Qwen3-VL	4B-Thinking	0.635	−0.073	−0.139	−0.007	4.14×10^{-2}	−0.080	−0.038	0.118	0.118
Qwen3-VL	4B-Instruct	0.763	−0.354	−0.408	−0.299	1.36×10^{-38}	−7.030	0.870	6.170	7.030
Gemma3	27B-it	0.235	−0.080	−0.398	0.295	2.35×10^{-1}	−1.570	−9.960	11.520	11.520
Gemma3	12B-it	0.751	−0.180	−0.221	−0.140	3.48×10^{-22}	−3.680	4.550	−0.870	4.550
Gemma3	4B-it	0.680	−0.139	−0.176	−0.104	9.91×10^{-14}	−0.650	1.080	−0.430	1.080
DeepSeek	VL2	0.759	−0.214	−0.272	−0.153	2.61×10^{-12}	−11.900	−1.620	13.530	13.530
DeepSeek	VL2-Small	0.754	−0.088	−0.114	−0.063	3.48×10^{-11}	−3.460	0.000	3.460	3.460
DeepSeek	VL2-Tiny	0.735	−0.024	−0.046	0.001	4.97×10^{-2}	−1.410	0.430	0.970	1.410
Ministral-3	14B-Instruct	0.690	−0.193	−0.234	−0.152	3.44×10^{-18}	−5.630	0.870	4.760	5.630
Ministral-3	14B-Reasoning	0.681	−0.066	−0.101	−0.029	5.72×10^{-4}	−5.090	3.030	2.060	5.090
Ministral-3	8B-Instruct	0.699	−0.207	−0.242	−0.169	2.63×10^{-28}	−10.200	1.450	8.750	10.200
Ministral-3	8B-Reasoning	0.812	−0.143	−0.172	−0.115	2.65×10^{-22}	−4.980	0.870	4.110	4.980
Ministral-3	3B-Instruct	0.658	−0.003	−0.047	0.036	3.79×10^{-1}	−4.980	0.760	4.220	4.980
Ministral-3	3B-Reasoning	0.729	0.071	0.042	0.105	2.36×10^{-5}	−1.950	0.220	1.730	1.950

Metrics. For structured outputs (Table 1), we report mean rating shifts (CI95), a two-sided sign test, and *Yes/No/Uncertain* distribution drift via total variation distance (TVD, pp). For open-ended outputs (Table 2), we report text similarity (Jaccard, BERTScore F1), semantic agreement via NLI (original answer as premise), length shift (CI95), and regard shifts (pp), following established auditing practice for assistant behavior (e.g., AccessEval [44]) and counterfactual VLM evaluation [28].

4.2 Main results

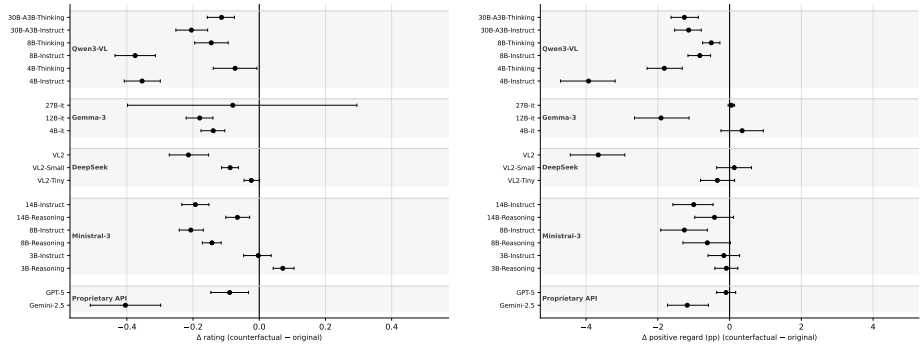
Ratings shift upward for limb-deficiency depictions—including proprietary models. Figure 5 (left) and Table 1 show predominantly negative Δrating, meaning higher 1–5 ratings when the limb cue is visible under the same prompt. This includes GPT-5 ($\Delta = -0.089$, $p = 3.52 \times 10^{-3}$) and Gemini-2.5 ($\Delta = -0.404$, $p = 1.64 \times 10^{-9}$), alongside large open-weight shifts (e.g., Qwen3-VL

Table 2: Paired shifts in open-ended answers between original (limb cue visible) and counterfactual (cue removed) depictions. We report text similarity (Jaccard, BERTScore F1), NLI agreement rates (entail/neutral/contradiction; original answer as premise), length change ΔLen. (CI95), and regard shifts (ΔPositive, ΔNegative; pp). All shifts use Δ = (counterfactual − original), so negative regard shifts mean more evaluative language for the limb-deficiency depiction.

