WHEN VISION-LANGUAGE MODELS LOOK BUT DON'T SEE: ANATOMICAL BIAS IN ENDOSCOPIC SPATIAL REASONING

Diego Bravo[a], Daniel Wolf[b], Juan Hurtado-Tobar[a], Martín Gómez[a], and Eduardo Romero[a].

[a] Computer Imaging and Medical Applications Laboratory (CIM@LAB)
Universidad Nacional de Colombia
[b] Visual Computing Group, Institute of Media Informatics, Ulm University, Germany

ABSTRACT

Reliable spatial understanding is essential for complete mucosal inspection during upper gastrointestinal endoscopy, particularly under the Systematic Screening Protocol for the Stomach (SSS). Vision-Language Models (VLMs) show promise for assisting navigation but often rely on textual priors rather than visual information, a limitation known as anatomical bias. To evaluate spatial reasoning, we introduce **EndoSSS-RP**, a benchmark based on the GastroHUN dataset (380 patients, 2,796 images, 3,678 binary questions). Each question, targeting left/right or above/below relationships between gastric surfaces (e.g., anterior vs. posterior wall), is tested on original, flipped, and rotated images to assess model robustness in determining relative position across different orientations. Prompts are structured across three levels: L1 (anatomical terms only), L2 (anatomical terms with visual markers), and L3 (visual markers only). We evaluate four VLMs (GPT-4o, Gemini-2.5-Flash, JanusPro-7B, and LLaMA-3.2) and find that accuracy declines under transformations when using anatomical prompts (L1, GPT-4o: 68% original, 61% flipped, 46% rotated) but perform consistently better with visual–marker only prompts (L3: 87%). These results reveal anatomical bias and underscore the need for models to rely more heavily on visual evidence in endoscopic AI systems. Code, data, prompts, and evaluation scripts are available at: https://github.com/DiegoBravoH/EndoNavQA.

***Index Terms*—** Endoscopy, Vision-Language Model, Spatial Reasoning, Anatomical Bias, Gastric Surfaces.

1. INTRODUCTION

Upper gastrointestinal (UGI) endoscopy (EGD) is the primary modality for diagnosing and screening gastric disease, where complete mucosal inspection under the Systematic Screening Protocol for the Stomach (SSS) is essential, as mislocalizing gastric walls or curvatures can produce blind spots that contribute to missed early cancers [1, 2]. In practice, reliable spatial navigation is challenged by dynamic camera motion, frequent view inversions (antegrade/retroflexed), and partially visible or occluded landmarks [3, 4]. Despite its widespread use, EGD performance varies with cognitive and technical factors [5], with reported miss rates of 20–25% for early gastric cancer and 11.3% for UGI cancers overall [6, 5]. These limitations have driven interest in technological support for navigation, including the use of vision-language models (VLMs). As vision-language models are increasingly used for reporting and guidance, robust visual spatial reasoning (e.g., left/right, above/below) becomes a prerequisite for safe clinical use.

Recent research shows that VLMs often exhibit *anatomical bias*: instead of analyzing image evidence, these models tend to depend on memorized anatomical priors (pre-existing textual knowledge), leading to unreliable answers when images are reoriented [7]. While this has been observed in relatively static modalities (e.g., CT), endoscopy is inherently orientation-variant: camera angles change continuously, and "left" and "right" are defined relative to the camera rather than fixed patient anatomy. Standard maneuvers such as the J-turn and U-turn invert the visual field, during retroflexion in the SSS protocol, the posterior wall may appear on the left side of the screen, whereas in antegrade viewing, it appears on the right [3]. In this context, where complete mucosal inspection depends on accurately localizing gastric surfaces (walls and curvatures) across rapid viewpoint changes, an evaluation robust to flips and rotations is essential to isolate visual evidence. It prevents text-based shortcuts and directly measures the visual reasoning required to avoid blind spots.

We present **EndoSSS-RP** (Endoscopic Systematic Screening of the Stomach – Relative Positioning), a benchmark for spatial reasoning in UGI endoscopy. It evaluates left/right and above/below relationships between gastric surfaces using anatomical and marker-based prompts, tested on original and transformed images to assess orientation robustness.

