Master's Thesis

HandVQA: Diagnosing Fine-Grained Spatial Reasoning Failures in Vision-Language Models via Hand Pose Question Answering

MD Khalrquzzaman Chowdhury Sayem

Computer Science and Engineering

Ulsan National Institute of Science and Technology

2025

HandVQA: Diagnosing Fine-Grained Spatial Reasoning Failures in Vision-Language Models via Hand Pose Question Answering

MD Khlalequzzaman Chowdhury Sayem

Supervised by:

Professor Seungryul Baek
Associate Professor, Artificial Intelligence Graduate School (AIGS)
Department of Computer Science and Engineering (CSE)
Ulsan National Institute of Science and Technology (UNIST), South Korea

Professor Binod Bhattarai
Lecturer, University of Aberdeen, Aberdeen, UK
Honorary Lecturer, University College London (UCL), UK

Computer Science and Engineering
Ulsan National Institute of Science and Technology

Abstract

Understanding the nuanced articulation of human hands is essential for high-stakes applications such as robot-assisted surgery, chip manufacturing, and human-AI interaction in AR/VR. Despite achieving near-human performance on general vision-language benchmarks, current vision-language models (VLMs) struggle with fine-grained spatial reasoning—especially in interpreting complex, articulated hand poses. We introduce HandVQA, a large-scale diagnostic benchmark designed to evaluate VLMs' understanding of detailed hand anatomy through visual question answering. Built upon high-quality 3D hand datasets (FreiHAND, InterHand2.6M, FPHA), our benchmark includes over 1.6M controlled multiple-choice questions that probe spatial relationships between hand joints, such as angles, distances, and relative positions. We evaluate several state-of-the-art VLMs (LLaVA, DeepSeek, Qwen-VL, mPLUG) in both base and fine-tuned settings, using lightweight fine-tuning via LoRA. Our findings reveal systematic limitations in current models, including hallucinated finger parts, incorrect geometric interpretations, and poor generalization. HandVQA not only exposes critical reasoning gaps but also offers a concrete path toward improving spatial grounding in multimodal language models.

Contents

I Introduction 1

II Related Works 4

2.1 Hallucinations and Alignment Failures in VLMs 4

2.2 Spatial and Relational Reasoning in Multimodal Models 4

2.3 VQA Benchmarks and Fine-Grained Reasoning Evaluation 5

2.4 Applications Requiring Precise Hand Pose Understanding 5

III Pipeline to generate HandVQA benchmark 7

3.1 Pose Descriptor Extraction 7

3.2 Complete template sentences 7

3.3 Form VQA pairs 8

IV Further Dataset Construction Details 8

4.1 Input to the HandVQA benchmark generation pipeline. 8

4.2 Calculating Pose Descriptors 9

4.3 Why cases with category label “aligned” in relative position are removed. 10

V Experiments 12

5.1 Dataset Construction 12

5.2 Dataset Statistics 12

5.3 Models, Evaluation Metrics and Training Setup 14

5.4 Results and Analysis 15

VI Further Analysis on Experiments 16

6.1 Systematic Preference for Ambiguous Angle Labels Across VLMs 17

6.2 Over-Reliance on "Close To" Reveals Weak Distance Reasoning 18

6.3 Near-Random Performance in Relative Position Prediction 19

VII Qualitative Results 20

VIII Conclusion 21

References 28

List of Figures

1 **Overview of HandVQA Question Format.** This figure illustrates the structure of our benchmark, which divides hand pose estimation into five sub-tasks: Angle, Distance, and Relative Position along X, Y, and Z axes. A hand image with annotated joint indices (top left) supports multiple-choice questions per task type, derived from 3D joint coordinates. Correct answers are shown in **green**, with instructions limiting reasoning to provided options—enabling precise evaluation of VLMs' joint-level spatial understanding. 1

2 **Overview of the pipeline to generate HandVQA benchmark:** The pipeline takes normalized ground-truth 3D hand keypoints, calculates pose descriptor values, and categorizes them. 6

3 The map of the hand skeleton used in our HandVQA benchmark generation pipeline. . . 10

4 Possible location of the 'aligned' Little Finger Proximal Interphalangeal (PIP) joint and Ring Finger Proximal Interphalangeal (PIP) joint underneath the index and middle finger. The relationship along the x-axis for the two PIP joints is ambiguous, making it necessary to drop the relative position X information of the two joints. 10

5 **Word cloud representation of the most frequently used terms in the caption options extracted from our dataset.** Prominent anatomical terms like "tip joint", "interphalangeal joint", and "metacarpophalangeal joint" highlight the fine-grained spatial and anatomical focus of the hand-centric question-answer pairs. 13

6 **Breakdown of question types across the training (left) and evaluation (right) splits for each dataset in the HandVQA benchmark.** Each dataset contains a balanced distribution of all five spatial reasoning tasks—angle, distance, and relative positions along the X, Y, and Z axes. This uniformity supports fair evaluation across all pose subtasks. 14

7 **Angle confusion matrix across four VLMs.** All models frequently predict *"bent slightly inward"* regardless of ground truth, revealing a strong prediction bias. Each model also follows a consistent preference ordering in its outputs, indicating difficulty in distinguishing fine-grained joint angles. 17

8 **Distance confusion matrix across four VLMs.** LLaVA, mPLUG-Owl, and Qwen-VL show a strong bias toward predicting *"close to"*, regardless of the ground truth. In contrast, DeepSeek produces a more balanced distribution, indicating better spatial discrimination. 17

9 **Relative Position X Confusion Matrix.** Most models show near-symmetric predictions with weak spatial grounding. Notably, mPLUG-Owl exhibits the strongest directional bias, overpredicting "left of" regardless of the ground truth. 18

10 **Relative Position Y Confusion Matrix.** LLaVA exhibits a strong bias toward predicting "above", regardless of the actual label, indicating poor vertical discrimination. 18

11 **Relative Position Z Confusion Matrix.** LLaVA and DeepSeek show a strong bias toward predicting "in front of," often misclassifying "behind" instances. 19

12 **Qualitative Comparison on FreiHAND. Examples comparing LLaVA (base) and LLaVA fine-tuned on FreiHAND.** Each row shows a question about hand pose from our proposed HandVQA benchmark on an image, with multiple-choice answers. While the base model frequently selects incorrect or spatially inconsistent options, the fine-tuned version consistently predicts the correct answers, demonstrating improved spatial reasoning and alignment with hand joint relationships. 24

13 **Qualitative Comparison on InterHand2.6M. Examples comparing LLaVA (base) and LLaVA fine-tuned on InterHand2.6M.** Each row shows a question about hand pose from our proposed HandVQA benchmark on an image, with multiple-choice answers. While the base model frequently selects incorrect or spatially inconsistent options, the fine-tuned version consistently predicts the correct answers, demonstrating improved spatial reasoning and alignment with hand joint relationships 25

14 **Qualitative Comparison on FPHA. Examples comparing LLaVA (base) and LLaVA fine-tuned on FPHA.** Each row shows a question about hand pose from our proposed HandVQA benchmark on an image, with multiple-choice answers. While the base model frequently selects incorrect or spatially inconsistent options, the fine-tuned version consistently predicts the correct answers, demonstrating improved spatial reasoning and alignment with hand joint relationships. 26

15 **Qualitative Results on In-the-Wild Images.** We evaluate spatial reasoning on challenging questions using in-the-wild images. The fine-tuned LLaVA outperforms the base model on tasks involving occlusion, depth, and inter-finger relationships, demonstrating improved generalization beyond the training data. 27

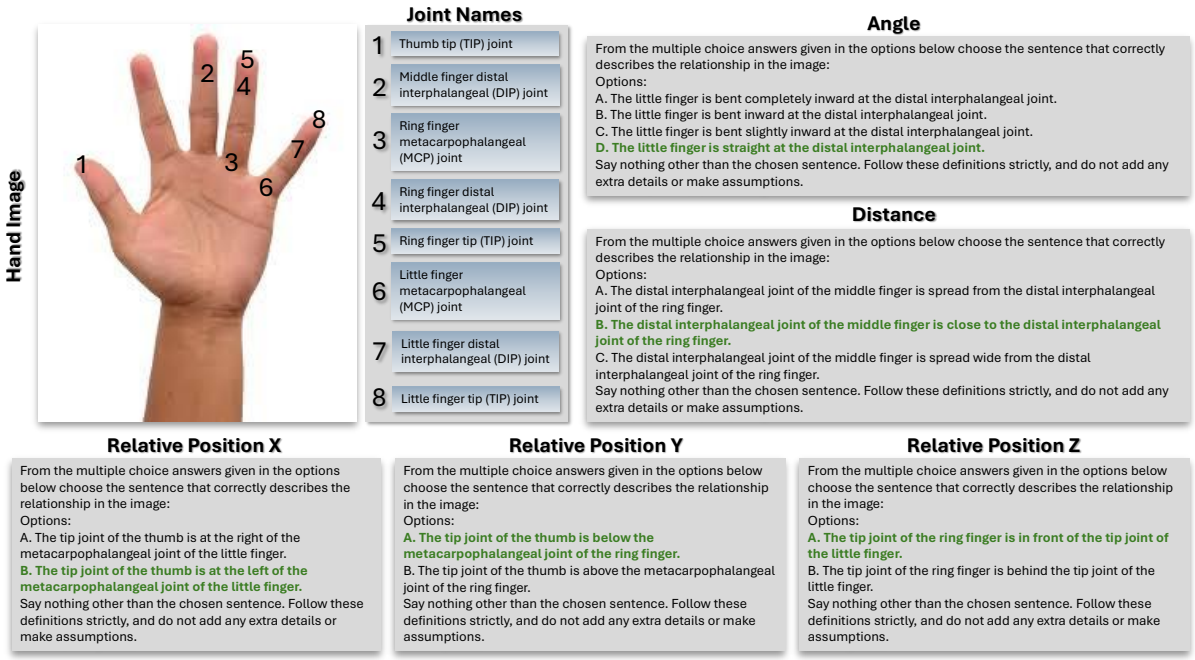


Figure 1: **Overview of HandVQA Question Format.** This figure illustrates the structure of our benchmark, which divides hand pose estimation into five sub-tasks: Angle, Distance, and Relative Position along X, Y, and Z axes. A hand image with annotated joint indices (top left) supports multiple-choice questions per task type, derived from 3D joint coordinates. Correct answers are shown in **green**, with instructions limiting reasoning to provided options—enabling precise evaluation of VLMs' joint-level spatial understanding.