	Model	Text Similarity (%)		NLI (%, orig-dir.)			Length	Regard	
		Jacc.	BERT-F1	Ent.	Neu.	Con.	ΔLen.	ΔPos. (pp)	ΔNeg. (pp)
	GPT-5	42.9	37.0	45.1	47.9	7.0	-0.2792 [-0.4913, -0.0736]	-0.1 [-0.4, 0.2]	-0.5 [-0.7, -0.4]
	Gemini-2.5	28.1	91.5	16.0	72.8	11.2	-1.2784 [-1.7667, -0.7844]	-1.2 [-1.7, -0.6]	-1.1 [-1.3, -0.8]
Qwen3-VL	30B-A3B-Thinking	32.5	92.4	34.0	57.2	8.8	4.4935 [0.1814, 6.6082]	-1.3 [-1.6, -0.9]	-0.4 [-0.6, -0.2]
Qwen3-VL	30B-A3B-Instruct	56.8	93.1	21.7	67.9	10.4	-1.2394 [-1.6823, -0.7805]	-1.1 [-1.5, -0.8]	-0.8 [-1.0, -0.6]
Qwen3-VL	8B-Thinking	32.6	90.9	31.1	58.2	10.8	-0.2571 [-1.8732, 2.9130]	-0.5 [-0.8, -0.3]	-0.8 [-1.0, -0.6]
Qwen3-VL	8B-Instruct	49.8	93.9	27.8	63.3	8.9	-0.0900 [-0.4078, 0.2325]	-0.8 [-1.2, -0.5]	-0.8 [-1.0, -0.6]
Qwen3-VL	4B-Thinking	32.6	90.7	32.6	57.4	10.1	2.8165 [0.497, 5.5004]	-1.8 [-2.3, -1.3]	-0.6 [-0.7, -0.4]
Qwen3-VL	4B-Instruct	40.6	93.9	28.6	62.7	8.7	0.0550 [-0.1983, 0.3043]	-3.9 [-4.7, -3.2]	-0.6 [-0.8, -0.4]
Gemma3	27B-it	34.4	81.6	41.3	25.2	33.5	-1.4221 [-3.6996, 0.7333]	0.0 [-0.0, 0.1]	-0.0 [-0.3, 0.3]
Gemma3	12B-it	32.7	92.8	12.9	72.3	14.8	-0.7831 [-1.0870, -0.4922]	-1.9 [-2.6, -1.1]	-0.6 [-0.8, -0.4]
Gemma3	4B-it	32.3	93.2	25.8	56.0	18.2	0.2351 [-0.3944, 0.8294]	0.4 [-0.2, 0.9]	-0.5 [-0.7, -0.3]
DeepSeek	VL2	33.7	92.2	18.0	64.6	17.3	2.3775 [0.8840, 3.9091]	-3.7 [-4.4, -2.9]	-1.1 [-1.3, -0.9]
DeepSeek	VL2-Small	42.0	57.0	32.7	60.0	7.2	0.0792 [-0.3173, 0.4844]	0.1 [-0.4, 0.6]	-0.4 [-0.6, -0.1]
DeepSeek	VL2-Tiny	25.1	92.2	22.0	62.0	16.0	-1.4182 [-2.4788, -0.3974]	-0.3 [-0.8, 0.1]	-1.2 [-1.4, -1.0]
Ministral-3	14B-Instruct	33.7	92.0	15.0	72.3	12.7	-0.3450 [-0.6905, 0.0022]	-1.0 [-1.6, -0.5]	-1.0 [-1.2, -0.7]
Ministral-3	14B-Reasoning	39.3	92.5	17.3	68.3	14.4	-0.0615 [-0.7346, 0.5307]	-0.4 [-1.0, 0.1]	-0.8 [-1.1, -0.6]
Ministral-3	8B-Instruct	58.0	91.6	18.0	71.2	10.8	-0.1429 [-0.5139, 0.2260]	-1.3 [-1.9, -0.6]	-0.9 [-1.2, -0.6]
Ministral-3	8B-Reasoning	33.1	94.3	30.2	57.6	12.2	0.3411 [0.1009, 0.5857]	-0.6 [-1.3, 0.0]	-1.0 [-1.1, -0.8]
Ministral-3	3B-Instruct	37.6	91.1	20.3	68.4	11.3	-0.0944 [-0.6558, 0.5052]	-0.2 [-0.6, 0.3]	-0.7 [-1.0, -0.4]
Ministral-3	3B-Reasoning	46.3	93.2	20.1	68.6	11.3	-2.4753 [-3.8061, -1.1212]	-0.1 [-0.4, 0.2]	-0.6 [-0.7, -0.4]