Our main contributions are: (1) A visual question answering (VQA) benchmark for assessing spatial reasoning in UGI endoscopy using diverse prompt styles. (2) A reproducible evaluation framework for analyzing model sensitivity to prompt structure and orientation changes, revealing anatomical bias.

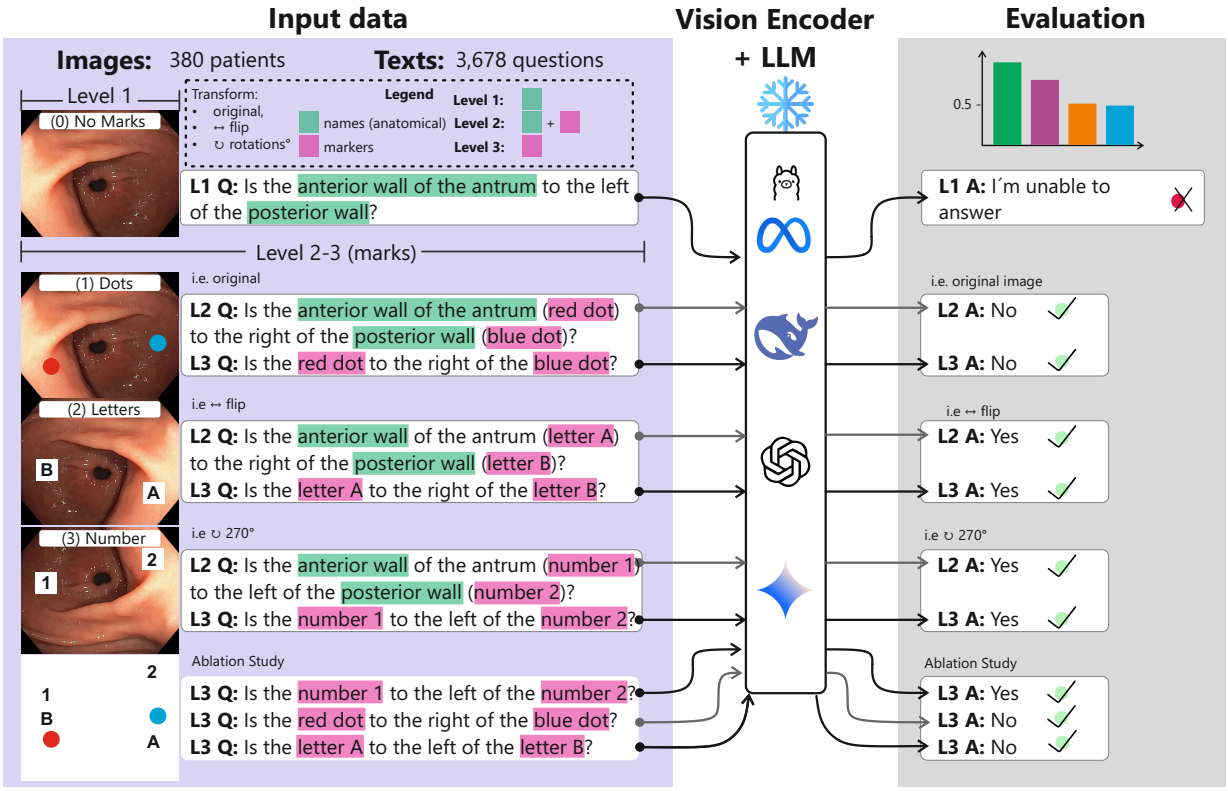

Fig. 1. Evaluation pipeline for visual spatial reasoning in vision-language models (VLMs). (I) **Input:** Each sample consists of an endoscopic image and a spatial question asking about the relative position of gastric surfaces (left/right or above/below), expressed using one of three prompt types: L1 (anatomical terms only), L2 (anatomical terms with visual markers), and L3 (visual markers only). (II) **VLM Processing:** The image-question pair is processed by a vision-language model. (III) **Evaluation:** The model's response ('Yes', 'No', or failure) is compared against the ground truth across original and transformed images (flipped/rotated) to assess accuracy and detect anatomical bias.