I Introduction

In high-stakes applications such as robot-assisted surgery or precision chip manufacturing, accurately interpreting subtle hand gestures—such as distinguishing whether a finger is slightly bent or fully extended—is essential. Even minor misunderstandings in these contexts can lead to severe consequences, ranging from medical errors to costly production defects. Vision-language models (VLMs), increasingly integral as perceptual and reasoning components in robotics, augmented and virtual reality (AR/VR), and multimodal assistants [4, 31, 42], must reliably interpret detailed hand poses. Hands serve as a primary means of communication and control [10], and misinterpretations of their configurations can cause critical semantic or functional failures.

Despite significant advancements in general visual question answering (VQA), where state-of-the-art VLMs often achieve near-human accuracy on tasks like VQAv2 [2], they notably falter on tasks involving detailed spatial reasoning [49]. Recent benchmarks highlight their particular weaknesses with basic spatial distinctions like left versus right, achieving only around 56% accuracy compared to human performance at 99% [21]. These shortcomings point to a reliance on superficial correlations rather than genuine geometric comprehension. Articulated hand poses, characterized by complex spatial relationships across 21 joints, remain a significant challenge.

To systematically address this challenge, we introduce **HandVQA**, a diagnostic benchmark specifically designed to evaluate VLMs' fine-grained understanding of hand pose and anatomy through targeted visual question answering. HandVQA is constructed using precise 3D annotations from widely-used datasets—FreiHAND [52], InterHand2.6M [33], and FPHA [14]. We disentangle the hand pose estimation task into five separate subtasks: joint angles (measured at a single joint) and distances and relative positions along the X, Y, and Z axes (measured between pairs of joints) (see Fig. 1). While we draw conceptual motivation from Delmas et al. [11], our formulation is tailored specifically to the spatial nuances of hand anatomy. Our pipeline generates controlled multiple-choice questions probing these specific joint angles, distances, and relative positions, thereby minimizing ambiguity and encouraging genuine spatial reasoning (e.g., *"Is the distal interphalangeal joint of the middle finger closer to that of the ring finger or the index finger?"*). By mapping 3D hand joint relations into structured natural language, HandVQA enables a diagnostic view into how well VLMs grasp spatial pose concepts.

We evaluate several open-source large VLMs, including LLaVA [26], Qwen-VL [3], mPLUG [47], and DeepSeek [30]. These models are fine-tuned using parameter-efficient LoRA adapters [15] and evaluated on held-out samples from the same datasets used for training. Our findings indicate significant performance gaps: base models generally perform poorly, often similar to or below random guess, particularly on distance-related questions, reflecting a fundamental lack of spatial grounding. While fine-tuning markedly improves performance—confirming that VLMs can acquire spatial awareness with sufficient data—significant limitations persist, especially regarding the intricate task of accurately interpreting joint angles.

Our detailed analysis reveals key insights: VLMs often settle for simplified answers (e.g., repeatedly predicting "close" for distances or "slightly bent" for angles) indicating superficial alignment rather than true understanding. Furthermore, we observe that performance superiority in one spatial reasoning task rarely generalizes across others, highlighting that current VLM architectures might require stronger vision encoders or full-model fine-tuning to robustly capture intricate spatial relationships.

The implications of our study extend beyond benchmarking. Accurate hand pose interpretation is crucial in robotics for safe human-robot interactions, in AR/VR for immersive user experiences, and in medical assistance for sterile and precise gesture-based device control [4, 12, 34, 46]. HandVQA addresses a critical gap in existing benchmarks by specifically focusing on granular spatial reasoning, thereby contributing directly to improving VLM spatial reasoning capabilities broadly.

Our primary contributions include:

- **A novel VQA benchmark (HandVQA)**: Consisting of more than 1.6 million controlled, anatomically grounded questions about hand joint relationships, this is the first benchmark dedicated explicitly to hand pose interpretation.
- **Automated question generation pipeline**: Enabling comprehensive evaluation of fine-grained spatial reasoning through systematic and unbiased queries involving joint angles, distances, and positions.

- **Detailed evaluation of state-of-the-art open-source VLMs**: We provide an extensive performance analysis of several leading VLMs, highlighting systematic limitations in accurately interpreting specific hand poses. Notably, DeepSeek performs comparatively better across subtasks in its base form without fine-tuning; however, after fine-tuning, LLaVA consistently achieves the best results across all subtasks.
- **Insights linking to broader VLM challenges**: Base VLMs often perform at or below the level of random guessing, especially on distance-related tasks, revealing poor spatial grounding. Fine-tuning improves performance, showing VLMs can learn spatial reasoning with sufficient data. Still, challenges persist in interpreting fine-grained joint angles, underscoring the need for future work on spatial understanding in VLMs.

II Related Works

2.1 Hallucinations and Alignment Failures in VLMs

Vision-language models (VLMs) often suffer from hallucinations—generating content not grounded in the image—such as nonexistent objects, incorrect attributes, or false spatial relations [28]. For example, a model might describe a ring that isn't there or misstate left/right positions [21, 24]. Even strong models like GPT-4V [1] and BLIP-2 [23] hallucinate objects [24], struggle with spatial terms [21], or fabricate colors and shapes [19]. These issues stem from weak visual-textual alignment and overreliance on language priors [18, 40, 45]. Liu et al. [25] emphasize the lack of grounding signals in training data. Recent solutions include evaluation frameworks like POPE [24], NOPE [29], and Hal-Eval [19], as well as mitigation strategies such as contrastive decoding, reinforcement learning [6], and robust visual encoders [27]. Still, hallucinations remain prevalent, even in top-performing models [32].

Our work exposes a distinct form of hallucination: *pose hallucination*, where models misinterpret joint-level configurations. For instance, they may infer a bent joint where it is straight, or consistently default to a "close" distance between fingers. Such errors reveal fine-grained alignment failures and highlight the need for diagnostics beyond object-level VQA.

2.2 Spatial and Relational Reasoning in Multimodal Models

Beyond identifying objects, true vision-language understanding requires reasoning about spatial and relational configurations. Kamath et al. [21] revealed that many VLMs fail even simple spatial tasks like distinguishing a dog under a table versus on a table. Similarly, Zhang et al. [50] present SPHERE, a hierarchical evaluation of spatial skills from basic (positions, distances) to complex (occlusion, physical plausibility). On SPHERE, the best model achieved only $\sim$68% accuracy—far below human performance of $\sim$93% [50]. These results show that high-level accuracy on generic VQA does not imply reliable spatial understanding [37, 41]. Researchers have begun addressing this gap. Yang et al. [44] proposed a direction on improving model training by incorporating a spatial relation graph into vision-language pretraining, improving performance on reasoning tasks like VCR [48] and NLVR2 [39]. Chen et al. [7] proposed SpatialVLM, adding estimated depth maps and 3D spatial cues during training to overcome 2D limitations. Nonetheless, evaluating spatial reasoning remains difficult—many benchmarks conflate spatial skills with priors or world knowledge. Modern models can sometimes exploit shortcuts in datasets like CLEVR [20] or GQA [16]. Thus, new diagnostics like What'sUp [21] and SPHERE [50] are crucial.

Our HandVQA benchmark contributes by focusing on spatial relations within a single object: the human hand. Unlike benchmarks evaluating inter-object relationships, we target part–whole spatial structure, requiring understanding of joint kinematics and structured geometry. Since our questions are grounded in the real 3D coordinates, we test the model's ability to grasp Euclidean spatial concepts like *distance* and *angle*.

2.3 VQA Benchmarks and Fine-Grained Reasoning Evaluation

The VQA field has progressed from basic object/attribute Q&A (e.g., VQAv1/v2) [2] to compositional and fine-grained reasoning tasks. CLEVR [20] introduced logical queries in synthetic scenes; GQA [16] used real images and scene graphs for multi-step reasoning. Recent benchmarks like A-OKVQA [36] and Encyclopedic-VQA [32] test domain knowledge, but fine-grained accuracy remains low: e.g., PaLI [8] achieves only 13% on Encyclopedic-VQA [32]. Our benchmark instead focuses on fine-grained spatial understanding. It aligns with works like NLVR2 [38] and CLEVR-3D [43] that emphasize logical consistency and physical reasoning. New tools like VERIFY [5] evaluate if models follow reasoning chains rather than shallow cues. II-MMR [22] categorizes multi-hop VQA questions and shows many "hard" questions are solved via superficial features.To maintain focus, we keep HandVQA questions multiple-choice and relatively constrained—allowing clean assessment of spatial reasoning accuracy. Future extensions could incorporate explanation-based evaluation, such as "why is this joint closer than that one?"

HandVQA fills a novel gap: previous benchmarks either test external knowledge or fine-grained classification (e.g., bird species), while we demand geometric reasoning grounded in image content. It thus adds to diagnostic tools assessing whether progress in general VQA translates to true fine-grained understanding.

2.4 Applications Requiring Precise Hand Pose Understanding

Precise hand articulation understanding is critical across domains. In robotics, accurate perception of hand gestures and manipulation is essential. Duan et al. [12] introduce AHA, a VLM that detects and explains robotic manipulation failures, though it assumes accurate hand pose input. Bao et al. [4] present HandsOnVLM, forecasting future hand trajectories in egocentric video based on language queries. However, their method depends on reliable hand pose recognition [4]. Egocentric vision also relies on hand pose to infer intent. Ego4D emphasizes hand-object interactions [42], while several works show that better 3D hand pose estimation improves activity recognition [9, 17, 34, 35]. More specifically, Ohkawa et al. [34] argue that hand pose is a compact representation of action. In augmented reality interfaces, devices like HoloLens or Meta Quest use hand tracking for gesture-based commands [34]. Misinterpreting a finger's position could be the difference between registering a "pinch" vs. a "point" gesture. Fine-grained evaluation like HandVQA can support the development of models that correctly interpret these nuances, reducing the risk of gesture misclassification (a form of hallucination in interaction). In surgical robotics or assistive tech, reliability is paramount: misclassification of a surgeon's or disabled user's gesture could have severe consequences [46].Thus, many real-world applications demand detailed, error-free hand pose understanding. HandVQA offers a way to benchmark and improve these capabilities. By forcing models to reason about joint-level spatial details, we aim to drive broader improvements in multimodal perception and grounding.