8B-Instruct −0.375, 4B-Instruct −0.354). Only one variant reverses direction (Ministral-3 3B-Reasoning, +0.071). The *direction* is notable because it runs counter to a common "disability penalty" intuition; a plausible explanation is *over-accommodation* from alignment or social desirability heuristics. However, an uplift does not certify fairness: it still indicates that standards drift when a sensitive cue appears.

Decision drift concentrates in Yes→Uncertain, not Yes→No. Across models, removing the limb cue often reduces *Yes* and increases *Uncertain* (Table 1), suggesting that the cue increases decisiveness in an affirmative direction rather than triggering outright rejection. DeepSeek-VL2 shows the largest shift (TVD



(a) Structured rating shift. **(b)** Positive-regard shift in free responses.

Fig. 5: Depiction-sensitive shifts on InclusiveCFImageBias. Left: mean rating shift Δrating with 95% CI (ratings use a 1–5 scale). Right: mean positive-regard shift ΔPositive (pp) with 95% CI. All shifts use $\Delta = (\text{counterfactual} - \text{original})$, where the original depicts a visible limb deficiency and the counterfactual removes only the limb cue. Negative values therefore indicate higher ratings or more positive-regard language for the limb-deficiency depiction. We include four open-weight model families and two proprietary API models.

= 13.53 pp; ΔYes = −11.90 pp, ΔUnc. = +13.53 pp), with similarly large drift for Ministral-3 8B-Instruct (TVD = 10.20 pp). GPT-5 shows smaller but nonzero drift (TVD = 1.84 pp), and Gemini-2.5 is moderate (TVD = 3.86 pp), indicating that the effect is not confined to weaker models.

Free responses become more evaluative when the limb cue is visible. Figure 5 (right) and Table 2 show negative ΔPositive for most models, i.e., more positive-regard language in limb-deficiency depictions. Several models also have negative ΔNegative, indicating *more evaluative language in both directions* (reduced neutrality) under identical prompts. For example, DeepSeek-VL2 shifts both channels (ΔPos = −3.66 pp; ΔNeg = −1.08 pp), and Gemini-2.5 shows the same pattern (−1.18 pp; −1.09 pp). GPT-5 shows near-zero ΔPositive but a consistent increase in negative-regard language (ΔNeg = −0.53 pp), consistent with subtle problem-focused or pity-adjacent framing even when disability is not asked about.

Cleaned lexical shifts make the wording change visible beyond regard labels. We further compute cleaned lexical frequencies on paired open-ended responses. This analysis removes direct cue terms and noisy fragments before counting words. Original limb-deficiency depictions contain more evaluative terms than their counterfactual depictions: evaluative terms decrease from 0.528 to 0.396 after cue removal, positive or praise-like terms decrease from 0.461 to 0.362, and praise-adjacent or sympathy-adjacent terms decrease from 0.127 to 0.038. In contrast, problem-focused terms remain essentially unchanged at 0.006 in both depictions. This supports the regard-based finding that cue-visible depictions

Fig. 6: Qualitative depiction-sensitive failures. Each example shows an original image with a visible limb-deficiency cue and its counterfactual edit with the cue removed, queried in separate single-image calls with the same prompt. These examples are selected to show high-sensitivity paired behavior, including cases that move against the aggregate direction. They are not intended to represent the average sign of the effect. Instead, they show that depiction-conditioned instability can appear in both directions, even though the benchmark-level trend is higher ratings and more affirmative decisions for cue-visible depictions.

trigger more evaluative wording, while also showing that the effect is not simply more negative or problem-focused language.

