INCLUSIVEVLM-LIMBEVIDENCE: EVALUATING VISION-LANGUAGE MODELS ON IMAGES OF INDIVIDUALS WITH LIMB DEFICIENCIES

Anonymous authors
Paper under double-blind review



Figure 1: **Illustration of INCLUSIVEVLM-LIMBEVIDENCE.** Open-world images of individuals with limb deficiencies support three evaluation requirements. LIMBEVIDENCE-PRESENCE tests evidence detection. LIMBEVIDENCE-ATTRIBUTION assigns it to body regions and sides through parallel Constrained and FreeForm formats. LIMBEVIDENCE-COMPATIBILITY tests category relations and complete option sets through parallel Category and Diversity probes. The arrows summarize increasingly specific requirements across separately evaluated tasks.

ABSTRACT

Strong performance on general benchmarks does not establish whether vision-language models (VLMs) correctly interpret images of individuals with limb deficiencies. For this population, applications such as preparing visual evidence for Paralympic athlete classification depend on correctly identifying residual limbs, prosthetic devices, and body-device configurations. We find that VLMs can miss this evidence or assign it to the wrong body regions, so later reasoning begins from incorrect visual information. To examine these failures, we introduce INCLUSIVEVLM-LIMBEVIDENCE, a test-only benchmark on about 17k open-world images of this population. It provides over 52k direct evidence queries and over 20k compatibility instances, with human reference results and prompt paraphrases. The design separates recognizing evidence from using it in structured decisions. LIMBEVIDENCE-PRESENCE tests evidence detection. Because detection alone does not establish where evidence belongs, LIMBEVIDENCE-ATTRIBUTION tests body-region assignment in parallel Constrained and FreeForm formats. To examine whether models can also use this evidence in structured decisions, LIMBEVIDENCE-COMPATIBILITY tests option matching under benchmark-defined relations through parallel Category and Diversity probes. Across widely used general-purpose VLMs, models often detect limb evidence but fail to identify its body region or side. Partial option matches often fall short of complete compatible-set recovery. Models can also repeat the same incorrect answer across paraphrases. The larger model in our scale comparison improves coarse detection

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but does not resolve attribution and set-recovery errors. These results reveal a gap between recognizing limb evidence and using its spatial and relational details correctly. They also show that response consistency cannot replace correctness. INCLUSIVEVLM-LIMBEVIDENCE makes these overlooked capability limits measurable and supports research toward visual AI that better serves individuals with limb deficiencies.

1 INTRODUCTION

Strong performance on general benchmarks does not establish whether vision-language models (VLMs) correctly interpret images of individuals with limb deficiencies. General-purpose benchmarks cover a broad range of visual recognition and reasoning tasks (Liu et al., 2024b; Yu et al., 2023). They rarely assess whether models correctly interpret residual limbs, prosthetic devices, and their locations on the body. LDPose (Ying et al., 2025) and ProGait (Yin et al., 2025) address this population in pose estimation and gait analysis. These resources support inclusive perception research, but they do not directly evaluate how VLMs use limb evidence in language-based outputs.

For this population, assistive technologies can support participation and everyday access to information (World Health Organization, 2022). Visual AI applications also require accurate recognition of residual limbs, prosthetic devices, and body-device configurations. Our work began with VLM prototypes for preparing visual evidence in Paralympic athlete classification workflows. These prototypes missed visible limb evidence or assigned it to the wrong body regions. Later reasoning therefore began from incorrect visual information.

To examine these failures, we introduce INCLUSIVEVLM-LIMBEVIDENCE, a test-only benchmark on 16,850 open-world images of individuals with limb deficiencies. It contains 52,812 direct evidence queries and 20,664 compatibility instances. We reuse images from LDPose and add new masks for residual limbs and prosthetic devices, body-region labels, and body-device configuration records. These annotations provide the targets for the evaluation tasks. The benchmark includes structured answer spaces, prompt paraphrases, and human reference results for reproducible auditing.

As shown in Figure 1, the design separates recognizing evidence from using it in structured decisions. The basic requirement is to detect which types of limb evidence are visible, which motivates LIMBEVIDENCE-PRESENCE. Detection alone does not establish where the evidence belongs on the body. LIMBEVIDENCE-ATTRIBUTION therefore tests assignment to the correct body region and side. Constrained and FreeForm are parallel formats for the same attribution target, so their comparison examines the role of output format. Identifying evidence also does not establish whether a model can apply body-device relations when choosing among candidate options. LIMBEVIDENCE-COMPATIBILITY tests this additional requirement under benchmark-defined relations. Its parallel Category and Diversity probes distinguish coarse relation recovery from complete-set recovery across visually varied options and distractors.

Across the evaluated general-purpose VLMs, we observe three recurring failure modes. Models can detect an evidence type yet assign it to the wrong body region or side. They can select matching options without recovering the complete target set. They can also repeat an incorrect answer across equivalent prompts. Our comparison of the 30B and 235B Qwen3-VL models (Bai et al., 2025) shows that better coarse detection does not resolve attribution and set-recovery errors. In the 235B model's attribution responses, evidence type is correct on 74.40% of prompts, compared with 40.11% for body region and 40.59% for side.

These findings reveal a gap between recognizing limb evidence and using its spatial and relational details correctly. They also show why response consistency must be considered together with correctness. Even large general-purpose VLMs still make errors on visual details that matter for this population. By making these capability limits measurable, INCLUSIVEVLM-LIMBEVIDENCE supports targeted research on individuals with limb deficiencies. Its broader goal is to help advances in visual AI benefit a population whose images remain underexamined in mainstream evaluation. Further details are provided on the anonymous project page. Our contributions are as follows:

- We introduce **INCLUSIVEVLM-LIMBEVIDENCE**, a test-only VLM resource with new limb-evidence annotations and structured evaluation targets. It brings individuals with limb deficiencies into VLM evaluation.
- We separate evidence detection, body-region attribution, and compatibility-style structured decisions through shared annotations and parallel probes. This design makes accessibility-relevant visual errors measurable for this population.
- We identify attribution errors, incomplete matching sets, and stable but incorrect answers despite improved coarse detection. These findings motivate targeted research toward visual AI that serves this population.

2 RELATED WORK

2.1 INCLUSIVE AND ACCESSIBILITY-FACING VISION

Inclusive vision examines whether visual systems represent the people they serve. LDPose represents residual-limb endpoints for pose estimation (Ying et al., 2025). ProGait supports video analysis of individuals with transfemoral prostheses (Yin et al., 2025), and AJAHR adapts mesh recovery to absent joints (Cho et al., 2025). The position paper on topological generalization argues for human-centric vision beyond fixed skeletal topologies (Du et al., 2026b). These works address body configurations that conventional representations often overlook.

VizWiz evaluates visual questions from blind users (Gurari et al., 2018). Du et al. (2026a) use counterfactual depictions to evaluate VLMs on individuals with limb deficiencies. INCLUSIVEVLM-LIMBEVIDENCE complements these directions by separating evidence detection, body-region attribution, and structured compatibility decisions in open-world images. Its focus is whether answers reflect visible limb evidence as the evaluation requirements become more specific. A population-centered benchmark can reveal errors that task diversity alone does not expose. The distinction is between representing body geometry, examining depiction-dependent responses, and testing the evidence needed for a particular answer.

2.2 VISION-LANGUAGE GROUNDING AND STRUCTURED EVALUATION

Visual question answering tests image-conditioned answers, while VQAv2 exposes reliance on language priors (Antol et al., 2015; Goyal et al., 2017). GQA evaluates compositional reasoning (Hudson & Manning, 2019). Referring-expression and region-language datasets connect words with image regions (Kazemzadeh et al., 2014; Yu et al., 2016; Plummer et al., 2015; Krishna et al., 2016). These resources motivate checking evidence alignment beyond answer fluency. Our attribution targets link evidence types to named body regions, connecting language outputs with the annotated image content.

MMBench, MMMU, and MME evaluate broad multimodal capabilities (Liu et al., 2024a; Yue et al., 2024; Fu et al., 2023). MathVista and ScienceQA examine mathematical and scientific reasoning (Lu et al., 2024; 2022). MM-Vet evaluates integrated capabilities through open-ended responses and model-based scoring (Yu et al., 2024). INCLUSIVEVLM-LIMBEVIDENCE uses fixed targets to examine limb evidence within one population. Its parallel attribution formats probe output-format effects, while category and set matching distinguish coarse relations from complete recovery. This separates the content a model recognizes from the demands of expressing and using it.

2.3 RELIABILITY AND CONSISTENCY

HallusionBench examines language hallucinations and visual illusions (Guan et al., 2024). Calibration and uncertainty estimation distinguish confidence from correctness (Guo et al., 2017; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Other work studies unfamiliar inputs (Hendrycks & Gimpel, 2017; Liang et al., 2018), selective prediction (Geifman & El-Yaniv, 2019), and prediction sets with coverage guarantees (Angelopoulos & Bates, 2021). These directions motivate examining behavior beyond average accuracy. INCLUSIVEVLM-LIMBEVIDENCE reports paraphrase consistency alongside correctness because repeated answers can remain wrong. The resulting diagnostics separate missed evidence, incorrect attribution, and incomplete option sets for an overlooked population.

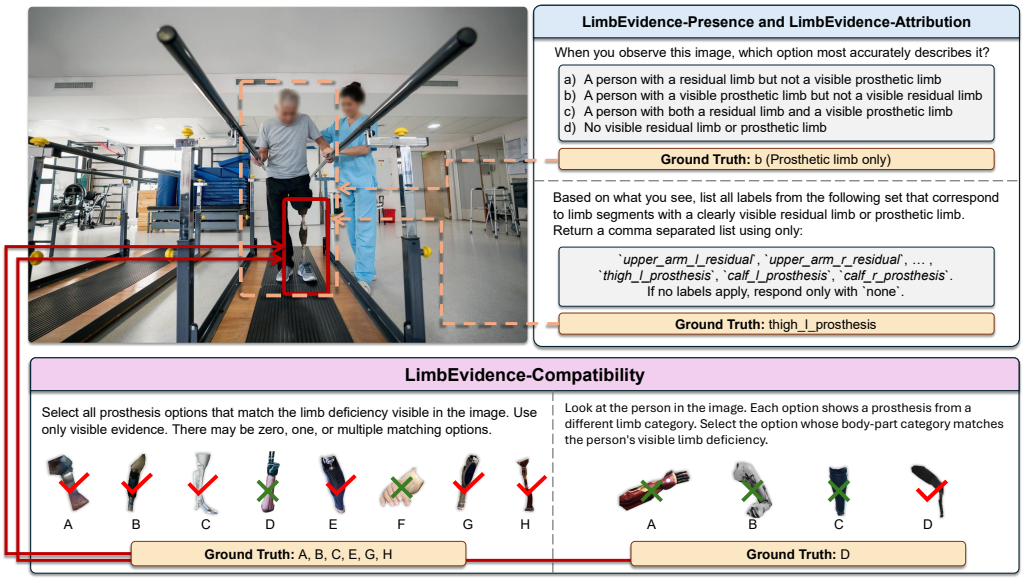


Figure 2: **Demonstration of evaluation tasks in INCLUSIVEVLM-LIMBEVIDENCE.** For the highlighted individual, LIMBEVIDENCE-PRESENCE identifies prosthetic-device evidence, while LIMBEVIDENCE-ATTRIBUTION specifies `thigh_l_prosthesis`. Below, the parallel probes of LIMBEVIDENCE-COMPATIBILITY use separate option grids. Diversity requires the complete set A, B, C, E, G, and H, while Category matches D. Checks mark target options and crosses mark distractors, not model predictions.