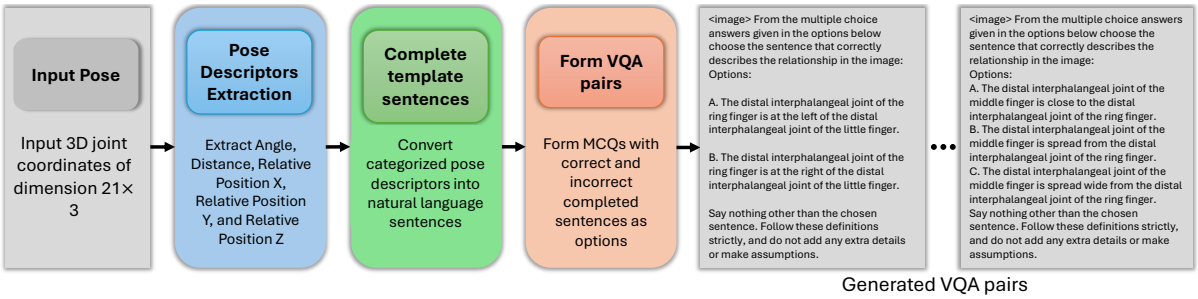
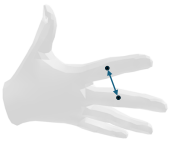
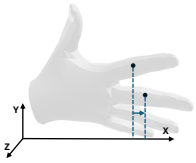
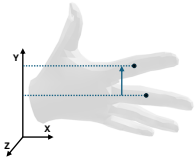
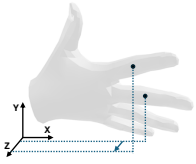


Figure 2: **Overview of the pipeline to generate HandVQA benchmark:** The pipeline takes normalized ground-truth 3D hand keypoints, calculates pose descriptor values, and categorizes them.

Table 1: Pose descriptor categorization with conditions and illustration.

Illustration	Pose Descriptor	Category Label	Condition
	**angle**	bent completely inward	$\theta < 105°$
		bent inward	$105° \leq \theta < 150°$
		bent slightly inward	$150° \leq \theta < 170°$
		straight	$\theta \geq 170°$
	**distance**	close to	$d < 0.1$
		spread from	$0.1 \leq d < 0.3$
		spread wide from	$d \geq 0.3$
	**rel. pos. (X)**	at the left of	$d_x < -0.15$
		aligned	$-0.15 \leq d_x < 0.15$
		at the right of	$d_x \geq 0.15$
	**rel. pos. (Y)**	below	$d_y < -0.15$
		aligned	$-0.15 \leq d_y < 0.15$
		above	$d_y \geq 0.15$
	**rel. pos. (Z)**	behind	$d_z < -0.15$
		aligned	$-0.15 \leq d_z < 0.15$
		in front of	$d_z \geq 0.15$

III Pipeline to generate HandVQA benchmark

In this section, the design of the automatic VQA generation pipeline is presented. The pipeline (Fig. 2) automates the VQA generation process, producing textual descriptions from normalized ground-truth 3D hand keypoints. The overall pipeline is divided into three steps: pose descriptor extraction, completing template sentences, and forming VQA pairs, which will be detailed below.

3.1 Pose Descriptor Extraction

The first step in the pipeline involves the extraction of pose descriptors, which are elementary units of information describing the relationship between body parts. Pose descriptors are derived from the 3D coordinates of keypoints in a hand pose, and they capture three main types of information:

- **Angle** pose descriptor categorizes the angle formed at a joint with two other adjacent joints into category labels, such as “bent completely inward”, “bent inward”, “bent slightly inward” or “straight”. It describes how the hand is bent at different hand joints.
- **Distance** pose descriptor measures the distances between pairs of joints and categorize them into category labels “close”, “spread” or “spread wide”.
- **Relative Position** pose descriptors describe spatial relationships between two hand joints along a given axis. Relative position along x-axis categorizes the relationship between two given joints into category labels “left of”, “aligned”, or “right of”. For y-axis, it’s “above”, “aligned”, or “below”. For z-axis, it’s “in front of,” “aligned”, or “behind”. This provides essential information on the pose’s structure. When the two joints are categorized with the label “aligned”, the 3D joint positions are deemed to be too close to each other along an axis for their relative position to be discernible from the image, i.e., part of the finger of one joint possibly occluding the other joint. Due to the ambiguity, we decided not to include samples having the “aligned” category (further details can be found in the Section IV).

The details on how to calculate each type of pose descriptor have been explained in more detail in Section IV. A map of a hand with joint names as well as the specific joints involved in each pose descriptor calculation have also been shown in Section IV. The thresholds for each category label in case of each pose descriptor type are shown in Table 1.

3.2 Complete template sentences

Finally, the categorized pose descriptors are converted into natural language sentences. Pre-determined sentence templates are filled in with joint and category information to form a complete sentence. For example, consider the following template for sentence involving the distance pose descriptor, *“The {joint A} joint of the {finger A} is {category label} the {joint B} joint of the {finger B}”*. Given that the distance category label is “close to" between ring finger distal interphalangeal joint and middle finger

distal interphalangeal joint, the template sentence looks like the following, *"The distal interphalangeal joint of the middle finger is close to the distal interphalangeal joint of the ring finger"*.

3.3 Form VQA pairs

For a particular pose descriptor, complete sentences involving the same joint(s) are made for every possible category label of the pose descriptor, such that only one sentence, represented by the true category label describes the true attribute/relationship of the joint/joints; while other sentences describe false attribute/relationship. The completed sentences are then arranged into a multiple choice question (MCQ) with a prompt asking to select the sentence with the correct category label for the corresponding joint(s) from the MCQ. Examples of an MCQ for each pose descriptor can be seen in Fig 1. For each image, pose descriptors are calculated for various joints/joint-pairs. For each pose descriptor, five different joints/joint-pairs are randomly selected to form five different MCQs. So, a total of twenty-five MCQs are generated for each image using the five pose descriptors (angle, distance, relative position X, relative position Y, and relative position Z). Before filtering out cases of "aligned" category label from relative position pose descriptors, a total of 107 MCQs are possible (details in Section IV) from the five pose descriptors for each image. We sample five different joint/joint-pairs per pose descriptor so that the dataset size remains reasonable to train on the dataset.
Through this process, the automatic VQA pipeline can generate descriptions for a vast number of poses in a fraction of the time it would take for manual annotations.

IV Further Dataset Construction Details

In Figure 3 and Table 2 we use the following shorthand for convenience: carpometacarpal - CMC, metacarpophalangeal - MCP, interphalangeal - IP, proximal interphalangeal - PIP, distal interphalangeal - DIP.

Figure 3 shows the joint names and their positions on the hand skeleton used in our HandVQA benchmark generation pipeline. Table 2 shows the list of joints/joint-pairs on which pose descriptors are calculated in our HandVQA benchmark generation pipeline. In total, there are 107 joints/joint-pairs being considered.

4.1 Input to the HandVQA benchmark generation pipeline.

The input to the pipeline is ground-truth normalized hand joint coordinates of dimension 21×3. The joint coordinates are normalized using coordinates of 3D hand mesh vertices. The joint coordinates are normalized such that the vertex of the mesh with maximum coordinate value along an axis is assigned a value of 1 and the vertex with minimum coordinate value along an axis is assigned a value of 0. Proportionate to this, the joints in between are assigned values between 0 and 1. Since FPHA [14] does not include hand mesh vertices or pose information from which a hand mesh can be extracted, the

Table 2: List of joints/joint-pairs on which angles, distances, and relative positions in X,Y, and Z axes are calculated. A total of 107 different joints/joint-pairs across all pose descriptors are considered.

Angle Pose Descriptors	**Distance Pose Descriptors**	**Relative Position Pose Descriptors (XYZ)**
Thumb-MCP	Thumb-MCP *vs.* Index-PIP	Thumb-MCP *vs.* Index-PIP
Index-PIP	Index-PIP *vs.* Middle-PIP	Index-PIP *vs.* Middle-PIP
Middle-PIP	Middle-PIP *vs.* Ring-PIP	Middle-PIP *vs.* Ring-PIP
Ring-PIP	Ring-PIP *vs.* Little-PIP	Ring-PIP *vs.* Little-PIP
Little-PIP	Thumb-Tip *vs.* Index-Tip	Thumb-Tip *vs.* Index-Tip
Thumb-IP	Index-Tip *vs.* Middle-Tip	Index-Tip *vs.* Middle-Tip
Index-DIP	Middle-Tip *vs.* Ring-Tip	Middle-Tip *vs.* Ring-Tip
Middle-DIP	Ring-Tip *vs.* Little-Tip	Ring-Tip *vs.* Little-Tip
Ring-DIP	Thumb-Tip *vs.* Index-DIP	Thumb-Tip *vs.* Index-DIP
Little-DIP	Thumb-Tip *vs.* Middle-DIP	Thumb-Tip *vs.* Middle-DIP
Little-MCP	Thumb-Tip *vs.* Ring-DIP	Thumb-Tip *vs.* Ring-DIP
Ring-MCP	Thumb-Tip *vs.* Little-DIP	Thumb-Tip *vs.* Little-DIP
Middle-MCP	Thumb-Tip *vs.* Index-MCP	Thumb-Tip *vs.* Index-MCP
Index-MCP	Thumb-Tip *vs.* Middle-MCP	Thumb-Tip *vs.* Middle-MCP
Thumb-CMC	Thumb-Tip *vs.* Ring-MCP	Thumb-Tip *vs.* Ring-MCP
	Thumb-Tip *vs.* Little-MCP	Thumb-Tip *vs.* Little-MCP
	Index-MCP *vs.* Index-DIP	Index-MCP *vs.* Index-DIP
	Index-DIP *vs.* Middle-DIP	Index-DIP *vs.* Middle-DIP
	Middle-DIP *vs.* Ring-DIP	Middle-DIP *vs.* Ring-DIP
	Ring-DIP *vs.* Little-DIP	Ring-DIP *vs.* Little-DIP
	Thumb-Tip *vs.* Middle-Tip	Thumb-Tip *vs.* Middle-Tip
	Middle-Tip *vs.* Little-Tip	Middle-Tip *vs.* Little-Tip
	Index-Tip *vs.* Ring-Tip	Index-Tip *vs.* Ring-Tip

normalization is performed using maximum and minimum coordinate values of joints instead of mesh vertices.

4.2 Calculating Pose Descriptors

Next, we discuss how the pose descriptors are calculated.

Figure 3: The map of the hand skeleton used in our HandVQA benchmark generation pipeline.