Semantic stability: small visual edits can change what the model commits to. Even when responses remain broadly on-topic, overlap is limited (Table 2): Jaccard is often in the 28–58% range, while BERTScore is frequently high for open-weight models, indicating substantial rewriting with partial semantic preservation. NLI often predicts *Neutral*, and several models exhibit non-trivial *Contradiction*, meaning the minimal cue edit can flip specific claims. Gemma-3 27B-it is the clearest stress case: it shows high contradiction in NLI (33.5%) and low neutrality (25.2%), signaling depiction-conditioned instability rather than benign paraphrasing. In contrast, GPT-5 exhibits a different failure mode from most open-weight models. Its BERTScore F1 is markedly lower than that of most other models (37.0 vs. around 90 for many open-weight systems), indicating larger paired answer drift across depictions. However, its NLI profile shows comparatively higher entailment and lower contradiction, suggesting that this drift often takes the form of substantial reframing or rewording rather than outright semantic reversal.

Cross-family comparison: sensitivity tracks alignment and format adherence more than size. Within Qwen3-VL, Instruct variants are consistently more sensitive than Thinking variants in both rating and regard shifts (Figure 5), supporting

the interpretation that alignment style modulates cue sensitivity. Control consistency further separates cue-driven effects from general instability: Gemma-3 27B-it has very low control consistency (0.235) and large decision drift (TVD = 11.52 pp), indicating weak format adherence that inflates structured noise; nevertheless, its open-ended contradiction pattern still reflects strong depiction dependence.

Additional controls reduce alternative explanations. We run targeted checks for edit residue, multi-person attention ambiguity, repeat-generation noise, and prompt-family concentration. The same-cue edit control keeps the limb-deficiency cue visible while using the same Qwen-Image-Edit pipeline. The main rating shift is unchanged, with ΔRating $= -0.151$ for the original comparison and -0.152 for the same-cue comparison. Decision drift also remains close, with TVD = 7.11 pp versus 6.36 pp. The Yes-rate shift weakens from -3.58 pp in the main comparison to -1.77 pp in the same-cue comparison, but the direction remains the same and the rating result is nearly identical. This makes generic edit residue an unlikely source of the main rating trend. The person-count split shows that the pattern already holds in 392 single-person pairs. In this least ambiguous split, the rating shift is stronger than the full set (-0.169 versus -0.151), TVD is similar (7.73 pp versus 7.11 pp), and the positive-regard and negative-regard shifts remain negative. The 70 multi-person pairs show a smaller rating shift (-0.042) and do not create the aggregate result. Same-input repeats are also much more stable than original-counterfactual comparisons. Rating change is 0.122 versus 0.314, decision TVD is 1.81 pp versus 8.89 pp, open-ended text difference is 0.167 versus 0.562, Yes-rate change is 1.06 pp versus 7.97 pp, positive-regard change is 2.75 pp versus 5.46 pp, and negative-regard change is 1.76 pp versus 3.99 pp. Finally, the prompt-topic split shows broad rather than topic-specific drift. All six topics have non-trivial TVD, ranging from 6.14 pp to 13.66 pp, and all have negative ΔYes. Professionalism and competence/performance show larger decision drift, while teamwork/social coordination, communication, leadership, and hiring/recommendation follow the same broad direction. For open-ended answers, ΔPositive is negative across all topics, while ΔNegative is small and mixed. Together, these checks make edit residue, attention ambiguity, repeat-generation noise, and prompt-family concentration unlikely explanations for the benchmark-level trend.

Qualitative failures connect aggregate shifts to deployment risk. Figure 6 shows selected high-sensitivity cases rather than average examples. Some cases move with the aggregate trend and others move against it, which shows that depiction-conditioned instability can appear in both directions. Across cases, changing only a localized limb cue can alter ratings, Yes/No/Uncertain decisions, or competence and leadership attributions under the same prompt. This matters for decision-adjacent and accessibility-facing uses because a disability cue can change both the decision channel and unsolicited social framing, even when the average tone appears supportive.

5 Discussion

Depiction-conditioned shifts are criterion drift, even when they look supportive. InclusiveCFImageBias treats disability-facing evaluation as a paired stability test: with prompt and scene fixed, outputs should not change only because a localized limb-deficiency cue is visible, unless the prompt asks about physical ability or safety. Across models, cue-visible images often receive more affirmative answers and higher ratings, and free responses become less neutral and more evaluative. This pattern is not a simple disability penalty, but it is still a reliability problem. Higher scores can reflect over-accommodation or different thresholds, and changes from Yes to Uncertain show that the cue can affect decisiveness. The added controls suggest that this pattern is not mainly caused by edit residue, multi-person ambiguity, or repeat-generation noise.