3 INCLUSIVEVLM-LIMBEVIDENCE: DATA, ANNOTATIONS, AND EVALUATION PROBES

3.1 BENCHMARK OVERVIEW

INCLUSIVEVLM-LIMBEVIDENCE evaluates how VLMs recognize and use visible limb evidence in images of individuals with limb deficiencies.[1] Its design follows errors observed in our VLM prototypes for preparing visual evidence in Paralympic athlete classification. These errors motivated separate tests for detecting evidence, assigning it to a body region, and using it in structured choices. We defined the evaluation targets before creating the benchmark-specific annotations.

LIMBEVIDENCE-PRESENCE and LIMBEVIDENCE-ATTRIBUTION assess directly visible evidence. LIMBEVIDENCE-COMPATIBILITY adds matching under benchmark-defined body-device relations. Constrained and FreeForm are parallel Attribution formats, while Category and Diversity are parallel Compatibility probes. The requirements share an annotation framework and are evaluated separately. Figure 2 illustrates example prompts and targets. Appendix B.1 provides the full specification of the evaluation units, answer spaces, scales, and metrics.

3.2 DATA SOURCE, SCALE, AND RELEASE

INCLUSIVEVLM-LIMBEVIDENCE contains 16,850 images from the LDPose source pool (Ying et al., 2025). Reusing this population-centered collection allows us to focus on limb-evidence annotations and VLM evaluation. LIMBEVIDENCE-PRESENCE and LIMBEVIDENCE-ATTRIBUTION use 8,802 images and support 52,812 structured queries through prompt paraphrasing. LIMBEVIDENCE-COMPATIBILITY uses 8,048 images and contains 20,664 instances, including 8,048 Category instances and 12,616 Diversity instances. Its candidate prosthesis options are extracted from 7,497 source images with visible prosthetic devices. A source image can provide multiple option crops.

The release includes annotations, prompts, derived metadata, source references, evaluation scripts, and machine-readable records. Annotations, prompts, and derived benchmark artifacts are released under

[1]This project has approval from an institutional Human Research Ethics Committee through November 2029.

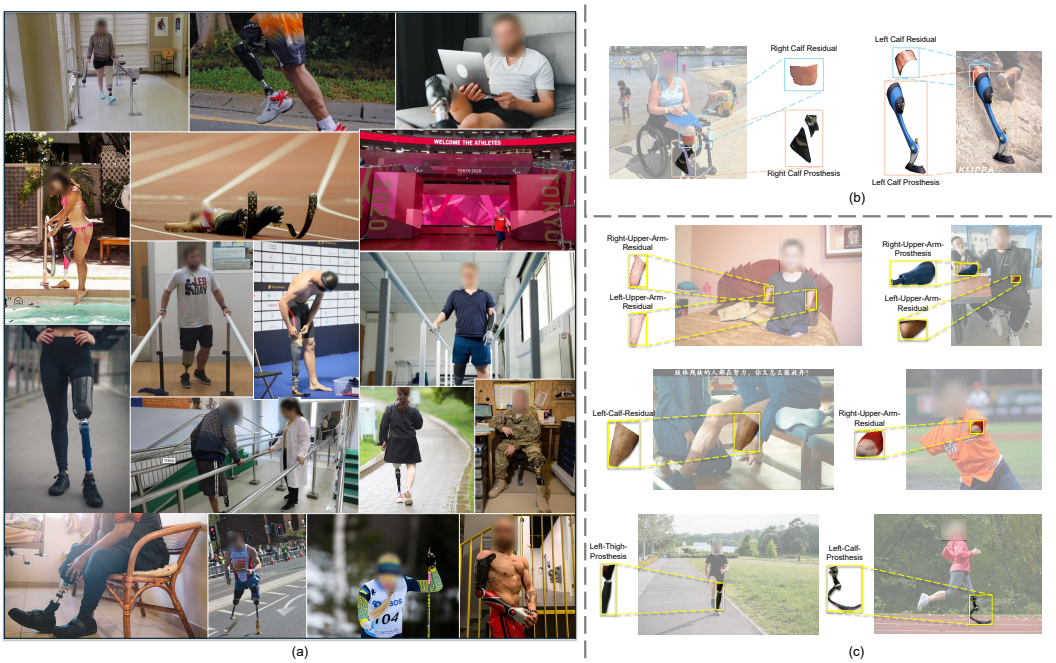


Figure 3: **Illustration of open-world images and limb-evidence annotations.** **(a)** Images of individuals with limb deficiencies cover everyday activities, rehabilitation, and sport. **(b)** Separate masks distinguish visible residual-limb and prosthetic-device regions, which can share a limb segment. **(c)** Labeled examples link evidence types to body regions and sides. These annotations support the evaluation targets in INCLUSIVEVLM-LIMBEVIDENCE.

CC BY-NC-SA 4.0.[2] Source images remain subject to LDPose access terms and are not redistributed as re-encoded files. Access follows a controlled request process and a Data Use Agreement for non-commercial research. The agreement prohibits re-identification and face recognition and supports takedown or correction requests. Appendix E provides the release and governance details.

3.3 LIMB-EVIDENCE ANNOTATIONS

We create new pixel-level masks for visible residual-limb and prosthetic-device regions, together with body-region labels and body-device configuration records. These annotations are new contributions of INCLUSIVEVLM-LIMBEVIDENCE rather than reused LDPose annotations. Figure 3 shows the image contexts, separated evidence regions, and labeled examples.

For direct evidence evaluation, masks are aggregated into eight limb segments: left and right upper arm, forearm, thigh, and calf. Crossing these segments with the two evidence types yields 16 segment-type labels. Residual-limb and prosthetic-device masks are disjoint at the pixel level, but both types can occur within one segment. The resulting labels support image-level recognition and body-region attribution without coordinate prediction. For example, the prosthetic-device annotation in Figure 2 gives a Presence target of prosthetic-device evidence only and an Attribution target of `thigh_l_prosthesis`. Only visible evidence is annotated, and intact limbs are outside the annotation scope.

A second annotator reviews each mask and derived label set, including the accepted categories and matching option sets. Cases with disagreement above the annotation tolerance undergo re-annotation and adjudication. An International Paralympic Committee certified classifier supports annotator guidance and case adjudication. Periodic spot audits check randomly sampled batches. Appendix B.3 and Appendix B.4 describe the annotation scope, guidelines, and review procedure.

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

3.4 EVALUATION PROBES

LIMBEVIDENCE-PRESENCE. Detection is the first requirement for using limb evidence. The model selects one of four image-level labels: residual-limb evidence only, prosthetic-device evidence only, both evidence types visible, or no visible limb-related evidence. The target describes what is visible in the image.

LIMBEVIDENCE-ATTRIBUTION. Detecting an evidence type does not establish where it belongs on the body. Constrained therefore requires all applicable segment-type labels from the 16-label space, with `none` representing the empty set. FreeForm addresses the same target through text responses that a deterministic parser normalizes to canonical labels. It uses 500 image-target pairs randomly sampled from the Constrained pool, with the image and target unchanged. This paired comparison examines output-format effects on attribution.

LIMBEVIDENCE-COMPATIBILITY. Identifying evidence does not establish whether a model can use body-device relations to select matching options. Matching uses a fixed correspondence between the body-part categories annotated for the image and the device options. Applying this relation requires identifying the relevant body part in both the image and the candidates. Category tests coarse body-part matching under the benchmark relation. Diversity tests complete-set recovery across visually varied prosthetic-device options and distractors. The probes are parallel, and Diversity refers to option appearance rather than demographic composition. In Figure 2, the Category target is D, while the separate Diversity grid requires A, B, C, E, G, and H. Exact recovery requires every target option and no distractors. Applicable grids include an insufficient-information option when the visible evidence does not support a decision. Appendix B.6 specifies the probes and option construction, and Appendix D.4 discusses their scope.

3.5 TEST-ONLY EVALUATION AND REPRODUCIBILITY

INCLUSIVEVLM-LIMBEVIDENCE is test-only and provides no benchmark-specific training split. Structured answer spaces and deterministic normalization support scoring against fixed targets. A paraphrase group contains prompts for the same image and target that differ only in wording. Consistency across these prompts measures response variation, not correctness. We interpret raw and correctness-gated consistency alongside correctness and retained-group coverage. The gated score retains groups with at least one correct response.

Each instance stores stable identifiers, prompt text, the answer space, target labels, and derived metadata. Direct evidence records include paraphrase-group identifiers. Compatibility records include option-grid identifiers and insufficient-information choices where applicable. These records support reconstruction of the evaluated inputs and aggregation across model runs.

4 EXPERIMENTAL EVALUATION

We evaluate how current VLMs detect limb evidence, assign it to body regions, and use it in structured choices in INCLUSIVEVLM-LIMBEVIDENCE. The evaluation includes GPT-5 (OpenAI, 2025), Gemini-2.5 (Comanici et al., 2025), Qwen3-VL (Bai et al., 2025), Gemma-3 (Gemma Team, 2025), DeepSeek-VL2 (Wu et al., 2024), and Ministral-3 (Mistral AI, 2025). We use structured decoding or deterministic normalization to score responses against fixed targets. Scores are percentages with available bootstrap 95% confidence intervals. For Qwen3-VL-235B-A22B-Instruct, the four reported primary intervals use 10,000 resamples of original images. Appendix C provides the model setup, scoring details, and additional comparisons.

Human reference. The human study provides reference performance across rater backgrounds using the reviewed benchmark targets. Five raters each completed 200 items sampled randomly within fixed per-probe quotas. Three had primarily computer-science backgrounds and two had extensive disability-research experience. The human reference reaches 98.0% Presence accuracy and 83.4% Constrained Attribution set mIoU. For Diversity, the disability-research group reaches 74.22% F1, compared with 47.03% for the computer-science group. Reporting the groups separately shows a

Table 1: **Results for direct limb-evidence recognition.** LIMBEVIDENCE-PRESENCE reports accuracy and macro-F1. LIMBEVIDENCE-ATTRIBUTION reports Constrained set mIoU and F1, and FreeForm set mIoU on 500 paired items. Consistency is correctness-gated. Scores are percentages with available 95% interval half-widths. Bold marks the highest model accuracy, overlap, or F1.

