

HandVQA: Diagnosing and Improving Fine-Grained Spatial Reasoning about Hands in Vision-Language Models

MD Khalequzzaman Chowdhury Sayem^{1*} Mubarrat Tajoar Chowdhury^{1*} Yihalem Yimolal Tiruneh¹
 Muneeb A. Khan¹ Muhammad Salman Ali¹ Binod Bhattarai^{2,3,4†} Seungryul Baek^{1†}
¹UNIST ²University of Aberdeen ³University College London ⁴Fogsphere (Redev.AI Ltd), UK

Abstract

Understanding the fine-grained articulation of human hands is critical in high-stakes settings such as robot-assisted surgery, chip manufacturing, and AR/VR-based human-AI interaction. 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 and 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 and Qwen-VL) 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 these critical reasoning gaps but provides a validated path to improvement. We demonstrate that the 3D-grounded spatial knowledge learned from our benchmark transfers in a zero-shot setting, significantly improving accuracy of model on novel downstream tasks like hand gesture recognition (+10.33%) and hand-object interaction (+2.63%). Project page, code, and dataset: <https://kcsayem.github.io/handvqa/>.

1. Introduction

Imagine a high-stakes environment like robot-assisted surgery, where a subtle misinterpretation of a surgeon’s hand gesture whether a finger is slightly bent or fully extended can mean the difference between a successful operation and a critical error [19, 41]. Or consider chip manufacturing, where robotic agents mimic human motions with microscopic precision, and a misjudged fingertip angle can cause costly defects [17, 53]. In

*Equal contribution.

†These authors jointly supervised this work.

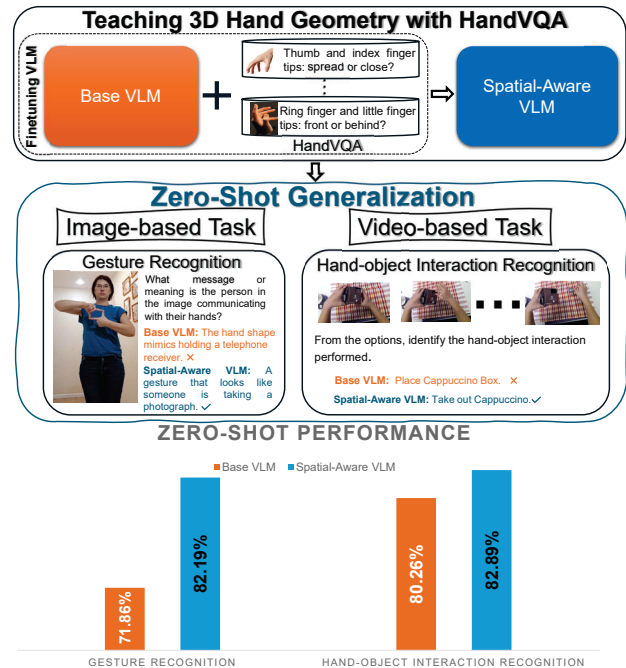


Figure 1. **Overview of HandVQA’s transfer effect.** Fine-tuning a base Vision-Language Model (VLM) on HandVQA teaches it explicit 3D hand geometry and joint-level spatial reasoning. The resulting Spatial-Aware VLM exhibits zero-shot generalization to novel downstream tasks: both image-based gesture recognition and video-based hand-object interaction recognition. Spatial-Aware VLM achieves consistent accuracy gains without task-specific training.

such scenarios, understanding the fine-grained articulation of the human hand is not just desirable, it is essential.

Vision Language Models (VLMs) have rapidly become the perceptual and decision-making core of embodied agents, AR/VR platforms, and multimodal assistants due to their ability to unify visual perception and natural-language reasoning within a single foundation model [42]. As these systems increasingly rely on such unified reasoning, fine-grained hand-pose understanding becomes crucial [4, 59]. Hands are often the primary medium through which humans convey actions, intent, and control [14]; misinterpreting whether a finger is curled, pointing, or relaxed