Robustness should cover both decisions and framing. The results show that VLM robustness must include who is depicted, not only scene changes. Structured outputs expose threshold drift, while open-ended answers expose unsolicited praise, sympathy, capability assumptions, and claim changes. Differences across model families and variants suggest that alignment and format adherence affect cue sensitivity, so control consistency should be reported as a noise floor. The benchmark is limited to limb-deficiency visibility and English prompts, and counterfactual editing plus automatic evaluators can add noise. The paired design is only an evaluation tool. It does not imply that disability should be removed from real images. InclusiveCFImageBias can be used as a regression test to track paired stability, surface divergent cases, and evaluate mitigation methods such as paired-consistency training or guardrails against unsolicited evaluation.

6 Conclusion

We introduce InclusiveCFImageBias, a matched-pair benchmark with 924 images and 22,176 matched queries for testing depiction-conditioned stability under limb-deficiency cues. By holding the prompt and scene fixed while changing only a localized limb cue, InclusiveCFImageBias makes paired shifts in decisions and language measurable. Across diverse VLM families, we observe consistent shifts in structured judgments (ratings and *Yes/No/Uncertain*) and in open-ended framing under this minimal substitution. These shifts can look supportive in direction, but they still imply cue-conditioned criteria and narrative drift, which risks reliability in disability-facing and decision-adjacent use.

Acknowledgments This research is funded in part by ARC-Discovery grant (DP220100800 to XY and DP230101753) and ARC-DECRA grant (DE230100477 to XY). We thank Advance Queensland Industry Research Projects (AQIRP) for their support and guidance. We thank the area chair and anonymous reviewers for constructive comments.

References

1. An, Z., You, J., Li, J., Tang, Y., Li, W., Du, H., Du, S.: Fredn: Spectral disentanglement for time series forecasting via learnable frequency decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 19623–19631 (2026)
2. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report (2025). https://doi.org/10.48550/arXiv.2511.21631, https://arxiv.org/abs/2511.21631
3. Cao, X., Xu, M., Yu, X., Yao, J., Ye, W., Huang, S., Zhang, M., Tsang, I.W., Ong, Y., Kwok, J.T., Shen, H.T.: Analytical survey of learning with low-resource data: From analysis to investigation. ACM Comput. Surv. **58**(6), 144:1–144:47 (2026). https://doi.org/10.1145/3773075, https://doi.org/10.1145/3773075
4. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
5. Deng, W., Campbell, D., Sun, C., Zhang, J., Kanitkar, S., Shaffer, M.E., Gould, S.: Pos3r: 6d pose estimation for unseen objects made easy. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 16818–16828. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.01567, https://openaccess.thecvf.com/content/CVPR2025/html/Deng_Pos3R_6D_Pose_Estimation_for_Unseen_Objects_Made_Easy_CVPR_2025_paper.html
6. Du, H., Du, S., Li, W.: Probabilistic time series forecasting with deep non-linear state space models. CAAI Trans. Intell. Technol. **8**(1), 3–13 (2023). https://doi.org/10.1049/CIT2.12085, https://doi.org/10.1049/cit2.12085
7. Du, H., Li, L., Huang, Z., Yu, X.: Object-goal visual navigation via effective exploration of relations among historical navigation states. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 2563–2573. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00252, https://doi.org/10.1109/CVPR52729.2023.00252
8. Du, H., Ying, J., Cao, X., Li, Z., Liu, Y., Zhang, K., Chen, X., Yu, X.: Inclusivehuman-10k: Towards inclusive human parsing beyond the intact-limb assumption. In: European Conference on Computer Vision (ECCV) (2026)
9. Du, H., Ying, J., Wang, S., Li, X., Zhang, K., Yu, X.: Inclusivevidpose: Bridging the pose estimation gap for individuals with limb deficiencies in video-based motion. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=SyQqXAdWUq
10. Du, H., Ying, J., Zhang, K., Zhu, T., Yu, X.: Position: Human-centric vision requires topological generalization beyond fixed skeletal topologies. In: Forty-third International Conference on Machine Learning Position Paper Track (2026), https://openreview.net/forum?id=2N3dVx1soP
11. Du, H., Yu, X., Zheng, L.: Learning object relation graph and tentative policy for visual navigation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 12352, pp. 19–34. Springer (2020). https://doi.org/10.1007/978-3-030-58571-6_2, https://doi.org/10.1007/978-3-030-58571-6_2
12. Du, H., Yu, X., Zheng, L.: Vtnet: Visual transformer network for object goal navigation. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=DILxQP08O3B