Family	Model	LIMBEVIDENCE-PRESENCE			LIMBEVIDENCE-ATTRIBUTION				
					Constrained				FreeForm
		Accuracy	**Macro F1**	Gated consistency	**Set mIoU**	**Micro F1**	**Macro F1**	Gated consistency	**Set mIoU**
Qwen3-VL	4B-Inst	41.9±0.6	30.1±0.5	60.7±0.7	17.0±0.3	25.4±0.4	18.4±0.4	45.8±3.0	37.7±2.2
	8B-Inst	41.4±0.6	32.0±0.4	61.7±0.7	18.8±0.4	28.2±0.4	21.5±0.5	51.3±2.5	35.2±2.1
	30B-A3B-Inst	23.9±0.5	21.8±0.3	84.3±1.0	24.0±0.5	32.4±0.5	21.3±0.6	68.3±1.4	30.2±1.9
	235B-A22B-Inst	74.2±0.8	43.8±0.3	91.6±0.8	24.8±0.7	34.2±0.3	23.6±0.4	68.5±4.5	**45.7±1.7**
	4B-Think	40.6±0.6	30.1±0.5	66.7±0.7	17.8±0.3	27.2±0.4	19.3±0.5	55.5±3.7	43.1±1.9
	8B-Think	35.3±0.6	28.8±0.4	66.0±0.7	16.2±0.3	25.3±0.4	20.4±0.5	47.6±2.5	44.7±2.1
	30B-A3B-Think	33.6±0.6	29.2±0.4	70.3±0.8	19.4±0.4	24.1±0.3	18.7±0.3	52.5±2.2	38.8±2.2
Gemma-3	4B	33.0±0.6	25.6±0.4	62.8±1.0	20.5±0.4	26.1±0.4	19.4±0.3	52.4±2.8	26.5±1.7
	12B	29.9±0.6	24.6±0.3	85.2±0.9	24.2±0.4	33.0±0.5	25.3±0.6	54.9±1.7	20.4±1.3
	27B	38.6±0.6	30.2±0.4	68.1±0.8	27.8±0.4	37.2±0.5	24.4±0.7	69.4±1.7	24.4±1.6
DeepSeek-VL2	Tiny	19.4±3.5	9.8±1.4	66.3±7.1	6.5±0.1	11.1±0.2	9.1±0.2	65.4±0.7	18.7±1.8
	Small	27.7±3.9	15.4±1.9	91.0±5.5	9.4±0.4	17.3±0.7	15.8±0.8	80.2±3.9	12.3±1.5
	27B-MoE	34.0±7.7	12.7±2.1	100.0±0.0	3.4±1.3	19.2±4.2	16.1±4.7	50.0±11.1	27.4±1.3
Ministral-3	3B-Inst	51.0±0.7	31.9±0.5	66.1±0.7	10.2±0.3	21.5±0.4	12.0±0.4	23.1±3.1	0.0±0.0
	8B-Inst	37.1±0.6	25.8±0.4	61.0±0.7	1.4±0.1	5.0±0.3	3.6±0.3	44.0±6.0	37.8±2.1
	14B-Inst	29.3±0.6	22.2±0.3	78.3±0.9	10.8±0.3	21.5±0.5	10.5±0.5	33.3±1.9	37.6±2.1
	3B-Reason	22.4±0.5	17.4±0.4	64.0±1.0	10.4±0.3	22.4±0.4	16.1±0.4	27.5±7.1	16.2±1.6
	8B-Reason	24.8±0.6	20.3±0.4	58.4±0.9	2.9±0.2	9.5±0.4	6.8±0.4	46.1±6.0	36.2±2.0
	14B-Reason	25.0±0.5	19.8±0.3	76.3±1.0	2.3±0.1	8.5±0.4	5.3±0.4	75.0±12.5	28.9±1.9
OpenAI	GPT-5	**77.2±1.6**	**50.0±1.4**	86.9±1.5	**29.4±1.5**	**38.5±1.6**	**32.8±3.3**	85.8±2.9	30.5±2.0
Gemini	Gemini-2.5	47.0±0.7	35.3±0.4	73.1±0.8	24.8±0.4	35.9±0.5	29.0±0.7	55.3±1.8	22.8±1.2
Human reference									
Human	5 raters	98.0±3.0	69.6±12.9	–	83.4±9.1	85.0±7.5	73.1±18.0	–	–

difference that the mixed-background average hides. Appendix C.3 gives the assignments, rater-level scores, and aggregation details.

4.1 EVIDENCE DETECTION DOES NOT ENSURE CORRECT ATTRIBUTION

Table 1 reports LIMBEVIDENCE-PRESENCE and LIMBEVIDENCE-ATTRIBUTION. GPT-5 reaches the highest model point estimates of 77.2% Presence accuracy and 29.4% Constrained set mIoU. Both remain below the corresponding human reference results. For the 235B model, recall is 87.57% for prosthetic-device evidence only but 25.07% when both evidence types are visible. Its overall Presence accuracy therefore hides a marked difference across evidence categories.

A component analysis examines the distinction between evidence type and spatial assignment within the same attribution responses. For Qwen3-VL-235B-A22B-Instruct, evidence type is correct on 74.40% of prompts. Body region and side are correct on only 40.11% and 40.59%, and the complete target is correct on 17.10%. Thus, coarse evidence-type recognition coexists with errors about where that evidence belongs. Appendix C.6 reports the component analysis and oracle corrections.

The scale comparison shows a related pattern. Moving from Qwen3-VL-30B-A3B-Instruct to 235B-A22B-Instruct raises Presence accuracy from 23.9% to 74.23%. Constrained set mIoU changes from 24.0% to 24.76%. The gain in detection is therefore not accompanied by a comparable gain in attribution.

Output-format effects. The paired 500-item comparison holds the images and attribution targets fixed. FreeForm set mIoU rises from 17.22% to 43.1% for Qwen3-VL-4B-Thinking and from 15.87% to 44.7% for 8B-Thinking. GPT-5 changes from 26.51% to 30.5%. Output format therefore affects measured attribution, but neither format resolves the remaining errors. Appendix Table 4(b) also reports parsing coverage and invalid outputs for these models.

Table 2: **Results for LIMBEVIDENCE-COMPATIBILITY.** Category credits one selection from the accepted target set. Diversity Exact requires the whole set, while Partial requires a target hit. Precision, recall, and F1 are sample averages. Consistency is correctness-gated. Scores are percentages with available 95% interval half-widths. Bold marks the highest model accuracy, Exact, or F1.

Family	Model	Category	Diversity					
		Accuracy	Exact	Partial	Precision	Recall	F1	Gated consistency
Qwen3-VL	4B-Inst	37.2±1.1	2.0±0.3	66.0±0.9	51.9±0.8	29.1±0.6	31.9±0.5	14.3±4.9
	8B-Inst	36.6±1.1	1.9±0.3	64.1±0.9	52.3±0.8	26.0±0.6	30.0±0.5	17.2±5.4
	30B-A3B-Inst	35.4±1.1	7.6±0.5	80.4±0.7	48.8±0.6	69.7±0.7	51.2±0.6	70.1±3.5
	235B-A22B-Inst	**39.7±1.1**	3.1±0.4	57.1±0.8	49.8±0.7	21.4±0.8	25.9±0.5	30.8±3.4
	4B-Think	25.4±0.5	7.5±0.6	78.3±0.9	48.1±0.9	67.2±0.9	49.3±0.7	65.8±4.2
	8B-Think	24.5±0.5	7.7±0.5	75.4±0.7	47.1±0.6	61.8±0.8	47.0±0.6	57.9±3.8
	30B-A3B-Think	22.6±0.4	2.7±0.4	62.9±1.0	50.4±0.9	29.0±0.8	29.7±0.7	19.5±5.3
Gemma-3	4B	16.3±0.9	0.7±0.1	57.2±0.8	47.7±0.8	16.9±0.3	22.9±0.4	25.7±9.3
	12B	11.6±0.8	1.0±0.2	57.4±0.9	48.2±0.8	17.7±0.4	23.4±0.4	19.9±7.1
	27B	10.5±0.7	0.9±0.2	56.0±0.9	49.0±0.9	15.8±0.4	21.9±0.5	36.5±11.2
DeepSeek-VL2	Tiny	3.0±1.1	**10.2±1.9**	84.5±2.2	49.6±2.1	81.3±2.3	**57.5±2.1**	87.0±7.2
	Small	30.6±3.0	5.8±1.5	70.3±2.8	58.3±2.7	40.5±2.1	43.8±2.1	50.6±13.7
	27B-MoE	37.3±5.5	8.7±3.2	69.3±5.2	55.6±4.9	44.8±4.6	42.6±4.0	48.8±18.0
Ministral-3	3B-Inst	10.0±0.3	2.8±0.4	50.4±1.1	44.6±1.0	16.1±0.5	21.3±0.6	18.5±5.0
	8B-Inst	18.0±0.9	1.0±0.2	54.7±0.9	49.3±0.8	15.6±0.4	21.6±0.4	13.3±7.6
	14B-Inst	17.5±0.9	1.1±0.2	55.5±0.9	49.4±0.8	15.9±0.4	21.9±0.4	14.4±7.0
	3B-Reason	17.6±0.9	0.6±0.2	64.1±0.9	47.8±0.8	22.8±0.5	28.2±0.5	64.6±12.5
	8B-Reason	17.9±0.9	0.4±0.2	64.9±0.9	48.2±0.8	24.5±0.5	29.5±0.4	42.3±15.7
	14B-Reason	15.6±0.9	1.1±0.2	70.4±0.8	48.8±0.7	34.8±0.6	35.8±0.5	13.4±6.6
OpenAI	GPT-5	32.5±0.8	2.9±0.3	57.6±0.9	53.3±0.9	18.9±0.5	24.8±0.6	17.7±4.7
Gemini	Gemini-2.5	34.7±0.5	3.5±0.3	63.5±0.9	48.2±0.8	27.0±0.6	30.4±0.5	31.9±4.8
Human reference								
Human	5 raters	56.0±13.0	12.0±9.0	68.0±13.0	–	–	57.9±3.4	–

4.2 PARTIAL OPTION MATCHES DO NOT ENSURE COMPLETE SET RECOVERY

Table 2 reports the parallel probes of LIMBEVIDENCE-COMPATIBILITY. Category credits a single selected category when it belongs to the accepted target set. This accepted-set hit at one reaches 39.70% for Qwen3-VL-235B-A22B-Instruct, compared with the human reference of 56.0%. Category gains do not imply better set prediction. In the comparison with the 30B Instruct model, the 235B model improves Category accuracy from 35.4% to 39.70%. Diversity F1 falls from 51.2% to 25.90%, and exact-set accuracy falls from 7.6% to 3.07%.

For Diversity, the main pattern is the difference between partial target hits and exact recovery. DeepSeek-VL2-Tiny reaches 84.5% Partial but only 10.2% Exact. GPT-5 reaches 57.6% Partial and 2.9% Exact. Partial records at least one target hit and does not require the rest of the selection to be correct. It therefore cannot establish complete understanding of the matching relation. Exact, precision, recall, and F1 distinguish the remaining set errors.

The precision and recall profiles show different error balances. DeepSeek-VL2-Tiny has 49.6% precision and 81.3% recall, so high target coverage coexists with incorrect inclusions. GPT-5 has 53.3% precision and 18.9% recall, indicating many missed targets. These results distinguish partial retrieval from recovering every target without distractors. The 235B model also has low recall, at 21.37% under sample averaging. It selects 1.63 options on average against a target average of 3.83, and its predicted set is smaller on 68.0% of instances. At the same time, it selects an option on 1,674 of 1,851 empty-target instances, or 90.44%. Under-selection when matches exist therefore coexists with incorrect selection when no option matches. Appendix C.7 gives the complete outcome counts. The six inference-style comparisons and the matched MedGemma control appear in Appendix C.4.

Figure 4 gives concrete examples of the observed errors. The two Presence cases show unsupported residual-limb evidence and missed residual-limb evidence. The Category case shows incorrect body-part matching. The supplied prompts, targets, and predictions make these errors inspectable at the example level.