2. MATERIALS AND METHODS

2.1. Public Endoscopy Dataset

We use the publicly available *GastroHUN* dataset [8], which contains high-resolution upper gastrointestinal (UGI) endoscopic images from 387 patients, acquired across 22 sites following the Systematic Screening Protocol for the Stomach (SSS) [9]. For benchmarking, we included only images that met the following criteria: (i) A consensus SSS site label assigned by four expert gastroenterologists. (ii) New spatial annotations of gastric surfaces (anterior and posterior walls, lesser and greater curvatures), manually performed by an expert endoscopist. For each surface, centroid coordinates were computed to support relative position analysis. (iii) Compliance with the centroid separation rule (left/right or above/below thresholds, see Section 2.3) and valid transformation pairing (flipped or rotated counterparts, see Section 2.6). After filtering, the benchmark set includes 380 patients, 2,796 images, and 3,678 visual question answering (VQA) items. All VQA items, images, and centroids are openly available. No additional ethics approval was required.

2.2. Pre-processing

Original images (1350×1080 and 900×720) are resized to 512×512 using aspect ratio–preserving letterboxing. The same affine transformation (scaling and padding) is applied to the gastric wall centroids to maintain spatial alignment. To evaluate orientation robustness, horizontal flips and random rotations (90°, 180°, 270°) are applied. All transformations are consistently applied to both the image and the corresponding centroid coordinates to preserve geometric consistency and simulate variations in viewing orientation.

2.3. Question Generation

We generate VQA items based on the relative position of gastric surfaces in SSS landmarks, focusing on two anatomical pairs: (i) gastric walls (anterior vs. posterior) and (ii) gastric curvatures (lesser vs. greater). Using image coordinates with the origin at the top-left corner, let (x_A, y_A) and (x_B, y_B) denote the centroids of gastric surfaces A and B. A question is considered valid if $|x_A - x_B| \geq \tau_x$ (left/right) or $|y_A - y_B| \geq \tau_y$ (above/below), with $\tau_x = \tau_y = 50$ px. This ≈10% threshold relative to image size ensures that spatial relationships re-

main visually unambiguous for expert verification. Excluding adjacent surface pairs and those with centroid separation below τ ensures unequivocal visual evidence, allowing the benchmark to isolate model failures driven by anatomical bias over geometric uncertainty. Ground truth labels are assigned as follows: A is to the left of B if $x_A < x_B - \tau_x$, to the right if $x_A > x_B + \tau_x$, above if $y_A < y_B - \tau_y$, and below if $y_A > y_B + \tau_y$.

2.4. VLM Evaluation Protocol

We evaluate four VLMs (three runs each): GPT-4o (OpenAI, `gpt-4o-2024-11-20`) [10], Gemini-2.5-Flash (Google DeepMind) [11] accessed via API; JanusPro-7B (DeepSeek) [12], and Llama-3.2-11B-Vision-Instruct (Meta) [13, 14] via Hugging Face. All imagess are resized to 512×512 pixels. Decoding parameters are fixed: temperature = 0, top-p = 1 (no nucleus sampling), and maximum output length = 64 tokens. Each item is prompted with: (i) fixed endoscopy/SSS context instructing the model to rely on visual evidence (ignore anatomical priors), (ii) a one-shot example illustrating the expected binary output format (`0` for False, `1` for True), and (iii) one of three prompt formats (see Figure 1):

- **Level 1 (anatomical names only).** "Is the anterior wall to the left of the posterior wall?"
- **Level 2 (anatomical + markers).** "Is the anterior wall (red) to the left of the posterior wall (blue)?"
- **Level 3 (markers only).** "Is the red dot to the left of the blue dot?"

Model outputs are mapped to binary answers: {False: 0, True: 1}. Markers follow fixed settings: colored dots with radius $r = 14$ px (red for A, blue for B). Letter and number labels are displayed in black text within white square boxes sized 28×28 px.

2.5. Ablation Study (AS)

To test spatial reasoning independent of endoscopy-specific visual features, we run Level 3 prompts on "phantom" images, white canvases with only visual markers (dots, letters and numbers), with no endoscopic content. This isolates marker-based reasoning by removing anatomical context.