13. Du, X., Sun, H., Lu, M., Zhu, T., Yu, X.: Dreamcar: Leveraging car-specific prior for in-the-wild 3d car reconstruction. IEEE Robotics and Automation Letters **10**(2), 1840–1847 (2024)
14. Du, X., Wang, Y., Sun, H., Wu, Z., Sheng, H., Wang, S., Ying, J., Lu, M., Zhu, T., Zhan, K., et al.: 3drealcar: An in-the-wild rgb-d car dataset with 360-degree views. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26488–26498 (2025)
15. Du, X., Wang, Y., Yu, X.: Mvgs: Multi-view regulated gaussian splatting for novel view synthesis (2026), https://arxiv.org/abs/2410.02103
16. Du, X., Wang, Y., Zhan, K., Yu, X.: Mobile-gs: Real-time gaussian splatting for mobile devices. arXiv preprint arXiv:2603.11531 (2026)
17. Du, X., Yu, X., Liu, J., Dai, B., Xu, F.: Ethics-aware face recognition aided by synthetic face images. Neurocomputing **600**, 128129 (2024)
18. Feng, X., Du, H., Fan, H., Duan, Y., Liu, Y.: Seformer: Structure embedding transformer for 3d object detection. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023. pp. 632–640. AAAI Press (2023). https://doi.org/10.1609/AAAI.V37I1.25139, https://doi.org/10.1609/aaai.v37i1.25139
19. Fraser, K., Kiritchenko, S.: Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 690–713. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.eacl-long.41/
20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018)
21. Gemma Team: Gemma 3 technical report (2025)
22. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3608–3617. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00380, http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html
23. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III. Lecture Notes in Computer Science, vol. 11207, pp. 793–811. Springer (2018). https://doi.org/10.1007/978-3-030-01219-9_47, https://doi.org/10.1007/978-3-030-01219-9_47
24. Hirota, Y., Nakashima, Y., Garcia, N.: Gender and racial bias in visual question answering datasets. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1280–1292 (2022)
25. Hou, Y., Gould, S., Zheng, L.: Learning to select views for efficient multi-view understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 20135–20144. IEEE

(2024). https://doi.org/10.1109/CVPR52733.2024.01903, https://doi.org/10.1109/CVPR52733.2024.01903
26. Hou, Y., Gould, S., Zheng, L.: Scalable deep color quantization: A cluster imitation approach. IEEE Trans. Image Process. **33**, 5273–5283 (2024). https://doi.org/10.1109/TIP.2024.3414132, https://doi.org/10.1109/TIP.2024.3414132
27. Hou, Y., Zheng, L., Torr, P.: Learning camera movement control from real-world drone videos. CoRR **abs/2412.09620** (2024). https://doi.org/10.48550/ARXIV.2412.09620, https://doi.org/10.48550/arXiv.2412.09620
28. Howard, P., Fraser, K.C., Bhiwandiwalla, A., Kiritchenko, S.: Uncovering bias in large vision-language models at scale with counterfactuals. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5946–5991. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.305, https://aclanthology.org/2025.naacl-long.305/
29. Howard, P., Madasu, A., Le, T., Moreno, G.L., Bhiwandiwalla, A., Lal, V.: Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11975–11985 (Jun 2024), https://openaccess.thecvf.com/content/CVPR2024/html/Howard_SocialCounterfactuals_Probing_and_Mitigating_Intersectional_Social_Biases_in_Vision-Language_Models_CVPR_2024_paper.html
30. Hu, X., Lin, Y., Wang, S., Wu, Z., Lv, K.: Agent-centric relation graph for object visual navigation. IEEE Trans. Circuits Syst. Video Technol. **34**(2), 1295–1309 (2024). https://doi.org/10.1109/TCSVT.2023.3291131, https://doi.org/10.1109/TCSVT.2023.3291131
31. Jiang, X., Zheng, J., Liu, R., Li, J., Zhang, J., Matthiesen, S., Stiefelhagen, R.: @ bench: Benchmarking vision-language models for human-centered assistive technology. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3934–3943. IEEE (2025)
32. Karamolegkou, A., Nikandrou, M., Pantazopoulos, G., Villegas, D.S., Rust, P., Dhar, R., Hershcovich, D., Søgaard, A.: Evaluating multimodal language models as visual assistants for visually impaired users. arXiv preprint arXiv:2503.22610 (2025)
33. Khan, M.W., Sheng, H., Zhang, H., Du, H., Wang, S., Coroneo, M.T., Hajati, F., Shariflou, S., Kalloniatis, M., Phu, J., Agar, A., Huang, Z., Golzan, S.M., Yu, X.: RVD: A handheld device-based fundus video dataset for retinal vessel segmentation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/3a71ee306d6991f2f87dd414e0bdf851-Abstract-Datasets_and_Benchmarks.html
34. Li, J., Du, H., Tang, Y., You, J., Du, S., Yu, W., Li, W.: Lgfnet: Local correlation and global context fusion for multivariate time series forecasting. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2231–2235. IEEE (2026)
35. Li, J., Du, H., Yu, W., Tang, Y., You, J., Du, S., Li, W.: Cadmamba: Clustering adaptive mamba for multivariate time series forecasting. In: ICASSP 2026-