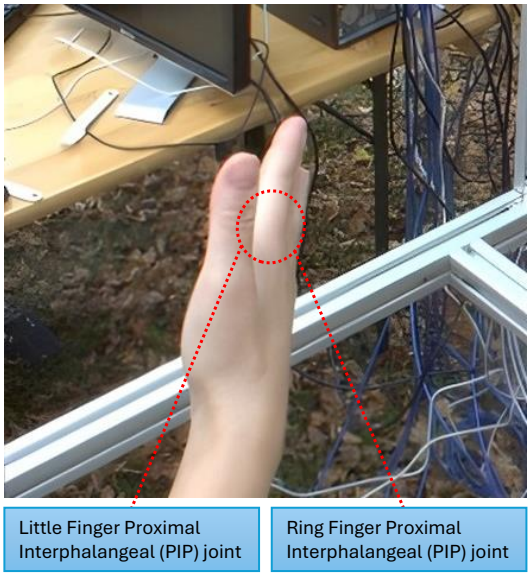


Figure 4: Possible location of the 'aligned' Little Finger Proximal Interphalangeal (PIP) joint and Ring Finger Proximal Interphalangeal (PIP) joint underneath the index and middle finger. The relationship along the x-axis for the two PIP joints is ambiguous, making it necessary to drop the relative position X information of the two joints.

- **Angle** pose descriptor describes the extent of bend at a particular joint. Given a set of joints (a,b,c) assume joint a and joint c are adjacent to joint b. Let the coordinates of joint a be C_a. The angle at b is calculated by taking the cosine similarity between the vectors $C_a - C_b$ and $C_c - C_b$.
- **Distance** pose descriptor is computed by taking the $L2$-distance $\|C_b - C_a\|$ between two given joints a and b.
- **Relative Position** pose descriptor is calculated by taking the difference between two joint positions along a particular axis. For example, along the y-axis, if $C_a^y > C_b^y$, joint a is categorized as being "*above*" joint b. Similarly, joint a is "*in front of*" joint b if $C_a^z > C_b^z$, and "*at the right of*" if $C_a^x > C_b^x$.

4.3 Why cases with category label "aligned" in relative position are removed.

In Figure 4, while the little finger proximal interphalangeal joint (PIP) and the ring finger proximal interphalangeal joint (PIP) are occluded by the index finger and the middle finger, it can be deduced from the posture that the two PIP joints lie somewhere around the marked oval region, and it can also be deduced that the joints are close enough along the x-axis for the category label to be deemed as "aligned". While access to ground-truth joint coordinates allows us to ascertain their relative left-right

relationship and generate corresponding MCQ data, the visual cue in itself is insufficient for a VLM to determine the relative left-right relationship of the two joints. The joints being too close along the x-axis makes their relationship ambiguous making it necessary to drop the relative position X information of the two joints when creating MCQ. Figure 4 shows an example of a scenario where aligned relative position makes the relationship ambiguous to interpret. Similarly in cases of all relative position pose descriptors, visual cues from cases where two joints are too close are deemed possibly ambiguous and dropped.

V Experiments

5.1 Dataset Construction

We use three hand datasets to construct our HandVQA benchmark: FreiHAND [52], InterHand2.6M [33], and FPHA [14]. To ensure training and evaluations do not take too long, we create our own train/test splits by sampling a subset of images from the official splits of each dataset. The only exception is the FreiHAND test set, which we use in full due to its relatively small size.

FreiHAND. We construct our VQA training set using the last 30,000 images in the original training split of FreiHAND, which yields 742,575 VQA pairs consisting of all five pose descriptors. For the test set, we use the entire FreiHAND test split of size 3,960, yielding 98,261 VQA pairs consisting of all five pose descriptors.

InterHand2.6M. To construct training set, we use the 5 FPS version of the dataset and take images from the official training split of InterHand2.6M. We take images of subjects 5 to 26 in all right-hand postures, from the viewing point "cam400053" and "cam400064", yielding 132,999 VQA pairs from 5,348 images. The test split is also made up of images from the official training split of InterHand2.6M. We use images of subjects 1 to 4 in all right-hand postures with the images being from the same viewing points as our training split. This yields 97,806 VQA pairs from 3,934 images.

FPHA. The training set is constructed using all video sequences of subjects 1,2,3, and 4 performing all the actions in the dataset, yielding 374,056 VQA pairs from 15,000 randomly selected images. The test set is constructed using video sequence 1 images of subjects 5 and 6 performing all the actions in the dataset, yielding 212,336 VQA pairs from 8,511 images.

5.2 Dataset Statistics

Balanced Coverage of Spatial Reasoning Tasks

The Figure 6 presents the distribution of question types—Angle, Distance, and Relative Position (X, Y, Z axes)—across the training and evaluation splits for each dataset used in HandVQA: FPHA [14], FreiHAND [52], and InterHand2.6M [33].

In both the training (left) and evaluation (right) plots, each dataset exhibits a balanced distribution across all five question types. This uniformity ensures that no particular spatial reasoning category is over- or under-represented, facilitating fair comparison and comprehensive evaluation across models.

The proportions of each question type are consistent across all datasets, making HandVQA a well-structured benchmark for studying fine-grained multimodal understanding across diverse datasets.

Word Cloud Analysis of Pose Descriptors

Figure 5 shows a word cloud visualization constructed from the textual pose descriptors used throughout the HandVQA benchmark. This visual highlights the most frequently occurring terms across the dataset's five question types—angle, distance, and relative positions in X, Y and Z axis.

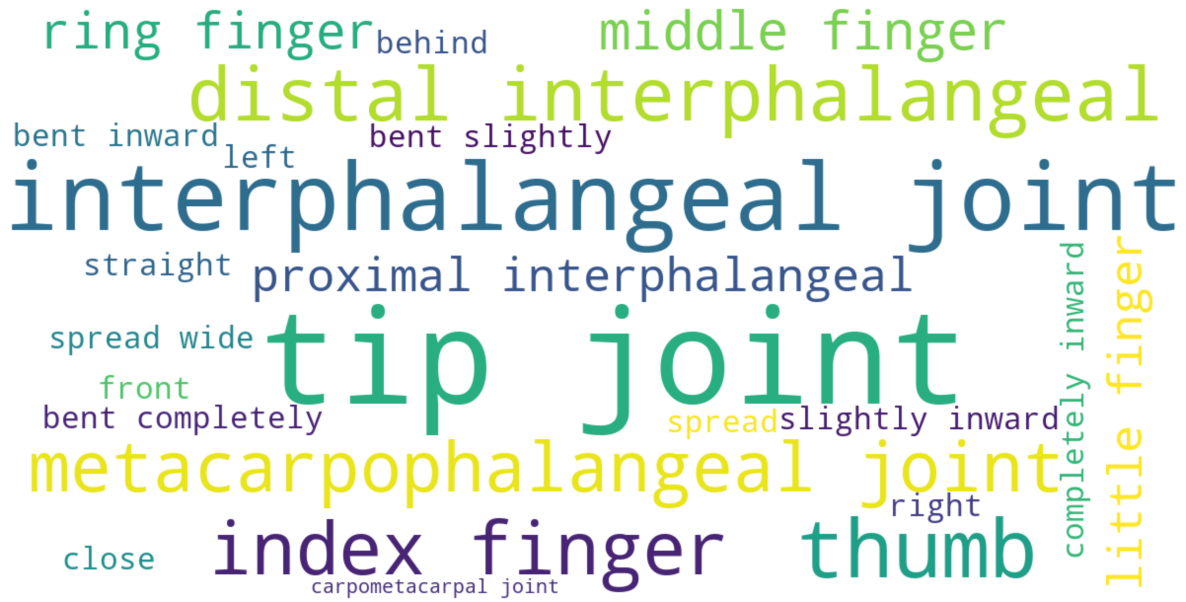


Figure 5: **Word cloud representation of the most frequently used terms in the caption options extracted from our dataset.** Prominent anatomical terms like "tip joint", "interphalangeal joint", and "metacarpophalangeal joint" highlight the fine-grained spatial and anatomical focus of the hand-centric question-answer pairs.

Many of the most prominent words (e.g., *interphalangeal joint*, *tip joint*, *metacarpophalangeal joint*, *thumb*, *index finger*) are directly tied to the anatomical joint names and relationships defined in our task design. As shown in Table 2, these joint names form the core of the five types of pose descriptions used to generate structured language annotations.

Specifically:

- **"Tip"** appears prominently because many distance and relative position comparisons involve tip joints, such as *Thumb-Tip vs. Index-Tip* or *Thumb-Tip vs. Ring-MCP*.
- **"Interphalangeal"** and its variations (e.g., proximal, distal) are common due to their presence in all pose descriptors as shown in Table 2.
- **Category label related terms** such as *bent inward*, *straight*, *spread wide*, *spread*, *left*, *behind*, and *completely inward* come from the classification vocabulary used to describe joint relationships across all five pose descriptors.
- **Finger names** like *thumb*, *index finger*, *ring finger*, and *little finger* occur frequently because they are used systematically across all pose descriptor types.

This word cloud highlights the anatomical precision and task consistency of our benchmark's language component, demonstrating that the generated textual annotations are grounded in structured, meaningful joint relationships.

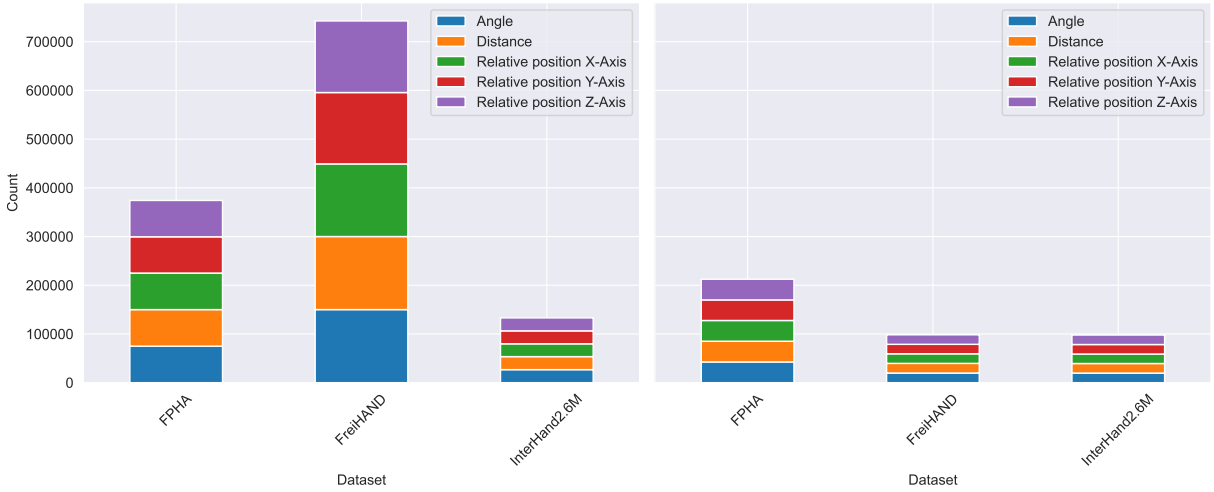


Figure 6: **Breakdown of question types across the training (left) and evaluation (right) splits for each dataset in the HandVQA benchmark.** Each dataset contains a balanced distribution of all five spatial reasoning tasks—angle, distance, and relative positions along the X, Y, and Z axes. This uniformity supports fair evaluation across all pose subtasks.