2.6. Pairwise Construction

Each VQA item is defined as (I, q, y), where I is the image, q the spatial question (e.g., left/right or above/below), and y the binary ground-truth answer (0 or 1). For each item, we generate a transformed counterpart (I', q', y') by applying a transformation $T \in \{\text{HF}, 90°, 180°, 270°\}$ to both the image and its centroids. Depending on T, the queried axis may invert (e.g., horizontal flip (HF) affects left/right) or switch (e.g., 90° and 270° rotate left/right into above/below, and vice versa). The ground-truth label y' is recomputed from the transformed centroids using the same spatial rule.

2.7. Evaluation Metrics

The primary metric was accuracy, defined as the percentage of model responses ('1' or '0') matching the ground truth derived from spatial annotations. To assess anatomical bias, we compared accuracy across three prompt levels on original, rotated, and flipped images to evaluate the impact of view transformations on model performance.

3. EVALUATION AND RESULTS

3.1. Experimental Setup

We evaluated each VLM using the EndoSSS-RP benchmark under three image conditions: original, randomly rotated (90°, 180°, 270°), and horizontally flipped. Prompts followed the formats described in Section 2.4, instructing the model to rely solely on visual evidence and respond with a binary output: '1' (Yes) or '0' (No). Performance was measured by accuracy, computed as the proportion of correct responses relative to ground truth labels derived from annotated gastric surface centroids.

	Method	Level 1	Level 2	Level 3	AS
Original	Gemini-2.5	66.78 ±0.06	97.46$^{\odot}$ ±0.06	97.71$^{\square}$ ±0.04	99.61$^{\odot}$ ±0.01
	GPT-4o	68.08 ±0.32	76.39$^{\odot}$ ±0.59	86.89$^{\diamond}$ ±0.2	97.43$^{\diamond}$ ±0.36
	GPT-4o Q: right/left	58.66 ±0.33	66.96$^{\odot}$ ±1.15	86.00$^{\diamond}$ ±0.52	97.35$^{\diamond}$ ±0.32
	JanusPro-7B	48.28 ±0.27	43.18$^{\odot}$ ±1.06	52.02$^{\odot}$ ±0.69	67.25$^{\odot}$ ±0.48
	LLaMA3.2	51.02 ±0.36	52.97$^{\odot}$ ±1.64	53.63$^{\odot}$ ± 0.70	55.22$^{\odot}$ ±0.07
Flip	Gemini-2.5	58.04 ±0.09	97.71$^{\odot}$ ±0.09	98.06$^{\odot}$ ±0.04	- -
	GPT-4o	61.48 ±0.01	68.06$^{\odot}$ ±0.26	87.43$^{\diamond}$ ±0.32	- -
	GPT-4o Q: right/left	47.25 ±0.09	51.61$^{\odot}$ ±0.43	87.13$^{\diamond}$ ±0.03	- -
	JanusPro-7B	48.30 ±1.14	41.78$^{\odot}$ ±1.12	54.86$^{\odot}$ ±0.45	- -
	LLaMA3.2	50.32 ±0.39	50.55$^{\odot}$ ±1.56	50.67$^{\square}$ ± 0.74	- -
Rotation	Gemini-2.5	42.15 ±0.03	97.30$^{\odot}$ ±0.04	97.75$^{\odot}$ ±0.10	- -
	GPT-4o	45.80 ±0.15	54.50$^{\odot}$ ±0.39	87.77$^{\odot}$ ±0.21	- -
	GPT-4o Q: right/left	47.18 ±0.03	53.20$^{\odot}$ ±0.19	86.39$^{\diamond}$ ±0.03	- -
	JanusPro-7B	46.33 ±1.28	41.77$^{\odot}$ ±0.57	52.44$^{\odot}$ ±0.68	- -
	LLaMA3.2	50.35 ±0.69	51.54$^{\odot}$ ±0.29	52.02$^{\odot}$ ± 0.08	- -

Table 1. Accuracy comparison across prompt levels and VLMs. Symbols indicate the best-performing marker type for each setting: $^{\odot}$ (*Dot*), $^{\diamond}$ (*Number*), and $^{\square}$ (*Letter*) in Levels 2–3 and the Ablation Study (AS, see Section 2.5). "Q: right/left" indicates the variant using binary right/left questions, shown for the best-performance comparison in Level 1.