2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3066–3070. IEEE (2026)
36. Li, J., Yu, W., Du, H., Tang, Y., You, J., Du, S., Li, W.: Pamnet: Patch-adaptive mixing network for multivariate time series forecasting. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5836–5840. IEEE (2026)
37. Luo, B., Guo, J., Yao, Y., Ding, X.: Adversarial style optimization: Enhancing vlm jailbreaks by grpo-based stylistic triggers optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11–19 (June 2026)
38. Lv, K., Li, Y., Chen, Z., Wang, S., Han, S., Lin, Y.: Infer the whole from a glimpse of a part: Keypoint-based knowledge graph for vehicle re-identification. In: Walsh, T., Shah, J., Kolter, Z. (eds.) Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025. pp. 5901–5909. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I6.32630, https://doi.org/10.1609/aaai.v39i6.32630
39. Melistas, T., Spyrou, N., Gkouti, N., Sanchez, P., Vlontzos, A., Panagakis, Y., Papanastasiou, G., Tsaftaris, S.A.: Benchmarking counterfactual image generation. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024), https://openreview.net/forum?id=0T8xRFrScB
40. Mistral AI: Introducing mistral 3. Blog post (2025), accessed 2025-12
41. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. arXiv preprint arXiv:1810.03993 (2018)
42. OpenAI: Introducing gpt-5. https://openai.com/index/introducing-gpt-5/ (Aug 2025), accessed: 2026-03-05
43. Pan, Y., Bareinboim, E.: Counterfactual image editing. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=OXzkw7vFIO
44. Panda, S., Agarwal, A., Patel, H.L.: AccessEval: Benchmarking disability bias in large language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 32504–32530. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1653
45. Panda, S., Yadav, S.S., Malviya, P.: Auditing disability representation in vision-language models. CoRR **abs/2601.17348** (2026). https://doi.org/10.48550/ARXIV.2601.17348, https://doi.org/10.48550/arXiv.2601.17348
46. Ruggeri, G., Nozza, D., et al.: A multi-dimensional study on bias in vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics (2023)
47. Sun, X., Gazagnadou, N., Sharma, V., Lyu, L., Li, H., Zheng, L.: Privacy assessment on reconstructed images: Are existing evaluation metrics faithful to human perception? In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/2082273791021571c410f41d565d0b45-Abstract-Conference.html