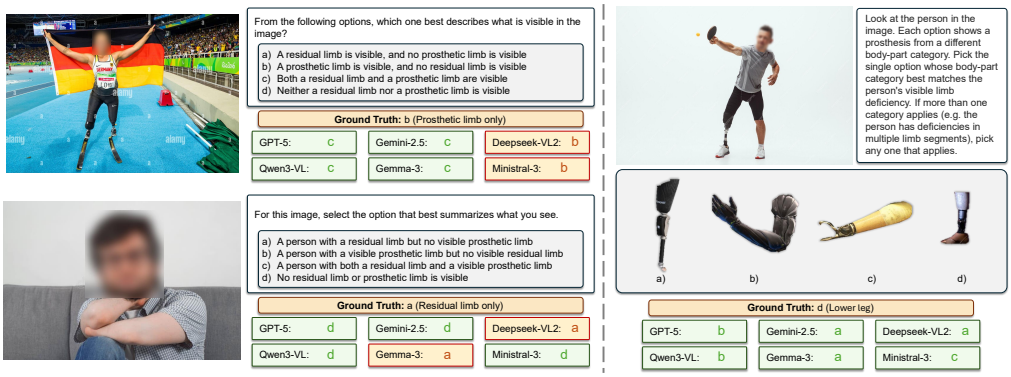


Figure 4: **Demonstration of model errors in INCLUSIVEVLM-LIMBEVIDENCE.** The two left examples show unsupported residual-limb evidence and missed residual-limb evidence in LIMBEVIDENCE-PRESENCE. The right example shows incorrect Category matching in LIMBEVIDENCE-COMPATIBILITY. Targets and recorded model selections are shown for each case. Red marks correct selections and green marks errors.

4.3 CONSISTENT ANSWERS CAN REMAIN INCORRECT

Paraphrase consistency measures response variation, not correctness. For the 235B model, Presence raw consistency is 92.22%. Gated consistency is 91.58% on 7,080 of 8,802 groups, with 80.44% coverage. On Constrained Attribution, Qwen3-VL-235B-A22B-Instruct has 17.10% exact correctness and 75.43% raw consistency. The model returns the same incorrect answer throughout 47.28% of paraphrase groups. Correctness-gated consistency is 68.51%, but the gate retains only 2,099 of 8,802 groups, or 23.85%. This score describes stability within the retained subset, rather than performance across all groups. Appendix C.8 reports the definitions and the complete available diagnostic summary.

The results distinguish detecting evidence, assigning it correctly, recovering complete option sets, and repeating an answer. Even large general-purpose models still make errors on these population-relevant visual details. INCLUSIVEVLM-LIMBEVIDENCE makes the errors measurable and supports more targeted evaluation of images of individuals with limb deficiencies.

5 DISCUSSION

Our results show why population coverage and task requirements both matter in VLM evaluation. Including images of individuals with limb deficiencies must be accompanied by tests of the visual details that matter for this population. A model can detect a prosthetic device yet assign it to the wrong limb, or select a matching option without recovering the complete set. In our scale comparison, improved coarse detection does not resolve these finer errors. Repeated answers can also remain incorrect. INCLUSIVEVLM-LIMBEVIDENCE makes these distinctions measurable through separate tests of evidence recognition and structured decisions. It provides a basis for examining whether model improvements extend to residual limbs, prosthetic devices, and their relation to the body. This coverage matters if individuals with limb deficiencies are to share in the benefits of visual AI. Appendix D provides the extended discussion.

6 CONCLUSION

We introduced INCLUSIVEVLM-LIMBEVIDENCE, a test-only benchmark for evaluating vision-language models on open-world images of individuals with limb deficiencies. It uses shared annotations to separate evidence detection, body-region attribution, and compatibility-style structured decisions. Our evaluation shows that even large general-purpose models still miss limb evidence, misassign its body region, and fail to recover complete matching sets. Consistent responses can also be incorrect. By making these errors measurable, INCLUSIVEVLM-LIMBEVIDENCE supports targeted research on accessibility-facing image understanding. Its broader aim is to help individuals with limb deficiencies benefit from progress in visual AI.

AI USE STATEMENT

We used ChatGPT for language polishing, manuscript organization, reference checking, and feedback on evaluation methodology. We did not use generative AI to originate the research idea or conduct autonomous research. Separately, the VLMs evaluated in this work are research subjects for examining capability limits on images of individuals with limb deficiencies. The authors review AI-assisted material against the source records and take responsibility for the final content.

ETHICS STATEMENT

This research received approval from an institutional Human Research Ethics Committee. Data use and future releases follow the approved protocol, source licenses, and privacy requirements. Source-image access is restricted to non-commercial research under an agreement that prohibits re-identification and face recognition. We support correction and removal requests. INCLUSIVEVLM-LIMBEVIDENCE aims to improve the evaluation and accurate representation of individuals with limb deficiencies. Appendix E documents the governance procedures.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe data provenance, annotation procedures, and benchmark construction. Appendix C documents inference settings, prompts, response parsing, evaluation metrics, human assessment, and confidence interval estimation. We will publicly release benchmark annotations, task records, evaluation code, and reconstruction instructions to support independent verification and reuse. Source images will remain subject to the documented access and licensing requirements.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. URL http://arxiv.org/abs/1505.00468.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hyunjin Cho, Giyun Choi, and Jongwon Choi. AJAHR: Amputated joint aware 3d human mesh recovery. *arXiv preprint arXiv:2509.19939*, 2025.

Matteo Cognolato, Manfredo Atzori, Roger Gassert, and Henning Müller. Improving robotic hand prosthesis control with eye tracking and computer vision: A multimodal approach based on the visuomotor behavior of grasping. *Frontiers in artificial intelligence*, 4:744476, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Creative Commons. Creative commons attribution-noncommercial-sharealike 4.0 international (cc by-nc-sa 4.0): Human-readable summary (deed). Creative Commons, 2013a. URL https://creativecommons.org/licenses/by-nc-sa/4.0/. Accessed: 2026-01-14.

Creative Commons. Creative commons attribution-noncommercial-sharealike 4.0 international public license (cc by-nc-sa 4.0): Legal code. Creative Commons, 2013b. URL https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode. Accessed: 2026-01-14.

Heming Du, Jiaying Ying, Xin Chen, Sen Wang, Xue Li, and Xin Yu. Same person, different depiction: Counterfactual evaluation of vision-language models on individuals with limb deficiencies. In *European Conference on Computer Vision*, pp. 685–705. Springer, 2026a.

Heming Du, Jiaying Ying, Kaihao Zhang, Tianqing Zhu, and Xin Yu. Position: Human-centric vision requires topological generalization beyond fixed skeletal topologies. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026b.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pp. 792–804, December 2021. doi: 10.1109/3DV53792.2021.00088.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.48550/ARXIV.2306.13394. URL https://doi.org/10.48550/arXiv.2306.13394.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Gemma Team. Gemma 3 technical report, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14375–14385. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01363. URL https://doi.org/10.1109/CVPR52733.2024.01363.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Muhammed Kocabas, Nikolaos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: part attention regressor for 3d human body estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 11107–11117. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01094. URL https://doi.org/10.1109/ICCV48922.2021.01094.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016. URL http://arxiv.org/abs/1602.07332.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: carrying location information in full frames into human pose and shape estimation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, volume 13665 of *Lecture Notes in Computer Science*, pp. 590–606. Springer, 2022. doi: 10.1007/978-3-031-20065-6_34. URL https://doi.org/10.1007/978-3-031-20065-6_34.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten

Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024a. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, 2024b. DBLP:conf/eccv/LiuDZLZZYWHLCL24.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. In *TOG*, 2015.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL https://amass.is.tue.mpg.de.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single RGB camera. *CoRR*, abs/1705.01583, 2017. URL http://arxiv.org/abs/1705.01583.

Mistral AI. Introducing mistral 3. Blog post, 2025. Accessed 2025-12.

OpenAI. Gpt-5.1, 2025. URL https://platform.openai.com. Proprietary vision–language model accessed via OpenAI API.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Bryan A. Plummer, Liwei Wang, Chris Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 11159–11168. IEEE, 2021. doi: 10.

1109/ICCV48922.2021.01099. URL https://doi.org/10.1109/ICCV48922.2021.01099.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

World Health Organization. Global report on assistive technology: Summary. World Health Organization, December 2022. URL https://www.who.int/publications/i/item/9789240049178. ISBN: 978-92-4-004917-8.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *CoRR*, abs/2412.10302, 2024. doi: 10.48550/ARXIV.2412.10302. URL https://doi.org/10.48550/arXiv.2412.10302.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fbb10d319d44f8c3b4720873e4177c65-Abstract-Conference.html.

Xiangyu Yin, Boyuan Yang, Weichen Liu, Qiyao Xue, Abrar Alamri, Goeran Fiedler, and Wei Gao. Progait: A multi-purpose video dataset and benchmark for transfemoral prosthesis users. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8984–8993, 2025.

Jiaying Ying, Heming Du, Kaihao Zhang, Lincheng Li, and Xin Yu. Ldpose: Towards inclusive human pose estimation for limb-deficient individuals in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9865–9875, October 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *CoRR*, abs/2308.02490, 2023. doi: 10.48550/arXiv.2308.02490. DBLP:journals/corr/abs-2308-02490.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KOTutrSR2y.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

A Extended Related Work

A.1 Population coverage and body representation

Human pose, motion, and reconstruction resources support a broad range of person-understanding tasks (Lin et al., 2014; Andriluka et al., 2014; Ionescu et al., 2014; Mahmood et al., 2019; Mehta et al., 2017). Pose estimators operate over the keypoint vocabularies defined by these resources (Wei et al., 2016; Cao et al., 2017; Cheng et al., 2020; Sun et al., 2019; Xu et al., 2022). Parametric body models and reconstruction methods also use predefined body structures (Loper et al., 2015; Pavlakos et al., 2019; Kanazawa et al., 2018; Kolotouros et al., 2019; Kocabas et al., 2020; 2021; Sun et al., 2021; Li et al., 2022; Feng et al., 2021). These representations do not by themselves provide a dedicated evaluation of residual-limb endpoints, prosthetic devices, or their association with body regions. The position paper on topological generalization argues for human-centric vision beyond fixed skeletal topologies (Du et al., 2026b). Its concern is which body structures a representation can accommodate. Our concern is whether a VLM's answer agrees with the visible limb evidence in the evaluated image.

LDPose represents residual-limb endpoints and distinguishes them from downstream joints that do not exist (Ying et al., 2025). Its evaluation examines localization and predicted body structure. ProGait provides videos, segmentation and pose annotations, and textual gait descriptions for individuals with transfemoral prostheses (Yin et al., 2025). AJAHR addresses mesh recovery when joints are absent (Cho et al., 2025). Work on prosthetics-oriented vision further motivates attention to device appearance and body-device relations (Cognolato et al., 2022). These resources strengthen inclusive perception, but their targets differ from the language labels and option sets scored in InclusiveVLM-LimbEvidence. We reuse images from LDPose and create new evidence masks, region labels, and task targets. The reuse concerns images rather than the new annotation assets.

A.2 Accessibility and counterfactual evaluation

VizWiz connects evaluation to photographs and questions contributed by blind users (Gurari et al., 2018). It demonstrates the importance of the population supplying the images and the information that users seek. InclusiveVLM-LimbEvidence shares this attention to population coverage but studies limb evidence in images of individuals with limb deficiencies. The target is the accuracy of the visual interpretation, rather than a general assessment of disability-related conversation.

Du et al. (2026a) evaluate VLMs through counterfactual depictions of individuals with limb deficiencies. Response changes across depictions and agreement with an image-specific target address complementary questions. InclusiveVLM-LimbEvidence evaluates the latter through evidence detection, body-region attribution, and compatibility-style structured decisions. An unchanged answer is not necessarily correct, and a changed image can require a changed answer. This distinction also motivates separating paraphrase consistency from correctness.