5.3 Models, Evaluation Metrics and Training Setup

Models. We evaluate four state-of-the-art 7B vision-language models—LLaVA Mistral[26], DeepSeek Janus Pro [30], Qwen 2.5 VL Instruct [3], and mPLUG-Owl 3 [47]—on the HandVQA benchmark, using both base and LoRA [15] finetuned versions. Only 7B models were evaluated in this work due to GPU resource constraints.

Evaluation Metrics. HandVQA comprises five sub-tasks derived from 3D hand joint annotations: angle, distance, and relative positions along the X, Y, and Z axes. For angle and distance, we report both accuracy and mean absolute error (MAE). While accuracy captures correct predictions, MAE reflects the average deviation from ground truth—crucial for ordinal categories where not all errors are equally severe (e.g., misclassifying "bent completely inward" joint as "straight" is worse than as "bent inward"). To compute MAE, we assign ordinal indices to each category based on increasing magnitude. For the angle task, the four categories—bent completely inward, bent inward, bent slightly inward, and straight—are mapped to class indices 0, 1, 2, and 3, respectively, reflecting increasing joint angles. For the distance task, the categories—close to, spread from, and spread wide from—are assigned indices 0, 1, and 2, corresponding to increasing joint distances. For relative position (X/Y/Z), we report accuracy only, as each is framed as a binary classification (e.g., left vs. right, below vs. above and behind vs. front). Ambiguous cases labeled "aligned" are excluded to ensure evaluation on clearly defined spatial relations.

Training Setup. We fine-tune all VLMs using LoRA[15] with rank 8 and alpha 32, targeting all linear layers. We use a learning rate of 1e-4. For all VLMs we train on FreiHAND VQA pairs for 1 epoch across 4 RTX 6000 ada GPUs and an Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz CPU with a per device batch size 2, utilizing gradient accumulation over 16 steps for all datasets, resulting in an

effective batch size of 128. We train InterHand2.6M VQA pairs for 3 epochs with a per device batch size of 1, resulting in an effective batch size of 64. In case of FPHA VQA pairs, we train for 1 epoch with a per device batch size of 1, resulting in an effective batch size of 64. All trainings are done on bfloat16 precision for speed, memory efficiency, and numerical stability. We use the SWIFT[51] package to fine-tune all VLMs.

5.4 Results and Analysis

Tables 3 and 4 summarize the experimental results. In the following analysis, we examine the performance and behavior of the evaluated VLMs on our proposed HandVQA benchmark. Further analysis of our experiments and benchmark are provided in section VI and qualitative results in Section VII.

Scarcity of data the cause for VLMs' poor performance. As per Table 3 and Table 4, the base VLMs (base model without any finetuning) perform poorly on all pose descriptors, most often having an accuracy around random choice or worse. However, after fine-tuning, there are generally massive improvements across all metrics in all datasets for all the VLMs. This proves it is possible to train VLMs on spatial awareness of hands given abundant proper training data.

VLMs struggle to grasp distance between joints. As Table 3 shows, base VLMs generally seem to perform poorly on distance pose descriptor with LLaVA, mPLUG-Owl and Qwen performing well below the accuracy of 33.3% accuracy that would have been achieved via random choice. Even the MAE remains high for all three of these base models with the lowest MAE being 1.208 for Qwen on the FreiHAND dataset. While DeepSeek achieves an accuracy of more than random choice, it still remains low with the highest being 45.55% on the InterHand2.6M dataset and the lowest MAE being a rather high 0.657 on the InterHand2.6M. The reasons for failure in case of base LLaVA, mPLUG-Owl and Qwen can be attributed to these models answering hand joints being "close" regardless of the situation. This is most severe in case of Qwen which answers "close" 93% of the time when the actual answer is "spread" and 91.3% of the time when the actual answer is "spread wide", as illustrated in the confusion matrix in Section VI. This leads to the worst performance in base Qwen model with accuracies of 1.04%, 5.69%, and 6.88%, on FPHA, FreiHAND, and InterHand2.6M, respectively, as well as highest MAEs among all base models. While base VLMs fail to grasp the concept of distance between joints of fingers, for all the models across all datasets, the performance sees a massive boost upon fine-tuning, with the lowest accuracy being 80.88% in case of Qwen fine-tuned on the FPHA dataset, with the largest jumps in performance seen in mPLUG-Owl across all three datasets.

VLMs struggle to grasp angle even after fine-tuning. According to Table 3, the performance of base VLMs across datasets excluding FPHA is generally substantially higher than the accuracy of 25% that would have been achieved via random choice, with the lowest being 34.10% for DeepSeek on the FPHA dataset. A common trend observed in the confusion matrix in Section VI across all base models is that they all choose the option involving "bent slightly inward" in most cases irrespective of the actual answer. On the FPHA dataset, performance is significantly lower across all base VLMs compared to the other two datasets in terms of both accuracy and MAE, which can be attributed to FPHA being

an egocentric dataset, indicating a bias in VLMs for allocentric viewing points. However, this trend is remedied after fine-tuning upon which FreiHAND is usually the dataset with the worst performance across all models. Unlike distance and relative position pose descriptors, where, upon fine-tuning, the accuracy generally jumps to above 80%, in case of angles, however, the accuracy of fine-tuned model is below 70% in most cases with the highest being 74.35% for LLaVA fine-tuned on InterHand2.6M. Angle at joints of hands being a more intricate feature and being more representative of the pose of the hand means freezing the vision encoder, as is the case when fine-tuning with LoRA, becomes more of a limitation than for other pose descriptors. This can be overcome with a more powerful backbone or by fine-tuning the whole model on more data instead of fine-tuning with LoRA. Similar concerns have been raised in domains of VLMs expressing human-body pose [13].

Superiority in one task does not translate to superior performance in other tasks. Among the base VLMs no model is superior to others across all tasks. In Table 3, for angle, among base models, while base LLaVA on average performs best in terms of accuracy and base mPLUG-Owl performs best in terms of MAE, base DeepSeek performs best in distance in terms of both MAE and accuracy. As can be seen in Table 4, in case of Relative Position X, Y, and Z as well no base model dominates across all pose descriptors over other base models. However, upon finetuning, LLaVA comes out to be the superior model among all fine-tuned models for all pose descriptors across all metrics in all the datasets.

Challenges in Interpreting Left/Right, Above/Below, and Front/Behind. Our results reveal that base VLMs lack a grounded understanding of fundamental spatial directions—left/right (X-axis), above/below (Y-axis), and front/behind (Z-axis). As shown in Table 4, the accuracy of all base models across datasets remains close to 50%, effectively equivalent to random guessing in these binary classification tasks. This suggests that, without targeted adaptation, VLMs are unable to consistently reason about relative positions between joints. However, after fine-tuning on hand pose data, performance improves dramatically across all spatial axes. For example, LLaVA achieves over 97% accuracy on all three axes in InterHand2.6M, and even lower-performing configurations exceed 70% accuracy. These findings confirm that while VLMs do not possess inherent spatial grounding in directional concepts, they can acquire precise spatial reasoning abilities when exposed to sufficient task-specific supervision. This highlights the importance of our HandVQA benchmark, which explicitly isolates and evaluates these fine-grained spatial relations, enabling effective diagnosis and improvement of directional understanding in VLMs.

VI Further Analysis on Experiments

Figures 7, 8, 9, 10, and 11 show confusion matrices for all four base VLMs (before fine-tuning) we consider: DeepSeek [30], LLaVA [26], Qwen-VL [3], and mPlug-Owl [47]. The confusion matrices are constructed from the evaluation sets of all three datasets combined.

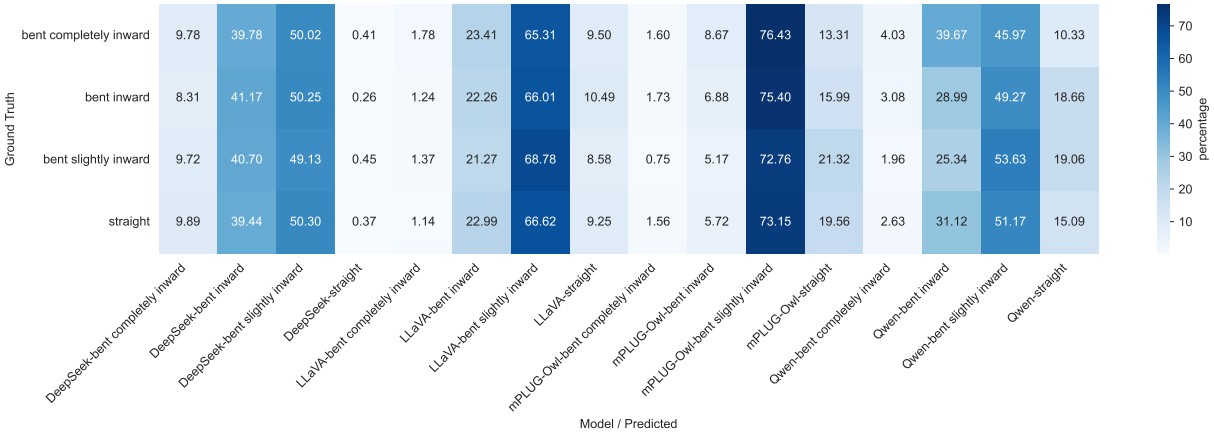


Figure 7: **Angle confusion matrix across four VLMs.** All models frequently predict *"bent slightly inward"* regardless of ground truth, revealing a strong prediction bias. Each model also follows a consistent preference ordering in its outputs, indicating difficulty in distinguishing fine-grained joint angles.