48. Sun, X., Hou, Y., Deng, W., Li, H., Zheng, L.: Ranking models in unlabeled new environments. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 11741–11751. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01155, https://doi.org/10.1109/ICCV48922.2021.01155
49. Sun, X., Leng, X., Wang, Z., Yang, Y., Huang, Z., Zheng, L.: Cifar-10-warehouse: Broad and more realistic testbeds in model generalization analysis. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=pw2ssoOTpo
50. Sun, X., Li, M., Yuan, K., Sun, M.W., Endo, M., Wu, S., Li, C., Zhang, Y., Wang, Z., Yeung-Levy, S.: Do vlms perceive or recall? probing visual perception vs. memory with classic visual illusions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25861–25870 (June 2026)
51. Sun, X., Yao, Y., Wang, S., Li, H., Zheng, L.: Alice benchmarks: Connecting real world object re-identification with the synthetic. CoRR **abs/2310.04416** (2023). https://doi.org/10.48550/ARXIV.2310.04416, https://doi.org/10.48550/arXiv.2310.04416
52. Tang, Y., Du, H., Du, S., Li, W.: Integrating deep learning into semiparametric network vector autoregressive models. Mathematics **14**(1), 38 (2025)
53. United Nations: Convention on the rights of persons with disabilities. OHCHR instrument page (2006), https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities, accessed 2026-01-05
54. Vo, A., Nguyen, K.N., Taesiri, M.R., Dang, V.T., Nguyen, A.T., Kim, D.: Vision language models are biased. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=DG4S2OlGQA
55. Wang, J., Lin, Y., Hu, X., Wang, S., Lv, K.: Local motion matters: A deconstruct-recompose paradigm for reinforcement learning pre-training from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9859–9868 (June 2026)
56. Wang, S., Wu, Z., Hu, X., Lin, Y., Lv, K.: Skill-based hierarchical reinforcement learning for target visual navigation. IEEE Trans. Multim. **25**, 8920–8932 (2023). https://doi.org/10.1109/TMM.2023.3243618, https://doi.org/10.1109/TMM.2023.3243618
57. Wang, S., Cao, X., Zhang, J., Yuan, Z., Shan, S., Chen, X., Gao, W.: Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model (2024). https://doi.org/10.48550/arXiv.2406.14194
58. World Health Organization: Disability and health. Fact sheet (2023), https://www.who.int/news-room/fact-sheets/detail/disability-and-health, accessed 2026-01-05
59. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
60. Wu, H., Song, S., Yao, C., Han, S., Wan, H., Lin, Y., Lv, K.: Think how your teammates think: Active inference can benefit decentralized execution. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence,

AAAI 2026, Singapore, January 20-27, 2026. pp. 29749–29757. AAAI Press (2026). https://doi.org/10.1609/AAAI.V40I35.40220, https://doi.org/10.1609/aaai.v40i35.40220

61. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., Ruan, C.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. CoRR **abs/2412.10302** (2024). https://doi.org/10.48550/ARXIV.2412.10302, https://doi.org/10.48550/arXiv.2412.10302
62. Yao, C., Lin, Y., Song, S., Wu, H., Yang, S., Ma, Y., Lv, K.: Role-level inductive bias for cross-task generalization in multi-agent reinforcement learning. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=oz8kFbXdpj
63. Yao, Y., Gedeon, T., Zheng, L.: Large-scale training data search for object re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15568–15578. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01494, https://doi.org/10.1109/CVPR52729.2023.01494
64. Yao, Y., Wen, Z., Tong, Y., Tian, X., Li, X., Ma, X., Xu, D., Gedeon, T.: Thought graph traversal for test-time scaling in chest x-ray vllms. Pattern Recognit. **179**, 113639 (2026). https://doi.org/10.1016/J.PATCOG.2026.113639, https://doi.org/10.1016/j.patcog.2026.113639
65. Yao, Y., Yang, R., Gedeon, T.: Bipartite mode matching for vision training set search from a hierarchical data server. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026. pp. 16128–16136. AAAI Press (2026). https://doi.org/10.1609/AAAI.V40I19.38648, https://doi.org/10.1609/aaai.v40i19.38648
66. Yao, Y., Zheng, L., Yang, X., Naphade, M., Gedeon, T.: Simulating content consistent vehicle datasets with attribute descent. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 12351, pp. 775–791. Springer (2020). https://doi.org/10.1007/978-3-030-58539-6_46, https://doi.org/10.1007/978-3-030-58539-6_46
67. Yao, Y., Zheng, L., Yang, X., Napthade, M., Gedeon, T.: Attribute descent: Simulating object-centric datasets on the content level and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **46**(4), 2489–2505 (2024). https://doi.org/10.1109/TPAMI.2023.3338291, https://doi.org/10.1109/TPAMI.2023.3338291
68. Yin, X., Yang, B., Liu, W., Xue, Q., Alamri, A., Fiedler, G., Gao, W.: Progait: A multi-purpose video dataset and benchmark for transfemoral prosthesis users. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8984–8993 (2025)
69. Ying, J., Du, H., Zhang, K., Li, L., Yu, X.: Ldpose: Towards inclusive human pose estimation for limb-deficient individuals in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9865–9875 (2025)
70. Ying, J., Du, H., Zhang, K., Tweedy, S.M., Yu, X.: Resihmr: Residual-limb aware single-image 3d human mesh recovery for individuals with limb loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13940–13950 (June 2026)

71. Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14830–14840 (2021)