A.3 Image-conditioned answers and region-language alignment

VQA establishes image-conditioned answering, while VQAv2 uses complementary image pairs to reduce reliance on question-only priors (Antol et al., 2015; Goyal et al., 2017). GQA examines compositional reasoning over real images (Hudson & Manning, 2019). These resources motivate checking an answer against image content rather than judging linguistic plausibility alone. ReferItGame and RefCOCO connect referring expressions with image regions (Kazemzadeh et al., 2014; Yu et al., 2016). Flickr30k Entities supplies region-to-phrase correspondences, and Visual Genome provides region descriptions and object relations (Plummer et al., 2015; Krishna et al., 2016). In LimbEvidence-Attribution, an output associates an evidence type with a named body region and side. Masks support target construction, but the model is scored on label sets rather than predicted coordinates or masks. Here, attribution means assigning evidence to the body, not explaining a model's internal computation.

A.4 Broad benchmarks and output protocols

MMBench, MMMU, and MME evaluate broad multimodal capabilities (Liu et al., 2024a; Yue et al., 2024; Fu et al., 2023). MathVista and ScienceQA examine mathematical and scientific reasoning (Lu

et al., 2024; 2022). MM-Vet evaluates integrated capabilities through open-ended responses and model-based scoring (Yu et al., 2024). These benchmarks use different answer formats and scoring mechanisms. Their common contribution is comparison under defined protocols, not one universal evaluation interface.

INCLUSIVEVLM-LIMBEVIDENCE applies structured evaluation to one population. Constrained and FreeForm address the same attribution target through parallel output formats. Category and Diversity address coarse category relations and complete matching sets through parallel probes. Fixed answer spaces enable deterministic comparison with targets, but do not establish that option positions, label frequencies, or answer-count preferences have no effect. The paired-format results and the analytical references in Appendix C.9 address specific aspects of this distinction.

A.5 CORRECTNESS, UNCERTAINTY, AND CONSISTENCY

HallusionBench examines language hallucinations and visual illusions (Guan et al., 2024). Calibration relates confidence to observed correctness (Guo et al., 2017). Dropout-based approximations and deep ensembles estimate predictive uncertainty (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Other work examines out-of-distribution inputs (Hendrycks & Gimpel, 2017; Liang et al., 2018), selective prediction (Geifman & El-Yaniv, 2019), and prediction sets with coverage guarantees under stated assumptions (Angelopoulos & Bates, 2021). These approaches concern distinct properties. INCLUSIVEVLM-LIMBEVIDENCE measures response variation across paraphrases alongside target correctness. It does not implement the cited uncertainty methods or infer calibrated confidence from stable answers. The scoring definitions appear in Appendix C.2, and the interpretation appears in Appendix D.

B BENCHMARK CONSTRUCTION AND SPECIFICATION

B.1 DESIGN ORIGIN AND EVALUATION REQUIREMENTS

INCLUSIVEVLM-LIMBEVIDENCE arose from errors in VLM prototypes for preparing visual evidence in Paralympic athlete classification workflows. Models missed visible residual-limb or prosthetic-device evidence or assigned it to the wrong body region. Later reasoning could then begin from an incorrect visual premise. We defined the evaluation targets before creating the evidence masks, region labels, and body-device records needed to score them. The authors selected the probes. An International Paralympic Committee certified classifier supported annotation guidance and case adjudication but did not select the probes. This was not a formal participatory co-design study.

The design distinguishes three questions. Which evidence is visible? Where does it belong on the body? Which presented options satisfy the benchmark relation? A correct response to the first question does not specify the answer to the other two. Constrained and FreeForm vary the output format within Attribution. Category and Diversity vary the matching requirement within Compatibility. Table 3 records the input, output, scale, and scoring target for each evaluation.

The progression is a decomposition of requirements, not a serial inference pipeline or a claim that every item is harder than every item in the preceding task. The model outputs of one probe are not supplied as inputs to another. The direct evidence tasks and Compatibility use different image subsets. Only the stated Constrained and FreeForm comparison uses a paired subset with an unchanged image and target. The common basis is the annotation framework, not an assertion that one atomic evidence item is tracked through all five evaluations.

B.2 SOURCE IMAGES, FILTERING, AND COUNTS

The 16,850 images come from the LDPose source pool (Ying et al., 2025). We reuse that collection rather than claim a separate image acquisition effort. The documented filtering process retains images with sufficient visible limb evidence or body-device configuration information for the assigned tasks. Near-duplicate removal uses perceptual hashes followed by visual review. Manual checks also remove redundant copies and highly similar frames. These steps reduce repetition but do not establish that all photographs depict different people or come from independent sources.

Table 3: **Specification of INCLUSIVEVLM-LIMBEVIDENCE evaluation requirements.** Presence and Attribution evaluate visible evidence, while Compatibility adds benchmark-defined matching. Constrained and FreeForm are parallel formats, and Category and Diversity are parallel probes. Counts describe task pools and the paired diagnostic, not completed coverage for every model. Source images and evaluation records are different units.

Evaluation	Records / source images	Required output	Primary score and scope
LIMBEVIDENCE-PRESENCE	26,406 / 8,802	One of four labels for residual-limb evidence only, prosthetic-device evidence only, both, or neither.	Accuracy. Evidence detection, not body-region assignment.
LIMBEVIDENCE-ATTRIBUTION Constrained	26,406 / 8,802	All applicable labels from eight segments crossed with two evidence types. `none` denotes an empty set.	Set mIoU. Label-set overlap, not pixel-mask overlap.
LIMBEVIDENCE-ATTRIBUTION FreeForm	500 paired image-target items	Text normalized to the same canonical attribution target.	Parsed set mIoU. Output-format diagnostic.
LIMBEVIDENCE-COMPATIBILITY Category	8,048 / 8,048	One category from the accepted set. More than one category can be accepted.	Accepted-set hit at one. No requirement to return the entire accepted set.
LIMBEVIDENCE-COMPATIBILITY Diversity	12,616 / 8,048	The complete set of matching options from the presented grid, including an empty set where applicable.	Sample-average F1. Exact, Partial, precision, and recall give separate diagnostics.

The direct evidence subset contains 8,802 images. Presence and Constrained Attribution each have 26,406 prompts, giving three prompts per image-task pair and 52,812 structured prompts in total. The 500-item FreeForm evaluation is a separate output-format diagnostic sampled from this attribution pool. It is not another collection of 500 independent people. Compatibility uses 8,048 original images. Category contains 8,048 evaluated items, and Diversity contains 12,616 evaluated items. Their combined task count is 20,664, not 12,616. Candidate device crops come from 7,497 source images, and one image can supply more than one crop. The candidate-source count is not added to the 16,850 benchmark images as a separate population sample.

An original image, a prompt, an option grid, a candidate crop, and a person are different counting units. The original-image counts define the sampling units used for the four new 235B confidence intervals. They do not establish person-disjoint or source-disjoint evaluation. The released task identifiers and mappings are needed to reproduce the item counts and their reuse.

B.3 NEW ANNOTATIONS AND VISIBLE-EVIDENCE SCOPE

We create pixel-level masks of visible residual-limb and prosthetic-device regions, region labels, and body-device configuration records. These are new benchmark annotations rather than reused LDPose keypoint targets. The records provide the targets for the language-based probes. They describe visible evidence, not the individual's identity, medical history, lived experience, or full anatomical state. The source population is individuals with limb deficiencies, and the tasks do not ask a model to infer whether the cause is congenital or acquired.

The direct evidence vocabulary contains upper arm, forearm, thigh, and calf on the left and right sides. Left and right refer to the individual, not the image viewer. Crossing these eight segments with residual-limb and prosthetic-device evidence yields 16 segment-type labels. Residual-limb and prosthetic-device masks do not overlap at the pixel level, but both evidence types can occur within the same segment. The aggregation therefore allows both segment-type labels where both regions are visible.

The model does not receive a pixel-localization score. For example, the target `thigh_l_prosthesis` associates prosthetic-device evidence with the left thigh. The corresponding Presence target can be prosthetic-device evidence only when no residual-limb evidence is visible elsewhere in the evaluated image. More detailed attribution is not obtained by scoring the same answer with a stricter pixel threshold. It requires a different label structure.

Intact limbs and full-body parsing are outside the annotation scope. No evidence mask on a region does not prove that an intact limb is present or absent. Clothing, occlusion, or a socket can conceal relevant regions. The target remains the annotated visible evidence rather than a reconstructed hidden body. This distinction applies to both annotation interpretation and downstream reuse of the released masks.

B.4 ANNOTATION GUIDANCE AND REVIEW

Annotators receive visual guidelines on residual-limb endpoints, prosthetic-device appearance, and the region vocabulary. The classifier supports onboarding and adjudication of cases that require additional guidance. A second annotator reviews each mask and derived label set. Disagreements above the annotation tolerance lead to re-annotation and adjudication. Periodic spot audits inspect randomly sampled batches, and recurring issues inform guideline updates.

The reviewed masks, region labels, accepted categories, and option sets form the reference targets for evaluation.

B.5 DIRECT EVIDENCE TARGETS AND PARALLEL OUTPUT FORMATS

LIMBEVIDENCE-PRESENCE. The answer space contains four labels. Label a denotes residual-limb evidence only. Label b denotes prosthetic-device evidence only. Label c denotes both evidence types. Label d denotes neither evidence type being visible. The task asks for one image-level answer, not a person-level diagnosis. The current 235B evaluation pool has gold labels a, b, and c but no gold label d. This pool can measure erroneous predictions of d, but does not estimate correct rejection on images whose target is d.

LIMBEVIDENCE-ATTRIBUTION. Constrained requires all applicable segment-type labels, with `none` denoting an empty label set. FreeForm asks for text that is normalized to the same canonical attribution target. The paired evaluation contains 500 randomly sampled image-target pairs from Constrained. Keeping the target fixed makes the comparison informative about the requested output format and its normalization. It is not a full assessment of unrestricted image description quality.

Equivalent wording. A paraphrase group contains prompts for one image and target that differ in wording. Groups support a response-stability diagnostic without changing the evidence target. The task score and consistency score answer different questions. Their definitions and denominators are given in Appendix C.8.

B.6 COMPATIBILITY RELATIONS, ANSWER SETS, AND OPTION CONSTRUCTION

LIMBEVIDENCE-COMPATIBILITY. Image annotations specify the target body-part categories, and candidate annotations specify the body part represented by each device. Matching follows the fixed correspondence between these categories. For example, an upper-limb device is excluded when the image has only lower-limb target categories. Devices with different appearances can share the same body-part relation. The accepted categories and matching option sets are recorded and reviewed during annotation. Category and Diversity use separate grids to assess category matching and complete option-set recovery.

Category. The model selects one category. Let A_i denote the accepted category set for item i and let $\hat{c}_i$ be the selected category. The response is correct when $\hat{c}_i \in A_i$. The four category names in the single-target diagnostic are `arm_lower`, `arm_upper`, `leg_lower`, and `leg_upper`. Some images support more than one accepted category. For these items, selecting any one accepted category satisfies the question. The model is not required to return the entire accepted set. Literal set equality is a separate diagnostic and is not substituted for the primary hit-at-one score.

Diversity. Let O_i be the presented option identifiers and $Y_i \subseteq O_i$ the matching target set. The model selects a set $\hat{Y}_i$. Exact recovery requires $\hat{Y}_i = Y_i$. Selecting one target is insufficient when other targets are omitted or distractors are included. This output demand explains why partial hits, exact equality, precision, and recall are reported separately. The target set can be empty.

Option pool. The documented construction uses device crops from the source pool, randomized option order per instance, and sampled distractors. Each evaluated grid has fixed option identifiers and a fixed target. Construction-time randomization is not an experimental test of robustness to option permutation. Once an instance is evaluated, its grid and ordering must remain identifiable for comparison. The full matching relation is determined by its annotation record, not by the illustrated example alone.