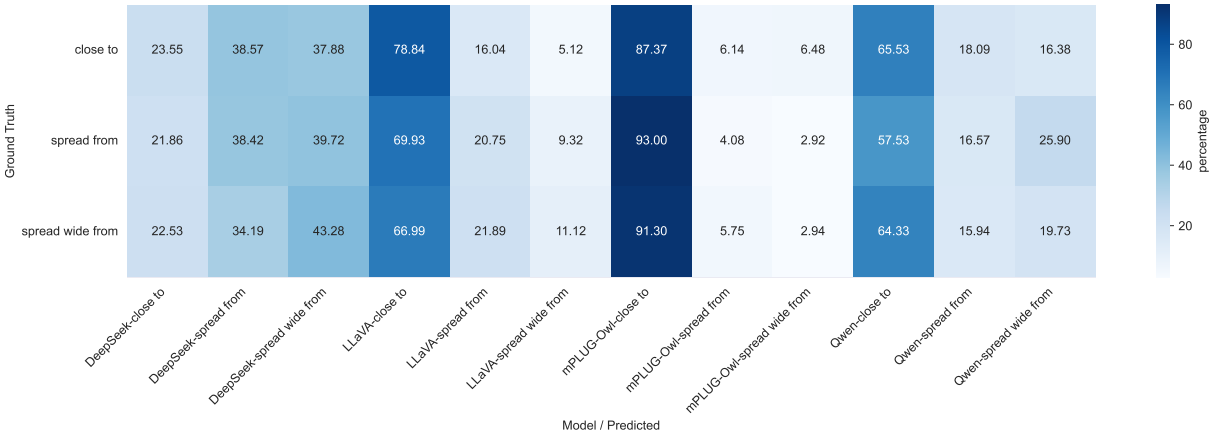


Figure 8: **Distance confusion matrix across four VLMs.** LLaVA, mPLUG-Owl, and Qwen-VL show a strong bias toward predicting *"close to"*, regardless of the ground truth. In contrast, DeepSeek produces a more balanced distribution, indicating better spatial discrimination.

6.1 Systematic Preference for Ambiguous Angle Labels Across VLMs

In Figure 7, we present the confusion matrix for the angle pose descriptor across four vision-language models (VLMs). A clear pattern emerges: all models exhibit a strong bias toward predicting the label "bent slightly inward", regardless of the actual ground truth. This bias dominates the prediction distribution across all ground truth categories.

In addition, each model shows a consistent preference ordering in its predictions across all ground truth classes. For instance, DeepSeek most frequently predicts "bent slightly inward", followed by "bent inward", then "bent completely inward", and finally "straight". This ordered bias persists even when the correct label is different. Similar trends are observed for LLaVA, mPLUG-Owl, and Qwen-VL, though the exact order of predicted preferences varies by model.

These results indicate that current VLMs lack the fine-grained spatial understanding required to

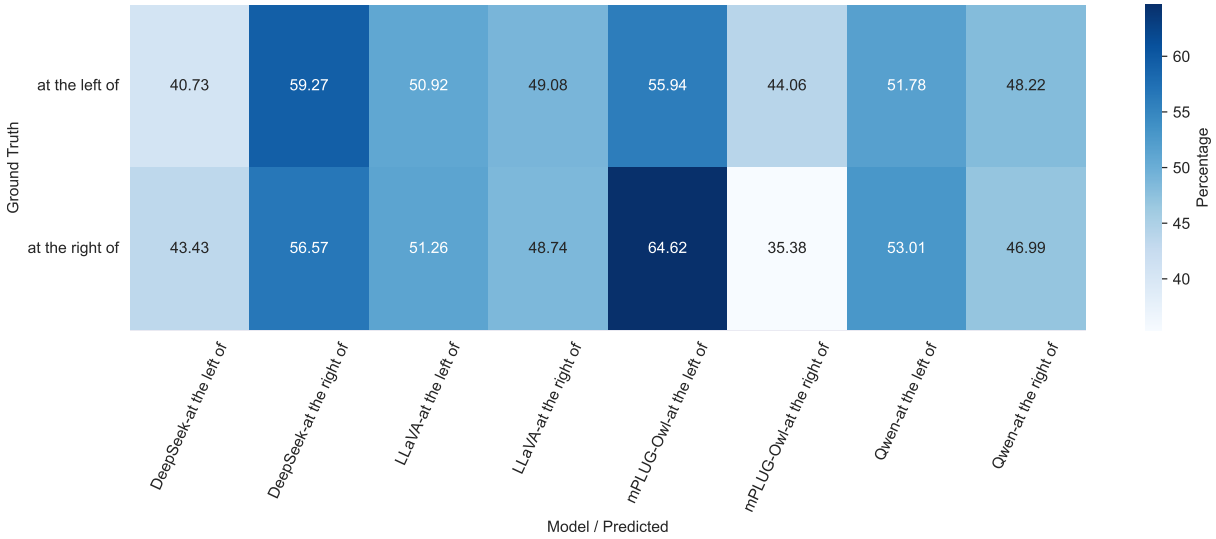


Figure 9: **Relative Position X Confusion Matrix.** Most models show near-symmetric predictions with weak spatial grounding. Notably, mPLUG-Owl exhibits the strongest directional bias, overpredicting "left of" regardless of the ground truth.

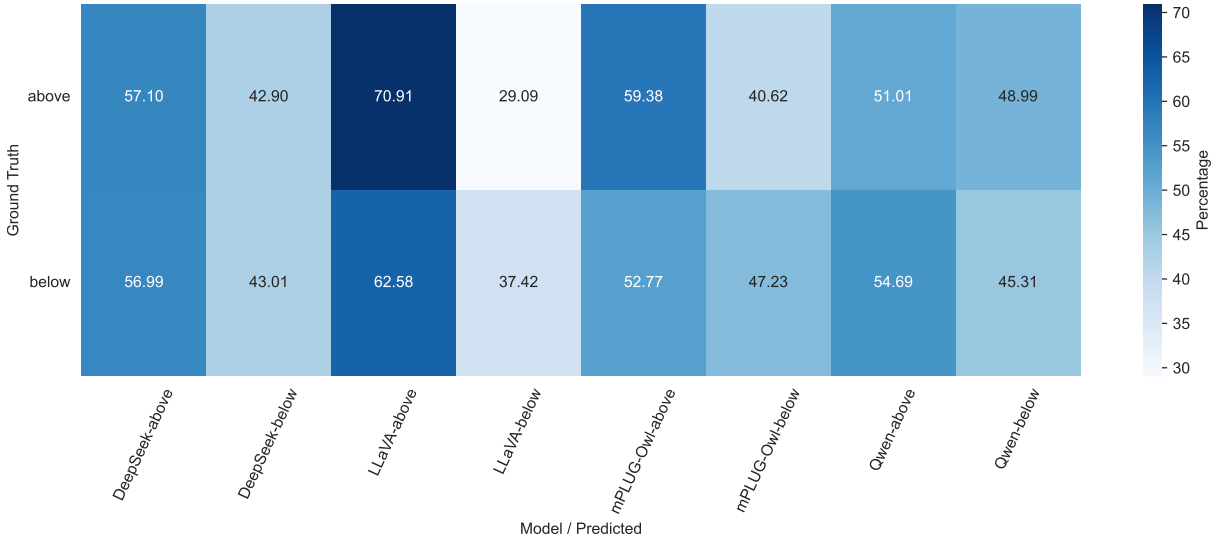


Figure 10: **Relative Position Y Confusion Matrix.** LLaVA exhibits a strong bias toward predicting "above", regardless of the actual label, indicating poor vertical discrimination.

accurately differentiate joint bending angles. Rather than interpreting the true angle from visual cues, models tend to default to mid-range or ambiguous options, revealing a limitation in their ability to reason about subtle variations in hand articulation.

6.2 Over-Reliance on "Close To" Reveals Weak Distance Reasoning

In Figure 8, we present the confusion matrix for the distance pose descriptor, comparing model predictions across three distance-related spatial relationships: close to, spread from, and spread wide from.

In the distance pose descriptor, we observe a general bias toward predicting "close to" regardless of

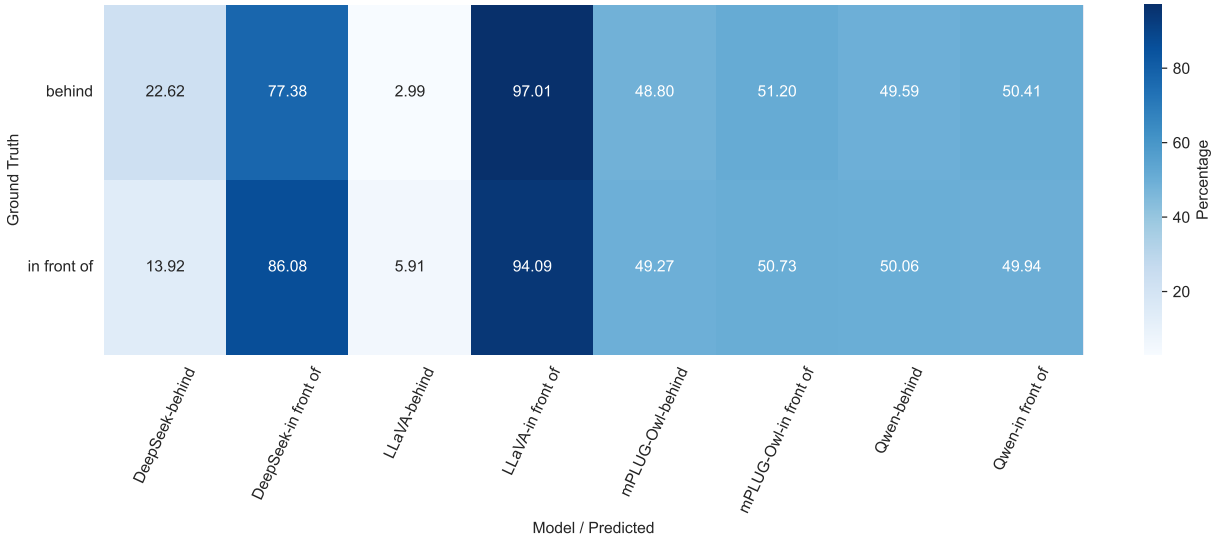


Figure 11: **Relative Position Z Confusion Matrix.** LLaVA and DeepSeek show a strong bias toward predicting “in front of,” often misclassifying “behind” instances.

the ground truth distance label. This tendency is especially pronounced in LLaVA, mPLUG-Owl, and Qwen-VL, which frequently default to “close to” even when the actual relationship is “spread from” or “spread wide from”. In contrast, DeepSeek demonstrates a more balanced prediction pattern across all three distance categories, indicating relatively better spatial discrimination.