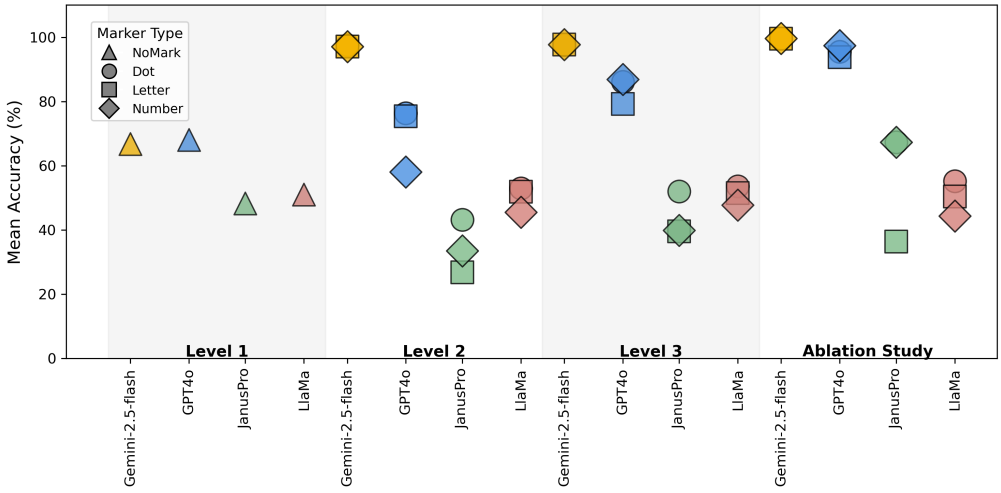

Fig. 2. Mean accuracy across VLMs for each prompt level (Levels 1–3) using original endoscopic images, and for the ablation study (AS) using marker-only phantom images without endoscopic content.

3.2. Quantitative Results: Impact of Anatomical Bias

We evaluated VLM performance across three prompt levels, ranging from anatomical terms (Level 1) to visual markers only (Level 3). Figure 2 summarizes the results on non-transformed images and illustrates how prompt design influences spatial reasoning. On original images, Level 1 accuracy hovered near chance (≈58% across models), indicating that spatial reasoning is unreliable when models rely solely on anatomical prompts. Performance improved at Levels 2 and 3, particularly for Gemini-2.5-Flash and GPT-4o, which achieved the highest accuracy with dot and number markers. The ablation study (marker-only inputs on blank backgrounds) further confirmed that Gemini-2.5-Flash and GPT-4o can reason spatially without anatomical context.

Observation 1 – Anatomical Bias (Level 1): GPT-4o achieved 68.08% on original images, but performance declined under transformations (61.48% flip, 45.80% rotation; see Table 1), revealing a reliance on anatomical priors rather than visual evidence. Accuracy near or below the random baseline (50%) on rotated views suggests poor generalization when orientation changes invalidate text-based assumptions.

Observation 2 – Effect of Visual Markers (Levels 2–3 and AS): Introducing visual markers (Level 2) modestly improved accuracy under transformed images, with the exception of Gemini-2.5-Flash, which maintained performance comparable to the original-image setting. Eliminating anatomical terms (Level 3 and AS) improved performance for both Gemini-2.5-Flash and GPT-4o, indicating that marker-only prompts strengthen spatial reasoning. The "left/right" accuracy lagged behind "above/below," reflecting that distinct patterns—such as the rugae along the greater curvature—provide clearer landmarks than the gastric walls.

Observation 3 – Limitations of Open-Source VLMs: JanusPro and LLaMA 3.2 showed no meaningful improvement across prompt levels, failing to exceed 56% accuracy. Clear visual evidence indicates that while anatomical priors limit top-performing models, open-source models are hindered by poor visual grounding. These markers serve as a diagnostic tool, not a clinical solution