Three distinct empty or uncertain outputs. An empty Diversity target states that none of the presented options matches the relation. An insufficient-information option, where provided, is an explicit answer for inadequate visible evidence. A parser failure means no valid canonical response was recovered. These cases must not be interpreted as interchangeable. Likewise, the Presence label d concerns absence of visible evidence, not a claim that an individual has no limb deficiency. The reported empty-target errors do not establish performance on uncertainty-based abstention.

B.7 Records and test-only use

Each instance is represented by a machine-readable record with stable identifiers, a task name, prompt text, answer space, target labels, and derived metadata. Direct evidence records contain paraphrase-group identifiers. Compatibility records identify the grid, option order, accepted answers, and an insufficient-information choice where applicable. These relationships distinguish repeated wording from repeated imagery and repeated imagery from new candidate grids. Display names in the paper do not justify changing the task identity or target in an existing record.

InclusiveVLM-LimbEvidence provides no benchmark-internal training split. Its role is to compare general-purpose VLM outputs against fixed evaluation targets. A test-only split does not establish that the source images are absent from a model's pretraining corpus. Future adaptation should use separate development data and document any overlap with the evaluation resource. The present work reports evaluation and diagnostic analyses rather than benchmark-specific model training.

C Evaluation Protocol and Additional Results

C.1 Models, inference, and computation

The main comparison includes GPT-5, Gemini-2.5, Qwen3-VL, Gemma-3, DeepSeek-VL2, and Ministral-3. The scale comparison adds Qwen3-VL-235B-A22B-Instruct to the same protocol used for Qwen3-VL-30B-A3B-Instruct. The medical-domain control compares MedGemma 1.5 4B with Gemma-3 4B on matched examples. Inst, Think, and Reason in the main tables denote Instruct, Thinking, and Reasoning. These are evaluated checkpoints, not new methods trained on the benchmark.

For fixed-answer open-weight evaluations, the documented pipeline uses constrained decoding where supported and deterministic normalization of the response. Closed-model evaluations use provider interfaces with the corresponding benchmark questions and option identifiers. Multi-select responses are converted to canonical sets. FreeForm uses deterministic parsing and aliases rather than a generative model judge. These procedures make scoring repeatable but do not remove possible effects of the requested output format.

The common environment recorded for most open-weight runs used nodes with two NVIDIA H100 GPUs, Python 3.13, CUDA 12.6, NVIDIA driver 580.126.09, and vLLM 0.12.0. This is the reported common environment, not a model-specific hardware allocation for every run. In particular, it does not specify the memory or serving configuration of the 235B evaluation. Account identifiers, credentials, institutional paths, and non-anonymous infrastructure details are not included in the paper. The reported model settings and scorer configuration should be distinguished from provider endpoints that may change after a run.

C.2 Metrics, averaging, and statistical units

Presence accuracy is the fraction of prompts with the correct image-level category. Category accuracy is accepted-set hit at one as defined in Appendix B.6. For set-valued targets, let Y_i and $\hat{Y}_i$ denote the target and predicted sets. Exact and Partial are

$$\text{Exact} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\hat{Y}_i = Y_i], \qquad \text{Partial} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[|\hat{Y}_i \cap Y_i| > 0].$$

Partial does not penalize additional selections once one target has been found. Exact includes both nonempty and empty targets. A correct empty prediction therefore contributes to Exact but cannot contain a target hit under this Partial definition.

For items with nonzero relevant denominators, the overlap quantities are

$$J_i = \frac{|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i \cup Y_i|}, \qquad P_i = \frac{|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i|}, \qquad R_i = \frac{|\hat{Y}_i \cap Y_i|}{|Y_i|}, \qquad F_i = \frac{2|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i| + |Y_i|}.$$

Constrained Attribution reports the mean set intersection over union, not pixel-mask overlap. Diversity's primary F1 averages item-level F_i values. Micro precision, recall, and F1 pool option-level counts before calculating their ratios. Macro metrics average label-level quantities over the evaluation inventory. Empty-set conventions and the inclusion of zero-support labels are part of that inventory and scorer configuration. They must remain fixed across comparisons and cannot be inferred from a printed aggregate alone.

The mean of item F1 scores need not equal the harmonic mean of the printed average precision and recall. Likewise, sample-average F1 and micro-F1 should not be substituted for each other. The 235B Diversity table reports 25.90% sample-average F1 and 31.80% micro-F1 as different quantities. Exact-set outcomes and target supports complement the overlap metrics.

Confidence intervals. For the earlier evaluations, confidence intervals use the recorded item-level or paraphrase-group bootstrap protocol. The four 235B primary intervals use 10,000 bootstrap resamples of original images. The resampling unit retains prompts or grids associated with each sampled original image. This addresses reuse of an image within an evaluation, not dependence between different photographs of the same person or source. Table 6(a) reports original interval limits and both prompt-level and image-level denominators.

The main tables use one-line $x \pm h$ cells, with $h = (U - L)/2$ for these four new intervals. This records interval width but does not reproduce small asymmetries around the point estimate. The original limits remain available in the appendix. A missing interval is not a zero-width interval. Point differences are descriptive unless a paired interval is explicitly reported. Overlap of marginal intervals is not used here as a test of a paired model difference. Human item evaluations, model prompts, original images, and distinct people are not interchangeable sample sizes.

C.3 Human reference, instructions, and rater background

Five raters each completed 200 items, producing 1,000 item evaluations. Items were sampled randomly within fixed quotas for Presence, Constrained Attribution, Category, and Diversity. The completed study did not use predefined easy and hard strata, and it did not evaluate FreeForm. Assignments overlapped partially but were not identical across all raters. Consequently, 1,000 responses should not be described as 1,000 independent images.

Three raters had primarily computer-science backgrounds and two had extensive disability-research experience. Lived-experience status was not established, and individuals with limb deficiencies were not recruited as a defined rater group. The study also did not recruit a professional prosthetics panel. Relevant research experience and clinical qualification are therefore not used as equivalent descriptions of the raters.

Raters received definitions of residual-limb and prosthetic-device evidence and the eight-segment taxonomy. They selected one answer for Presence and Category, all supported segment-type labels for Constrained Attribution, and the matching option set for Diversity. Instructions asked them to use visible evidence and not infer labels from clothing, posture, wheelchairs, or external context alone. For a partially visible region or uncertain side, they made the most visually supported judgment under the existing task instructions. They were told not to search for the source photograph, person, dataset, or paper. These instructions describe the completed study and do not introduce a new human abstention protocol.

Raters were informed of the task purpose, expected time commitment, image content, optional lived-experience field, and use of aggregated results in the paper. The interface recorded worker identifiers, responses, optional background information, and timing metadata. No identifying rater information is disclosed. Human responses were scored against the reviewed benchmark targets.

Table 4: **Human reference and paired output-format diagnostics.** (a) Diversity F1 is averaged over items within raters and then equally over raters. The group ratio is 1.58 with 95% interval $[1.16, 2.22]$, and the overall F1 is 57.9%. (b) Each model uses the same 500 image-target pairs. Scores are set mIoU percentages. Parse coverage indicates recoverable labels, not correct attribution.

(a) Human reference by rater background

Background	Raters	Per-rater mean F1	Group mean F1
Disability research	2	77.78, 70.67	74.22
Computer science	3	43.17, 39.02, 58.89	47.03

(b) Paired Attribution output formats

Model	Constrained	FreeForm	Parse coverage	Invalid outputs
Qwen3-VL-4B-Think	17.22	43.1	467/500 (93.4%)	33/500 (6.6%)
Qwen3-VL-8B-Think	15.87	44.7	452/500 (90.4%)	48/500 (9.6%)
GPT-5	26.51	30.5	417/500 (83.4%)	83/500 (16.6%)

Performance by rater background. The human reference reaches 98.0% Presence accuracy and 83.4% Constrained Attribution set mIoU. The corresponding Attribution micro-F1 and macro-F1 are 85.0% and 73.1%. Presence macro-F1 is 69.6%, which weights categories differently from aggregate accuracy.

The two raters with disability-research experience achieve Diversity F1 values of 77.78% and 70.67%. Their mean is 74.22%, compared with 47.03% for the other three raters. The reported ratio is 1.58 with a 95% interval of $[1.16, 2.22]$. Table 4(a) presents the individual and group scores. The background-specific results show the higher performance of the experienced participants that the mixed-background average obscures.

A bird-identification task can produce different scores for experienced birdwatchers and general participants even when both groups use the same answer rules. The analogy explains why background composition is informative, rather than establishing a cause of the difference here. Because the groups did not answer identical item sets, the observed contrast is not a controlled estimate of the effect of experience. No whole-study agreement coefficient is reported. Partial assignment overlap could support agreement analysis on overlapping items, but aggregate group scores do not provide the required response matrix. The F1 contrast does not establish an expert Exact score or a professional performance ceiling.

Overall human aggregation. The overall Diversity F1 is 57.9%, obtained by equally averaging the five rater-level mean F1 scores. Table 2 reports this estimate and its confidence interval. Background-specific results are reported separately in Table 4(a).

C.4 Scale, inference style, and the medical-domain control

Table 5(a) compares Qwen3-VL-30B-A3B-Instruct and 235B-A22B-Instruct. Presence accuracy rises by 50.33 percentage points, while Constrained Attribution set mIoU rises by 0.76 points and Category accuracy by 4.30 points. Diversity F1 decreases by 25.30 points, and Exact decreases by 4.53 points. The improvement is concentrated in image-level detection rather than shared across all evaluated requirements. The comparison concerns these checkpoints and settings. It does not isolate parameter count from differences in training or other model properties.

Table 5(b) gives six same-family, same-scale comparisons of Thinking or Reasoning with Instruct. No pair improves all four primary probe scores. At Qwen3-VL 4B and 8B, Thinking improves Diversity F1 and Exact while lowering Presence and Category accuracy. At 30B-A3B, it improves Presence but lowers the other listed scores. For Ministral-3, Reasoning increases Diversity F1 at all three scales, but Exact decreases at 3B and 8B and is unchanged at the displayed precision at 14B. These are checkpoint and inference-style comparisons, not a controlled intervention on reasoning alone.

Table 5(c) reports MedGemma 1.5 4B relative to Gemma-3 4B on matched examples. The medical model is lower by 7.87 points in Presence, 12.31 points in Attribution set mIoU, and 4.56 points in

Table 5: **Model scale, inference style, and medical-domain comparisons.** (a) Qwen3-VL Instruct results and their differences. (b) Same-family, same-scale differences relative to Instruct. (c) Matched medical-domain differences relative to Gemma-3. Scores are percentages and differences are percentage points. Only the magnitude of the medical Category difference is reported, and its paired interval includes zero. Dashes indicate unavailable results. The differences do not imply statistical significance.

(a) Model scale. Qwen3-VL Instruct scores and 235B minus 30B differences

Model	Presence accuracy	Attribution set mIoU	Category accuracy	Diversity Exact	Diversity F1
30B-A3B-Instruct	23.9	24.0	35.4	7.6	51.2
235B-A22B-Instruct	74.23	24.76	39.70	3.07	25.90
Difference	+50.33	+0.76	+4.30	−4.53	−25.30

(b) Inference style. Thinking or Reasoning minus Instruct

Pair	Presence accuracy	Attribution set mIoU	Category accuracy	Diversity Exact	Diversity F1
Qwen3-VL 4B	−1.3	+0.8	−11.8	+5.5	+17.4
Qwen3-VL 8B	−6.1	−2.6	−12.1	+5.8	+17.0
Qwen3-VL 30B-A3B	+9.7	−4.6	−12.8	−4.9	−21.5
Ministral-3 3B	−28.6	+0.2	+7.6	−2.2	+6.9
Ministral-3 8B	−12.3	+1.5	−0.1	−0.6	+7.9
Ministral-3 14B	−4.3	−8.5	−1.9	+0.0	+13.9

(c) Medical-domain control. MedGemma 1.5 4B minus Gemma-3 4B

Comparison	Presence Acc.	Attribution mIoU	Category Acc.	Diversity Exact	Diversity F1
Matched examples	−7.87	−12.31	$\lvert\Delta\rvert = 1.12$	–	−4.56

Diversity F1. The Category difference is reported only by its magnitude of 1.12 points, with a paired interval that includes zero.