While DeepSeek shows a slightly more distributed prediction pattern, it still tends to overpredict spread wide from. The results suggest that VLMs struggle to distinguish varying levels of inter-joint distances from visual input alone.

This over-reliance on the close to class indicates that current models may not be effectively grounding physical separation between joints. Instead, they default to the most semantically neutral or "safe" spatial label when uncertain, mirroring the trends observed in the angle classification task.

6.3 Near-Random Performance in Relative Position Prediction

The confusion matrices for the relative position tasks along the X, Y, and Z axes (Figures 9, 10, and 11) show near-uniform distributions with weak diagonal patterns for most of the cases, indicating behavior similar to random guessing—consistent with around 50% accuracy observed across datasets in Table 3 of the main paper.

On the X-axis, predictions for “left of” and “right of” are nearly symmetric across models except mPLUG-Owl, which tends to overpredict “at the left of”. For the Y-axis, LLaVA overpredicts “above” and mPLUG-Owl favors “below,” while others show slightly more balanced but still unreliable outputs. The Z-axis shows the strongest bias: LLaVA and DeepSeek consistently overpredict “in front of,” failing to capture “behind.”

Overall, these results highlight the lack of spatial grounding in base VLMs, and no model reliably distinguishes binary spatial relations without fine-tuning.

VII Qualitative Results

This section showcases qualitative examples using unseen images and questions from the HandVQA benchmark. These results highlight the strengths and limitations of each model, with a side-by-side comparison of responses from both the base and fine-tuned models. This further demonstrates the benchmark's ability to reveal generalization gaps and reasoning failures in current VLMs. We compare base (before fine-tuning) LLaVA [26] against fine-tuned LLaVA. We choose LLaVA since, after fine-tuning, LLaVA is the best performing VLM among the all VLMs we compared.

Figure 12, 13, and 14 show qualitative results for FreiHAND, InterHand2.6M and FPHA, respectively. The figures show MCQs in the form of HandVQA benchmark that were asked to base (before fine-tuning) LLaVA and fine-tuned LLaVA, and their responses, along with the image the MCQs were about. The MCQs feature all five pose descriptors in all three figures. From the results shown, base LLaVA fails to choose the correct option, sometimes even in the case of easy questions, while LLaVA fine-tuned on our dataset gets them right. Figure 15 uses LLaVA fine-tuned on FreiHAND dataset and base LLaVA, with in-the-wild images, and questions that are different in style and content to the ones we obtain using HandVQA benchmark generation pipeline. While we use LLaVA fine-tuned on fine-grained join-level data from our HandVQA benchmark generation pipeline, the questions in Figure 15 are at finger-level. The base model gets the answers wrong while the fine-tuned model gets them right, showcasing that joint-level training transfers to higher-level geometric reasoning and to images well outside the training distribution. Because no public datasets annotate hand geometry at finger-level granularity, we rely on these qualitative examples to showcase the fine-tuned model's broader spatial competence.

VIII Conclusion

We present HandVQA, a large-scale diagnostic benchmark that evaluates vision-language models on fine-grained spatial reasoning through structured questions about 3D hand poses. Our results show that base VLMs often default to biased predictions and perform at or below random accuracy—especially for joint distances and spatial directions—revealing poor inherent spatial grounding. While LoRA-based fine-tuning significantly improves performance, challenges persist for complex tasks like joint angle interpretation. HandVQA offers a scalable framework for targeted spatial evaluation and highlights the need for stronger spatial inductive biases in VLMs. We hope HandVQA not only serves as a robust evaluation tool but also catalyzes future research aimed at improving spatial grounding and geometric reasoning in vision language models.

Table 3: Angle and Distance Results for all four models. The best, second-best, and third-best models for each dataset in each metric are highlighted as **Gold**, **Silver**, and **Bronze**, respectively.

Model	Tuned	Eval	Angle		Distance	
			Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
Base model (no tuning)						
DeepSeek Janus Pro 7B	–	InterHand2.6M	34.10	0.883	45.55	0.657
DeepSeek Janus Pro 7B	–	FreiHAND	35.31	0.830	44.15	0.668
DeepSeek Janus Pro 7B	–	FPHA	26.46	0.991	39.02	0.819
Finetuned Models						
DeepSeek Janus Pro 7B	InterHand2.6M	InterHand2.6M	68.00	0.334	88.02	0.122
DeepSeek Janus Pro 7B	FreiHAND	FreiHAND	61.30	0.402	85.23	0.151
DeepSeek Janus Pro 7B	FPHA	FPHA	66.08	0.438	81.60	0.184
Base model (no tuning)						
LLaVA Mistral 7B	–	InterHand2.6M	40.08	0.739	16.20	1.293
LLaVA Mistral 7B	–	FreiHAND	42.48	0.678	13.18	1.342
LLaVA Mistral 7B	–	FPHA	23.38	1.011	13.57	1.353
Finetuned Models						
LLaVA Mistral 7B	InterHand2.6M	InterHand2.6M	**74.35**	**0.263**	**90.79**	**0.094**
LLaVA Mistral 7B	FreiHAND	FreiHAND	**62.91**	**0.382**	**86.19**	**0.141**
LLaVA Mistral 7B	FPHA	FPHA	**68.37**	**0.401**	**83.99**	**0.161**
Base model (no tuning)						
mPLUG-Owl 3 7B	–	InterHand2.6M	38.47	0.745	6.88	1.512
mPLUG-Owl 3 7B	–	FreiHAND	45.08	0.626	5.69	1.519
mPLUG-Owl 3 7B	–	FPHA	20.61	1.043	1.04	1.706
Finetuned Models						
mPLUG-Owl 3 7B	InterHand2.6M	InterHand2.6M	70.64	0.304	89.33	0.108
mPLUG-Owl 3 7B	FreiHAND	FreiHAND	58.51	0.437	84.48	0.158
mPLUG-Owl 3 7B	FPHA	FPHA	66.34	0.425	81.20	0.188
Base model (no tuning)						
Qwen 2.5 VL 7B Instruct	–	InterHand2.6M	37.92	0.779	19.58	1.247
Qwen 2.5 VL 7B Instruct	–	FreiHAND	38.70	0.746	20.48	1.208
Qwen 2.5 VL 7B Instruct	–	FPHA	24.22	1.055	18.03	1.306
Finetuned Models						
Qwen 2.5 VL 7B Instruct	InterHand2.6M	InterHand2.6M	67.08	0.341	88.56	0.116
Qwen 2.5 VL 7B Instruct	FreiHAND	FreiHAND	54.55	0.483	82.16	0.182
Qwen 2.5 VL 7B Instruct	FPHA	FPHA	62.94	0.481	80.88	0.192

Table 4: Relative Position Results for all four models. The best, second-best, and third-best models for each dataset are highlighted as **Gold**, **Silver** and **Bronze**, respectively.

Model	Tuned	Eval	Rel. Pos. X Accuracy ↑	Rel. Pos. Y Accuracy ↑	Rel. Pos. Z Accuracy ↑
Base model (no tuning)					
DeepSeek Janus Pro 7B	–	InterHand2.6M	50.41	52.46	51.16
DeepSeek Janus Pro 7B	–	FreiHAND	49.80	51.55	50.03
DeepSeek Janus Pro 7B	–	FPHA	43.02	52.64	61.73
Finetuned Models					
DeepSeek Janus Pro 7B	InterHand2.6M	InterHand2.6M	92.58	96.40	92.16
DeepSeek Janus Pro 7B	FreiHAND	FreiHAND	79.87	85.35	71.53
DeepSeek Janus Pro 7B	FPHA	FPHA	89.94	86.45	88.12
Base model (no tuning)					
LLaVA Mistral 7B	–	InterHand2.6M	49.72	66.26	40.87
LLaVA Mistral 7B	–	FreiHAND	50.25	59.95	50.66
LLaVA Mistral 7B	–	FPHA	50.27	56.33	56.73
Finetuned Models					
LLaVA Mistral 7B	InterHand2.6M	InterHand2.6M	**97.14**	**98.77**	**96.82**
LLaVA Mistral 7B	FreiHAND	FreiHAND	**92.60**	**93.20**	**88.17**
LLaVA Mistral 7B	FPHA	FPHA	**93.81**	**92.80**	**90.25**
Base model (no tuning)					
mPLUG-Owl 3 7B	–	InterHand2.6M	46.52	66.71	49.27
mPLUG-Owl 3 7B	–	FreiHAND	48.47	55.36	50.51
mPLUG-Owl 3 7B	–	FPHA	49.55	49.95	49.77
Finetuned Models					
mPLUG-Owl 3 7B	InterHand2.6M	InterHand2.6M	95.21	97.93	94.65
mPLUG-Owl 3 7B	FreiHAND	FreiHAND	86.18	87.24	81.14
mPLUG-Owl 3 7B	FPHA	FPHA	92.46	88.83	88.75
Base model (no tuning)					
Qwen 2.5 VL 7B Instr.	–	InterHand2.6M	48.98	49.78	49.33
Qwen 2.5 VL 7B Instr.	–	FreiHAND	49.17	49.60	50.19
Qwen 2.5 VL 7B Instr.	–	FPHA	50.98	48.53	49.79
Finetuned Models					
Qwen 2.5 VL 7B Instr.	InterHand2.6M	InterHand2.6M	94.90	97.49	94.11
Qwen 2.5 VL 7B Instr.	FreiHAND	FreiHAND	76.67	80.12	70.23
Qwen 2.5 VL 7B Instr.	FPHA	FPHA	93.45	90.61	87.63

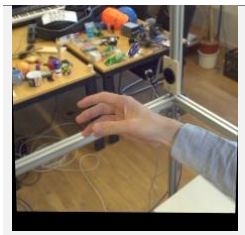
Image	Question	LLaVA (base)	LLaVA (fine-tuned on FreiHAND)
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The index finger is bent completely inward at the proximal interphalangeal joint. B. The index finger is bent inward at the proximal interphalangeal joint. C. The index finger is bent slightly inward at the proximal interphalangeal joint. D. The index finger is straight at the proximal interphalangeal joint.	B ✗	C ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the thumb is spread from the tip joint of the index finger. B. The tip joint of the thumb is close to the tip joint of the index finger. C. The tip joint of the thumb is spread wide from the tip joint of the index finger.	A ✗	C ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the index finger is at the right of the tip joint of the middle finger. B. The tip joint of the index finger is at the left of the tip joint of the middle finger.	A ✗	B ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the thumb is below the tip joint of the middle finger. B. The tip joint of the thumb is above the tip joint of the middle finger.	B ✗	A ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the thumb is in front of the distal interphalangeal joint of the little finger. B. The tip joint of the thumb is behind the distal interphalangeal joint of the little finger	A ✗	B ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the index finger is below the tip joint of the middle finger. B. The tip joint of the index finger is above the tip joint of the middle finger.	A ✗	B ✓

Figure 12: **Qualitative Comparison on FreiHAND. Examples comparing LLaVA (base) and LLaVA fine-tuned on FreiHAND.** Each row shows a question about hand pose from our proposed HandVQA benchmark on an image, with multiple-choice answers. While the base model frequently selects incorrect or spatially inconsistent options, the fine-tuned version consistently predicts the correct answers, demonstrating improved spatial reasoning and alignment with hand joint relationships.