4. CONCLUSION AND FUTURE WORK

Current AI applications in gastroenterology mostly target isolated visual tasks, but the long-term goal is integrated multimodal assistance. A prerequisite for this vision yet rarely tested is reliable spatial reasoning. We introduced EndoSSS-RP, a benchmark designed to assess spatial consistency through paired geometric transformations. By testing three prompt levels, we isolate visual grounding from reliance on anatomical priors. Results show that anatomical terms degrade performance under geometric transformations, whereas marker-only prompts reveal a more robust underlying spatial reasoning capacity. This suggests that foundational anatomical priors override visual evidence, a phenomenon recently identified across medical imaging domains [7, 15]. Crucially, while these markers are a diagnostic tool for model analysis rather than a clinical solution, our findings underscore the need for models that prioritize visual evidence over memorized cues. Establishing such benchmarks should become a minimum standard for VLM-based navigation support within SSS protocols. **Future work:** (i) Extend to spatio-temporal video evaluation (tracking and temporal consistency), (ii) explore mitigation strategies (orientation-aware training, transform-consistency losses, marker-grounded supervision), (iii) fine-tune VLMs with endoscopy-specific vision and language data, (iv) assess clinical utility in prospective studies (time-to-localization, blind-spot rate) while adhering to WHO ethics guidance [16].

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [8]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENT

This work was partially supported by project with code 110192092354 and entitled "Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño" of call No. 920 of 2022 of MinCiencias.

7. REFERENCES

[1] Kenshi Yao, "The endoscopic diagnosis of early gastric cancer," *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 26, no. 1, pp. 11, 2013.

[2] Chisato Hamashima, Kazuei Ogoshi, Rintarou Narisawa, Tomoki Kishi, Toshiyuki Kato, Kazutaka Fujita, Masatoshi Sano, and Satoshi Tsukioka, "Impact of endoscopic screening on mortality reduction from gastric cancer," *World journal of gastroenterology: WJG*, vol. 21, no. 8, pp. 2460, 2015.

[3] Seung-Hwa Lee, Young-Kyu Park, Sung-Min Cho, Joon-Koo Kang, and Duck-Joo Lee, "Technical skills and training of upper gastrointestinal endoscopy for new beginners," *World Journal of Gastroenterology: WJG*, vol. 21, no. 3, pp. 759, 2015.

[4] Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James E East, Xin Lu, and Jens Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *Medical image analysis*, vol. 68, pp. 101900, 2021.

[5] Shyam Menon and Nigel Trudgill, "How commonly is upper gastrointestinal cancer missed at endoscopy? a meta-analysis," *Endoscopy international open*, vol. 2, no. 02, pp. E46–E50, 2014.

[6] Mitsuru Kaise, "Advanced endoscopic imaging for early gastric cancer," *Best Practice & Research Clinical Gastroenterology*, vol. 29, no. 4, pp. 575–587, 2015.

[7] Daniel Wolf, Heiko Hillenhagen, Billurvan Taskin, Alex Bäuerle, Meinrad Beer, Michael Götz, and Timo Ropinski, "Your other left! vision-language models fail to identify relative positions in medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 691–701.

[8] Diego Bravo, Juan Frias, Felipe Vera, Juan Trejos, Carlos Martínez, Martín Gómez, Fabio González, and Eduardo Romero, "Gastrohun an endoscopy dataset of complete systematic screening protocol for the stomach," *Scientific Data*, vol. 12, no. 1, pp. 102, 2025.

[9] Diego Bravo, Martín Gómez, Juan Frías, Carlos Martínez, Felipe Vera Polanía, Fabio A. González, Eduardo Romero, and Juan Naranjo, "GastroHun: an endoscopy dataset of complete systematic screening protocol for the stomach," Figshare. `https://doi.org/10.6084/m9.figshare.27308133`, 2025.

[10] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[12] DeepSeek-AI, "Janus-pro-7b," `https://huggingface.co/deepseek-ai/Janus-Pro-7B`, 2025, Accessed: 2025-10-01.

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al., "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[14] Meta, "Llama-3.2-11b-vision-instruct," `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct`, 2024, Accessed: 2025-10-01.

[15] Leon Mayer, Piotr Kalinowski, Caroline Ebersbach, Marcel Knopp, Tim Rädsch, Evangelia Christodoulou, Annika Reinke, Fiona R Kolbinger, and Lena Maier-Hein, "6 fingers, 1 kidney: Natural adversarial medical images reveal critical weaknesses of vision-language models," *arXiv preprint arXiv:2512.04238*, 2025.

[16] World Health Organization, *Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance*, World Health Organization, 2024.