MedGemma's documented medical-image and text domains differ from the open-world limb evidence studied here (Sellergren et al., 2025). The comparison shows that this specialization does not ensure improvement on the reported tasks. It is not a conclusion about every medical VLM or a causal measurement of training-data composition.

C.5 Paired Attribution output formats

The paired-format diagnostic uses the same 500 image-target pairs for Constrained and FreeForm evaluation within each model. Table 4(b) contains the available matched summaries for Qwen3-VL-4B-Thinking, Qwen3-VL-8B-Thinking, and GPT-5. The Constrained scores are 17.22%, 15.87%, and 26.51% set mIoU. Their FreeForm scores are 43.1%, 44.7%, and 30.5%, respectively. These paired Constrained values must not be replaced by full-pool values from the main table. The diagnostic does not create a second full-scale attribution dataset.

The parser recovers a valid canonical prediction from 467, 452, and 417 of the 500 responses. Invalid outputs account for the remaining 33, 48, and 83 responses. Valid parsing indicates a recoverable label set, not correct attribution. The observed changes reflect the requested language format and its deterministic normalization. They establish that the output interface matters to measurement, rather than showing that label constraints are irrelevant.

C.6 235B recognition, attribution, and output validity

Table 6(a) reports the four primary results for Qwen3-VL-235B-A22B-Instruct and their image-level bootstrap intervals. For Presence, the model answers 19,601 of 26,406 prompts correctly, giving 74.23% accuracy. Target counts are 7,302 for residual-limb evidence only, 17,373 for prosthetic-device evidence only, and 1,731 for both types. Label d has no target examples in this evaluation pool. Macro precision, recall, and F1 are 49.74%, 41.70%, and 43.81%, respectively. Parsing and protocol error rates are both 0%, with 100% output coverage.

For Constrained Attribution, set mIoU is 24.76% and exact correctness is 17.10%. Micro precision, recall, and F1 are 35.30%, 33.08%, and 34.16%. Macro precision, recall, and F1 are 29.17%, 21.99%,

Table 6: **Qwen3-VL-235B scores and recognition diagnostics.** (a) Original 95% limits from 10,000 image-level bootstrap resamples. (b) Presence labels a, b, c, and d denote residual only, prosthesis only, both, and neither. (c) Component correctness and separate target-assisted oracle gains. (d) Joint region-type F1. (e) Raw consistency uses all groups, while gated consistency retains groups with at least one correct response. Coverage and wrong-and-stable rates use all groups. Scores are percentages unless marked otherwise.

(a) Primary scores and original-image bootstrap intervals

Probe	Primary metric	Score [95% interval]	Instances	Original images
Presence	Accuracy	74.23 [73.39, 75.06]	26,406	8,802
Attribution-Constrained	Set mIoU	24.76 [24.05, 25.49]	26,406	8,802
Compatibility-Category	Accepted-set hit at one	39.70 [38.61, 40.76]	8,048	8,048
Compatibility-Diversity	Sample-average F1	25.90 [25.41, 26.40]	12,616	8,048

(b) Presence classification and confusion counts. Rows are targets

Target	Support	Precision	Recall	F1	Predicted label			
					a	b	c	d
a	7,302	97.82	54.15	69.71	3,954	1,136	556	1,656
b	17,373	86.46	87.57	87.01	51	15,213	1,966	143
c	1,731	14.68	25.07	18.52	37	1,246	434	14
d	0	–	–	–	0	0	0	0

(c) Attribution components and separate oracle corrections

Scored component	Correctness (%)	Oracle mIoU gain (points)
Evidence type	74.40	–
Body region	40.11	+16.32
Side	40.59	+21.41
Complete target	17.10	–

(d) Attribution label F1. Left and right refer to the individual

Body region	Residual left	Residual right	Prosthesis left	Prosthesis right
Upper arm	27.83	24.98	32.04	17.50
Forearm	19.83	17.26	26.43	24.88
Thigh	25.54	25.72	29.02	30.02
Calf	16.87	16.82	40.12	49.18

(e) Prompt correctness and paraphrase-group consistency

Probe	Prompt correctness	Raw consistency	Gated consistency	Retained / all groups	Coverage	Wrong and stable
Presence	74.23	92.22	91.58	7,080 / 8,802	80.44	–
Attribution-Constrained	17.10	75.43	68.51	2,099 / 8,802	23.85	47.28

and 23.56%. The error rate is 0%, with 100% output coverage. Table 6(d) presents the segment-type F1 scores. Right-calf prosthetic-device evidence has the highest listed F1 at 49.18%.

The component analysis evaluates evidence type, body region, and side within the same Attribution responses. Their correctness rates are 74.40%, 40.11%, and 40.59%, respectively, as shown in Table 6(c). The model identifies the evidence type more accurately than its body region or side. This distinction localizes errors within the attribution output.

The oracle analysis replaces one predicted component with its target value and re-scores the output. Correcting body region increases set mIoU by 16.32 percentage points. Correcting side increases it by 21.41 points. Each correction is evaluated separately. These target-assisted corrections quantify the gains available from accurate body-region and side assignment.

C.7 235B ACCEPTED CATEGORIES AND SET-SELECTION OUTCOMES

The Category accepted-set hit-at-one score is 39.70%, compared with 32.63% literal set equality. The difference of 7.07 points reflects the distinct acceptance conditions. There are 878 items with multiple accepted categories and 7,170 with one accepted category. Their hit-at-one scores are 64.81% and 36.62%. For the single-target subset, macro precision, recall, and F1 are 33.06%, 35.00%, and 33.18%. The abstention and non-definitive-response rates are both zero. This records behavior under the given prompt, not a calibrated ability to decide when to abstain.

Table 7: **Qwen3-VL-235B compatibility diagnostics.** (a) Category hit at one accepts any single category in the target set. Literal set equality is diagnostic only. (b) Sample averages and micro scores use different aggregation rules. Set sizes and errors are measured in options. (c) Outcome counts cover all 12,616 Diversity instances. The model selects options on 1,674 of 1,851 empty-target instances, or 90.44%. Other scores are percentages.

(a) Category accepted answers and single-target classification

Evaluation subset or quantity	Instances	Score
All items, accepted-set hit at one	8,048	39.70
All items, literal exact match	8,048	32.63
Multiple accepted categories, hit at one	878	64.81
One accepted category, hit at one	7,170	36.62
Single-target macro precision	7,170	33.06
Single-target macro recall	7,170	35.00
Single-target macro F1	7,170	33.18
`arm_lower` F1	–	25.16
`arm_upper` F1	–	26.60
`leg_lower` F1	–	43.22
`leg_upper` F1	–	37.72

(b) Diversity overlap and selection size

Aggregation	Precision (%)	Recall (%)	F1 (%)
Sample average	49.75	21.37	25.90
Micro	53.36	22.64	31.80

Quantity	Value	Quantity	Value
Exact (%)	3.07	Partial (%)	57.07
Mean predicted size	1.63	Mean target size	3.83
Mean target minus predicted	2.21	Mean absolute size error	2.98
Predicted size smaller (%)	68.0	Predicted size larger (%)	20.5
Predicted size equal (%)	11.4		

(c) Diversity prediction outcomes

Outcome	Instances	Percentage
Nonempty proper subset of target	4,916	38.97
Missed and extra options	4,143	32.84
Nonempty prediction for empty target	1,674	13.27
Empty prediction for nonempty target	1,048	8.31
Strict superset of nonempty target	448	3.55
Exact nonempty set	210	1.66
Exact empty set	177	1.40
Total	12,616	100.00

Diversity sample-average precision, recall, and F1 are 49.75%, 21.37%, and 25.90%. Micro aggregation gives 53.36%, 22.64%, and 31.80%. The mean predicted set contains 1.63 options, compared with 3.83 for the target. The reported mean target-minus-predicted size is 2.21 and the mean absolute size error is 2.98 options. These means were rounded separately. Predicted sets are smaller, larger, or equal in size on 68.0%, 20.5%, and 11.4% of instances. Equal cardinality is not exact set equality.

The seven outcome counts in Table 7(c) sum to 12,616 instances. Nonempty exact predictions number 210 and empty exact predictions number 177, giving 387 exact predictions and 3.07% overall Exact. Among 1,851 empty-target items, 1,674 have a nonempty prediction, giving 90.44% false selection. Under-selection on some items therefore coexists with selection when no presented option matches. These outcome counts separate set-recovery errors from the effect of returning a valid output string.

C.8 Paraphrase consistency and retained coverage

For an eligible paraphrase group g with $|g| > 1$, let u_g be the number of distinct normalized predictions. The documented group score is

$$c_g = \max\left(0, 1 - \frac{u_g - 1}{|g| - 1}\right).$$

It equals one when the predictions are identical, including when all are incorrect. For all eligible groups $\mathcal{G}$, define

$$\mathcal{G}_+ = \{g \in \mathcal{G} : \exists i \; \hat{y}_{g,i} = y_g^*\}.$$

Raw consistency averages c_g over $\mathcal{G}$. Gated consistency averages it over $\mathcal{G}_+$, and retained coverage is $|\mathcal{G}_+|/|\mathcal{G}|$. For set tasks, correctness in this gate requires the exact normalized target. The wrong-and-stable rate counts groups with one repeated prediction different from the target, using all eligible groups as the denominator. If no group is retained, a gated mean is undefined rather than evidence of perfect stability.

Table 6(e) reports the available 235B diagnostics. Presence raw consistency is 92.22%, while gated consistency is 91.58% on 7,080 of 8,802 groups, or 80.44% coverage. Attribution raw consistency is 75.43%, with gated consistency of 68.51% on 2,099 of 8,802 groups, or 23.85% coverage. The reported Attribution wrong-and-stable rate is 47.28%. These quantities use group denominators, while 17.10% exact Attribution correctness uses prompts. They should not be read as one common-denominator accuracy breakdown.

The gate uses target correctness and is an analysis filter, not an inference-time abstention rule. A high value on retained groups says nothing by itself about excluded groups. Category has one prompt per group in the recorded protocol and has no informative paraphrase consistency score. FreeForm consistency is not reported.

C.9 Analytical references for prior-sensitive metrics

The following values are closed-form calculations from the supplied target counts and metric definitions. They are not additional model runs or completed perturbation experiments. For the 235B Presence pool, a uniform selector over the four options has expected accuracy of 25%. A constant label-b response has accuracy $17{,}373/26{,}406 = 65.79\%$. The observed 74.23% accuracy exceeds this constant-label reference by 8.44 percentage points. This comparison makes the class prior explicit and does not measure image use in isolation. It also does not transfer to a human or model subset with a different target distribution.

For valid canonical set outputs in the 235B Diversity pool, an always-empty prediction obtains Exact of $1{,}851/12{,}616 = 14.67\%$. Selecting every presented option obtains Partial of $10{,}765/12{,}616 = 85.33\%$ under the target-hit definition. Neither response uses the image to distinguish matches. Thus, the observed 3.07% Exact and 57.07% Partial should not be interpreted as above-prior evidence on their own. These references make clear why the benchmark reports overlap metrics, prediction sizes, and empty-target outcomes rather than treating one score as sufficient.