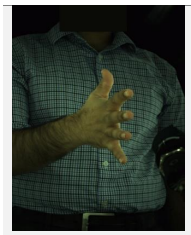
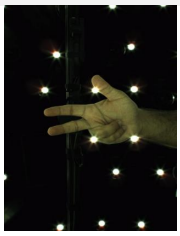
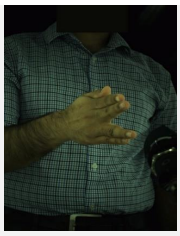
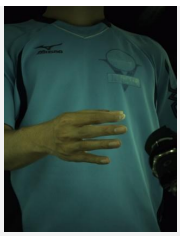
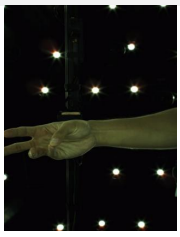
Image	Question	LLaVA (base)	LLaVA (fine-tuned on InterHand2.6M)
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The middle finger is bent completely inward at the distal interphalangeal joint. B. The middle finger is bent inward at the distal interphalangeal joint. C. The middle finger is bent slightly inward at the distal interphalangeal joint. D. The middle finger is straight at the distal interphalangeal joint.	C ✗	D ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The distal interphalangeal joint of the ring finger is close to the distal interphalangeal joint of the little finger. B. The distal interphalangeal joint of the ring finger is spread from the distal interphalangeal joint of the little finger. C. The distal interphalangeal joint of the ring finger is spread wide from the distal interphalangeal joint of the little finger.	B ✗	A ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the thumb is at the right of the tip joint of the index finger. B. The tip joint of the thumb is at the left of the tip joint of the index finger.	A ✗	B ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The distal interphalangeal joint of the index finger is above the distal interphalangeal joint of the middle finger. B. The distal interphalangeal joint of the index finger is below the distal interphalangeal joint of the middle finger.	A ✗	B ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The tip joint of the thumb is behind the metacarpophalangeal joint of the index finger B. The tip joint of the thumb is in front of the metacarpophalangeal joint of the index finger.	B ✗	A ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The metacarpophalangeal joint of the thumb is at the left of the proximal interphalangeal joint of the index finger. B. The metacarpophalangeal joint of the thumb is at the right of the proximal interphalangeal joint of the index finger.	A ✗	B ✓

Figure 13: **Qualitative Comparison on InterHand2.6M. Examples comparing LLaVA (base) and LLaVA fine-tuned on InterHand2.6M.** Each row shows a question about hand pose from our proposed HandVQA benchmark on an image, with multiple-choice answers. While the base model frequently selects incorrect or spatially inconsistent options, the fine-tuned version consistently predicts the correct answers, demonstrating improved spatial reasoning and alignment with hand joint relationships

Image	Question	LLaVA (base)	LLaVA (fine-tuned on FPHA)
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship of right hand in the image: A. The thumb is bent completely inward at the interphalangeal joint. B. The thumb is bent inward at the interphalangeal joint. C. The thumb is bent slightly inward at the interphalangeal joint. D. The thumb is straight at the interphalangeal joint.	C ✗	D ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship right hand in the image: A. The proximal interphalangeal joint of the middle finger is spread from the proximal interphalangeal joint of the index finger. B. The proximal interphalangeal joint of the middle finger is close to the proximal interphalangeal joint of the index finger. C. The proximal interphalangeal joint of the middle finger is spread wide from the proximal interphalangeal joint of the index finger.	A ✗	B ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship of the right hand in the image: A. The carpometacarpal joint of the thumb is at the right of the proximal interphalangeal joint of the index finger. B. The carpometacarpal joint of the thumb is at the left of the proximal interphalangeal joint of the index finger.	A ✗	B ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship of the right hand in the image: A. The tip joint of the thumb is below the tip joint of the index finger. B. The tip joint of the thumb is above the tip joint of the index finger.	B ✗	A ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship of the right hand in the image: A. The metacarpophalangeal joint of the thumb is in front of the proximal interphalangeal joint of the index finger. B. The metacarpophalangeal joint of the thumb is behind the proximal interphalangeal joint of the index finger.	B ✗	A ✓
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship of the right hand in the image: A. The tip joint of the index finger is below the tip joint of the middle finger. B. The tip joint of the index finger is above the tip joint of the middle finger.	B ✗	A ✓

Figure 14: **Qualitative Comparison on FPHA. Examples comparing LLaVA (base) and LLaVA fine-tuned on FPHA.** Each row shows a question about hand pose from our proposed HandVQA benchmark on an image, with multiple-choice answers. While the base model frequently selects incorrect or spatially inconsistent options, the fine-tuned version consistently predicts the correct answers, demonstrating improved spatial reasoning and alignment with hand joint relationships.

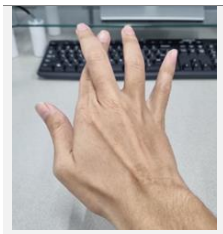
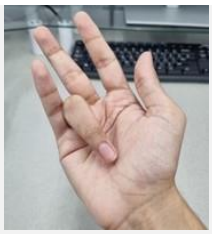
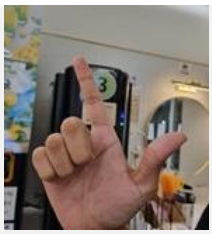
Image	Question	Options	LLaVA (base)	LLaVA (fine-tuned on FreiHAND)	Why the question matters
	Are any fingers crossing each other?	A) Yes B) No	B ✗	A ✓	Detects self-occlusion patterns
	Which pair of fingers are spread widest at their tips?	A) Index–Middle B) Middle–Ring C) Ring–Little	A ✗	B ✓	Compares distance between fingers
	Which fingertip lies closest to the palm center?	A) Index B) Ring C) Little	A ✗	B ✓	Combines distances from a reference point.
	Is the index finger crossing over the middle finger?	A) Yes B) No	B ✗	A ✓	Requires identifying both depth and X-ordering (left/right)
	Is the thumb right or left of the index finger?	A) Right B) Left	B ✗	A ✓	X-order (left/right) reasoning
	Are the two fingers touching?	A) Yes B) No	B ✗	A ✓	Detect contact between fingers.

Figure 15: **Qualitative Results on In-the-Wild Images.** We evaluate spatial reasoning on challenging questions using in-the-wild images. The fine-tuned LLaVA outperforms the base model on tasks involving occlusion, depth, and inter-finger relationships, demonstrating improved generalization beyond the training data.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Chen Bao, Jiarui Xu, Xiaolong Wang, Abhinav Gupta, and Homanga Bharadhwaj. Handsonvlm: Vision-language models for hand-object interaction prediction. *arXiv preprint arXiv:2412.13187*, 2024.

[5] Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, et al. Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity. *arXiv preprint arXiv:2503.11557*, 2025.

[6] Beitao Chen, Xinyu Lyu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. In *NeurIPS*, 2024.

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023.

[9] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *CVPR*, 2023.

[10] Sharice Clough and Melissa C Duff. The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience*, 2020.

[11] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: 3D Human Poses from Natural Language. In *ECCV*, 2022.

[12] Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In *ICLR*, 2025.

[13] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024.

[14] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[17] Elkhan Ismayilzada, MD Khalequzzaman Chowdhury Sayem, Yihalem Yimolal Tiruneh, Mubarrat Tajoar Chowdhury, Muhammadjon Boboev, and Seungryul Baek. Qort-former: Query-optimized real-time transformer for understanding two hands manipulating objects. In *AAAI*, 2025.

[18] Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. NEFTune: Noisy embeddings improve instruction finetuning. In *ICLR*, 2024.

[19] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *ACM MM*, 2024.

[20] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[21] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023.

[22] Jihyung Kil, Farideh Tavazoee, Dongyeop Kang, and Joo-Kyung Kim. II-MMR: Identifying and improving multi-modal multi-hop reasoning in visual question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *ACL*, 2024.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

[25] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[27] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *ICLR*, 2025.

[28] Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. In *EMNLP*, 2024.

[29] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang, editors, *ALVR*, 2024.

[30] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

[31] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

[32] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *ICCV*, 2023.

[33] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020.

[34] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *CVPR*, 2023.

[35] Dominick Reilly, Rajatsubhra Chakraborty, Arkaprava Sinha, Manish Kumar Govind, Pu Wang, Francois Bremond, Le Xue, and Srijan Das. Llavidal: A large language vision model for daily activities of living. In *CVPR*, 2025.

[36] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022.

[37] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *EMNLP*, 2024.

[38] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *ACL*, 2019.

[39] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.

[40] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *ICLR*, 2024.

[41] Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. In *ICML*, 2024.

[42] Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, and Qin Jin. Do egocentric video-language models truly understand hand-object interactions? In *ICLR*, 2025.

[43] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Shuguang Cui, and Zhen Li. Comprehensive visual question answering on point clouds through compositional scene manipulation. *IEEE Transactions on Visualization & Computer Graphics*, 2023.

[44] Cheng Yang, Rui Xu, Ye Guo, Peixiang Huang, Yiru Chen, Wenkui Ding, Zhongyuan Wang, and Hong Zhou. Improving vision-and-language reasoning via spatial relations modeling. In *WACV*, 2024.

[45] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *ICLR*, 2025.

[46] Kivanc Yangi, Thomas J On, Yuan Xu, Arianna S Gholami, Jinpyo Hong, Alexander G Reed, Pravarakhya Puppalla, Jiuxu Chen, Jonathan A Tangsrivimol, Baoxin Li, et al. Artificial intelligence integration in surgery through hand and instrument tracking: a systematic literature review. *Frontiers in surgery*, 2025.

[47] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024.

[48] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019.

[49] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[50] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. *arXiv preprint arXiv:2412.12693*, 2024.

[51] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024.

[52] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *CVPR*, 2019.