The outcome counts also give exact recovery of $210/10{,}765 = 1.95\%$ on nonempty targets and $177/1{,}851 = 9.56\%$ on empty targets. These are conditional calculations from the recorded counts, not new evaluations. They show that the low overall Exact does not arise solely from combining the two target types. No numerical result is claimed for an unexecuted random-set selector, image-removal control, option permutation, or newly trained mitigation. Those controls could address different explanations, but they are not replaced by the analytical references above.

D Interpretation, Coverage, and Intended Use

D.1 What the evidence establishes

The contribution is an evaluation resource centered on individuals with limb deficiencies and an analysis of the visual requirements on which current models make errors. Population coverage and task design address different needs. An image collection can include an individual without testing whether a model notices their residual limb, assigns device evidence correctly, or returns a complete matching set. The benchmark makes these distinctions visible in scores and inspectable examples. Its social motivation is that progress in visual AI should include this population rather than leave its relevant visual details outside evaluation.

The results support output-level conclusions under the recorded protocol. The attribution component analysis identifies spatial assignment errors within the same responses. The matching diagnostics distinguish missed targets, incorrect inclusions, and errors on empty targets. The consistency analysis shows why repeated answers cannot replace target correctness. These findings do not depend on treating the numerical difference between Presence accuracy and Attribution mIoU as a measured amount of lost information. Nor do they establish a serial failure rate through a system that performs all probes in sequence.

Annotation review establishes the reference targets used for evaluation. The human results describe performance against these targets and show higher Diversity F1 among participants with disability-research experience. We report these background-specific results alongside the overall human reference.

D.2 Task demands and alternative explanations

Recognition, vocabulary use, output format, and relation application can each affect performance. The benchmark separates observable requirements, but the scores do not isolate a latent visual capability from every other contributing factor. This is why the evidence-type, region, and side analysis, the paired output-format comparison, and the option-set diagnostics are reported separately. Each addresses a narrower question rather than attributing every error to one mechanism.

Compatibility uses accepted categories and option sets defined for the presented image and grid. Multiple accepted categories are handled explicitly by hit at one. Complete-set recovery is evaluated separately and has different error costs from retrieving any one target. Both sample composition and answer-count behavior affect the resulting metrics. The analytical constant-output references quantify this issue for the current 235B pools without claiming that all option-prior effects have been eliminated.

The scale and medical-domain comparisons show that a larger or differently specialized model does not uniformly improve the reported requirements. They concern selected checkpoints, not every model within a family or domain. The exact pretraining corpora are not available for estimating how often images of individuals with limb deficiencies were encountered. Insufficient population coverage in training is a possible explanation, not a prevalence estimate or causal result of this study. The empirical claim is that the evaluated capabilities still leave observable errors on the population-specific tasks.

D.3 Source coverage, visual conditions, and dependence

The source pool includes everyday, rehabilitation, and sport imagery, with variation in viewpoint, clothing, device appearance, and background. It is not a demographic census or a representative sample of all individuals with limb deficiencies. Public-source selection can favor particular activities, geographic contexts, photographic styles, or repeatedly depicted individuals. Filtering and near-duplicate removal reduce repetition without establishing population representativeness. No unreported demographic proportions or person-disjoint guarantees are inferred from image counts.

Visibility also limits the evidence available from a photograph. A sleeve, socket, crop, or occlusion can hide a region that would be observable from another view. The annotations and task instructions constrain the target to the evidence visible in the evaluated image. They do not recover medical

history or hidden anatomy. The eight-segment vocabulary is the annotation scope for this version, not a complete model of human anatomical variation.

Image-level resampling accounts for repeated prompts and grids associated with the same sampled image. It does not account for every form of dependence between different images or reused candidate crops. The earlier and additional intervals use their documented sampling units and should not be presented as if all had been recomputed under one new bootstrap procedure. The fixed benchmark version and evaluation records define the population of items to which a reported score applies.

D.4 Intended use and research implications

InclusiveVLM-LimbEvidence is intended for auditing VLM outputs on images of individuals with limb deficiencies. Compatibility means agreement with the benchmark's visual relation, not a recommendation of which device an individual should use. A photograph and option grid do not contain the individual assessment, preferences, and contextual information required for a personalized device decision. The evaluation does not determine clinical suitability, rehabilitation plans, eligibility for sport classification, or readiness for assistive deployment. The athlete-classification prototype explains the origin of the visual questions, not an application validated by the benchmark scores.

The observed errors suggest concrete research targets, including evidence-to-region assignment, laterality, complete-set recovery, and no-selection behavior when no option matches. The oracle results identify gains available with target-assisted correction, not a practical mitigation method that achieved those gains. Likewise, a correctness-based consistency gate is not an abstention policy available without a target. The present work does not report new training interventions or improved outcomes for end users. Its contribution is to make the relevant errors measurable so that later work can assess those improvements directly.

E Ethics, Access, and Reproducibility

E.1 Approval, participation, and privacy

The project received approval from an institutional Human Research Ethics Committee. Institution names, approval identifiers, and identifying administrative details are omitted for anonymous review. The documented protocol covers secondary use of source imagery, annotation, secure storage, controlled access, and risk mitigation. This approval is not presented as a transfer of image copyright or evidence of individual consent from every depicted person. The human rating study and secondary use of existing photographs are distinct activities.

Where direct individual consent to secondary use is not feasible, the documented approach relies on institutional review, proportionate risk assessment, access restrictions, and a correction or removal pathway. Illustrations are selected to support scientific interpretation while reducing identifiability where possible. Reduced identifiability is not a guarantee that a person cannot be recognized from remaining context. The task labels do not require personal identity, employability, productivity, or capability judgments about depicted individuals.

Disability-related imagery can be sensitive even when publicly accessible. The access agreement restricts non-commercial research use and prohibits re-identification, face recognition, profiling, and attempts to contact depicted individuals. It also requires secure storage and limits redistribution that would increase identifiability. The benchmark's intended use does not extend to surveillance or stigmatizing inference. Author-identifying funding and competing-interest information is reserved for the appropriate non-anonymous declarations.

E.2 Licenses and source-image continuity

Benchmark annotations, prompts, and derived metadata are designated for release under CC BY-NC-SA 4.0 (Creative Commons, 2013a;b). Source photographs remain subject to their own rights and the LDPose access conditions. The benchmark does not treat the annotation license as permission to redistribute third-party source images. Public benchmark artifacts and controlled source-image access are therefore documented separately.

The release plan will provide stable image identifiers, a benchmark-to-image manifest, checksummed annotations and records, scoring code, and reconstruction instructions. It will preserve a documented route to the exact image snapshot used for evaluation through the permitted source-access process. This is preferable to silently substituting a new download or a visually similar photograph when an original reference changes. Unavailable or removed records must be identified in later evaluations, with the resulting denominator changes reported.

Option pools and target definitions will be tied to versioned releases. A new device appearance, corrected target, or changed grid belongs to a documented update rather than an unannounced modification of a published instance. Prior results retain their original version identifiers. This supports comparison while allowing justified corrections and removal requests. We will validate the download and scoring path from a non-owner account before claiming end-to-end access verification.

A correction and takedown pathway accepts requests from rights holders, depicted individuals, and affected parties. Validated requests lead to removal or suppression of relevant records in future releases and a documented change to the usable manifest. Change logs should preserve the interpretation of earlier results without republishing sensitive identifying material. The controlled release is intended to support both evidence traceability and the rights of those represented.

E.3 Anonymous artifacts and reproduction steps

The anonymous artifact is intended to expose the benchmark definitions, annotations, prompts, sample records, scoring procedures, and source-access instructions without identifying authors or infrastructure. It does not re-host third-party photographs. Dataset-level metadata, including the described Croissant representation, must agree with the task counts, accepted-answer definitions, and access conditions in this appendix.

Independent reproduction requires the same image manifest, option grids and ordering, prompt set, normalization rules, target labels, and aggregation configuration. A reproduction should first recover the evaluation inputs, then normalize model outputs, and finally aggregate with the recorded denominators and resampling units. Matching a family name alone is insufficient when the checkpoint, prompt, or output interface differs. All inputs necessary for scoring are distinct from confidential access credentials or provider-account information.

The release will include benchmark annotations, task records, evaluation code, and reconstruction instructions. A complete execution record also needs the actual model-specific inference settings and scoring configuration.

E.4 AI assistance and evidential provenance

ChatGPT assisted manuscript language and organization, reference checking, and feedback on evaluation methodology. It was not used to originate the research idea or conduct autonomous research. The evaluated VLMs are research subjects, while benchmark targets follow the separate human annotation and review process. A generative model judge does not replace the deterministic target-based scoring described here.

The teaser is an illustration of the evaluation design, not an additional benchmark image, annotation source, or recorded model response. The quantitative claims rely on the reported experimental records and tables. The authors are responsible for checking the final descriptions, citations, and figures against their sources.

F Datasheet for InclusiveVLM-LimbEvidence

This datasheet follows the question families of the datasheets framework (Gebru et al., 2021). It summarizes the resource and points to the detailed specifications above rather than restating a second evaluation protocol.

Purpose and population. InclusiveVLM-LimbEvidence evaluates VLM interpretation of images of individuals with limb deficiencies. Its three requirements are evidence detection, body-region attribution, and compatibility-style structured decisions. The resource supports population-

centered auditing rather than measurement of the depicted individuals' abilities or experiences. Appendix B.1 explains why the requirements were chosen.

Sources, units, and scale. The resource uses 16,850 images from the LDPose pool and adds benchmark-specific evidence annotations and targets. Presence and Constrained Attribution each contain 26,406 prompts over 8,802 images. The paired FreeForm diagnostic contains 500 image-target pairs. Category contains 8,048 instances, and Diversity contains 12,616 instances over 8,048 original images. These counts refer to evaluation records and imagery, not distinct people. The 7,497 candidate-source images describe option provenance rather than an extra participant population. Appendix B.2 gives the counting rules.

Annotations and target interpretation. Annotations mark visible residual-limb and prosthetic-device evidence and its body region and side. They support a 16-label segment-type vocabulary and benchmark-defined accepted categories or option sets. Intact limbs, full-body parsing, medical history, and personal identity are not annotation targets. Review and adjudication are described in Appendix B.4. A no-match set, an insufficient-information choice, and an invalid response have different meanings.

Evaluation and human reference. Metrics include category accuracy, set overlap, exact and partial set recovery, and paraphrase consistency. The human reference comprises five raters with different research backgrounds and partially overlapping assignments. Their 1,000 responses are not 1,000 independent images. The paired-format results, rater-level summaries, and model diagnostics appear in Appendix C. No benchmark-internal training split is provided.

Coverage and limitations. The source distribution can favor particular activities, settings, or repeatedly depicted individuals. The current direct evidence evaluation has no target examples of Presence label d. Constrained outputs, option composition, scoring inventories, and parser behavior affect interpretation. Analytical prior-based references are distinguished from executed model runs in Appendix C.9. The benchmark does not certify person-disjoint generalization, absence of pretraining overlap, or end-user outcomes.

Distribution and maintenance. Benchmark annotations and metadata are designated for CC BY-NC-SA 4.0 release, while original imagery retains its source-access requirements. The code, record manifests, and reconstruction instructions are part of the release plan. Non-commercial access restrictions, correction and removal procedures, immutable option-pool versions, and review anonymity are described in Appendix E. The documentation distinguishes available results from planned access and maintenance checks. The intended-use limits in Appendix D.4 apply to every release.