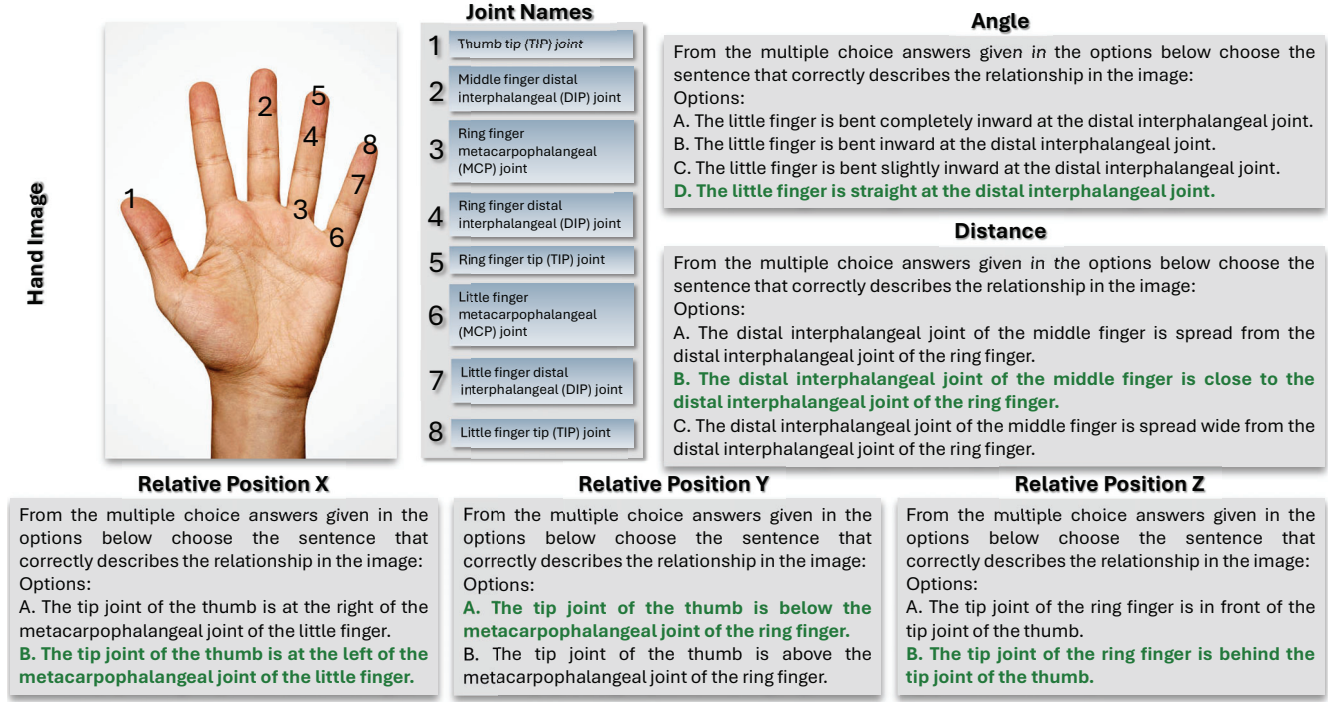


Figure 2. **Overview of HandVQA Question Format.** This figure illustrates the structure of our benchmark, which divides hand pose estimation into five sub-tasks: Angle, Distance, and Relative Position along X, Y, and Z axes. A hand image with annotated joint indices (top left) supports multiple-choice questions per task type, derived from 3D joint coordinates and the correct answers are shown in **green**.

can result in semantic confusion or physical failure. For VLMs to be safely deployed in these real-world contexts, robust, fine-grained hand understanding is a foundational requirement. 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 [10, 31, 48, 65]. 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% [27]. 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 [67], InterHand2.6M [45], and FPHA [20]. 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. 2). While we draw conceptual motivation from Delmas et al. [15], 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 [34], Qwen-VL [3], and DeepSeek [40]. These models are fine-tuned using parameter-efficient LoRA adapters [21] 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 archi-

tures 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, 16, 46, 63]. HandVQA addresses a critical gap in existing benchmarks, but our work demonstrates it is more than just a diagnostic tool; it is a robust training resource. We hypothesize that by mastering the 3D-grounded spatial reasoning in HandVQA, models acquire a transferable skill. We validate this by showing that models trained with HandVQA achieve significant zero-shot generalization to novel, unseen downstream tasks, including static Gesture Recognition and dynamic, video-based Hand-Object Interaction (See Fig. 1). HandVQA thus provides a concrete path to improving VLM spatial reasoning capabilities broadly. To summarize our contributions include:

- We introduce a large-scale benchmark with over 1.6 million questions, created via an automated pipeline that generates anatomically-grounded VQA queries about 3D hand joint angles, distances, and relative positions.
- We provide a Comprehensive evaluation of leading VLMs (e.g., LLaVA, DeepSeek, Qwen-VL), revealing their systemic failures in base spatial grounding (often performing at or below random guess) and highlighting persistent challenges, such as interpreting fine-grained joint angles, even after fine-tuning.
- We demonstrate HandVQA is more than a diagnostic tool; it is a robust training resource. We show that the 3D-grounded spatial knowledge learned from our benchmark **transfers in a zero-shot setting**, significantly improving performance on novel, unseen downstream tasks. This includes a **+10.33%** absolute accuracy gain on static **Gesture Recognition** and a **+2.63%** gain on video-based **Hand-Object Interaction**.

2. Related Works

Hallucinations and Alignment Failures in VLMs. Vision-language models (VLMs) often suffer from hallucinations, *i.e.*, generating content not grounded in the image; such as nonexistent objects, incorrect attributes, or false spatial relations [38]. For example, a model might describe a ring that isn’t there or misstate left/right positions [27, 33]. Even state-of-the-art models like GPT-4V [1] and BLIP-2 [32] hallucinate objects [33], struggle with spatial terms [27], or fabricate colors and shapes [25]. These issues stem from weak visual-textual alignment and overreliance on language priors [24, 57, 62]. Liu et al. [35] emphasizes the lack of grounding signals in training data. Recent solutions include evaluation frameworks such as POPE [33], NOPE [39], and Hal-Eval [25], as well as mitigation strategies, including contrastive decoding, reinforcement learning [8], and robust visual encoders [36]. Still, hallucinations remain prevalent, even in top-performing models [6, 11, 13, 43]. Our work exposes a distinct form of

hallucination: *pose hallucination*, where models misinterpret joint-level configurations. For instance, a model may predict a bent joint where the ground truth is straight, or repeatedly label finger pairs as ‘close’ despite substantial separation between them. Such errors reveal fine-grained alignment failures and highlight the need for diagnostics beyond object-level VQA.

Spatial and Relational Reasoning in Multimodal Models.

Beyond identifying objects, true vision-language understanding requires reasoning about spatial and relational configurations. Kamath et al. [27] revealed that many VLMs fail even simple spatial tasks like distinguishing a dog under a table versus on a table. Similarly, Zhang et al. [66] presents SPHERE, a hierarchical evaluation of spatial skills from basic (positions, distances) to complex (occlusion, physical plausibility). On SPHERE, the best model achieved only 68% accuracy far below human performance of 93% [66]. These results show that high-level accuracy on generic VQA does not imply reliable spatial understanding [52, 58]. To address these limitations, recent work has explored ways to explicitly inject spatial structure into multimodal models. For example, Yang et al. [61] proposed a direction on improving model training by incorporating a spatial relation graph into vision-language pretraining, improving performance on reasoning tasks like VCR [64] and NLVR2 [54]. Chen et al. [9] 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. Because VLMs [3, 40, 44] frequently exploit shortcut cues in benchmarks such as CLEVR [26] and GQA [22], their reported accuracy can obscure underlying spatial failures. This motivates the development of diagnostic evaluations like What’sUp [27] and SPHERE [66]. 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*.

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 [26] introduced logical queries in synthetic scenes; GQA [22] used real images and scene graphs for multi-step reasoning. Recent benchmarks like A-OKVQA [50] and Encyclopedic-VQA [43] test domain knowledge, but fine-grained accuracy remains low: e.g., PaLI [11] achieves only 13% on Encyclopedic-VQA [43]. Our benchmark instead focuses on fine-grained spatial understanding. It aligns with works like NLVR2 [54] and CLEVR-3D [60] that emphasize logical consistency and physical reasoning. Recent diagnostic frameworks such as VERIFY [5] evaluate if models follow reasoning chains rather than shallow cues. II-MMR [29] categorizes multi-hop VQA questions and shows many “hard” questions are solved via superficial features. To maintain focus, we keep HandVQA

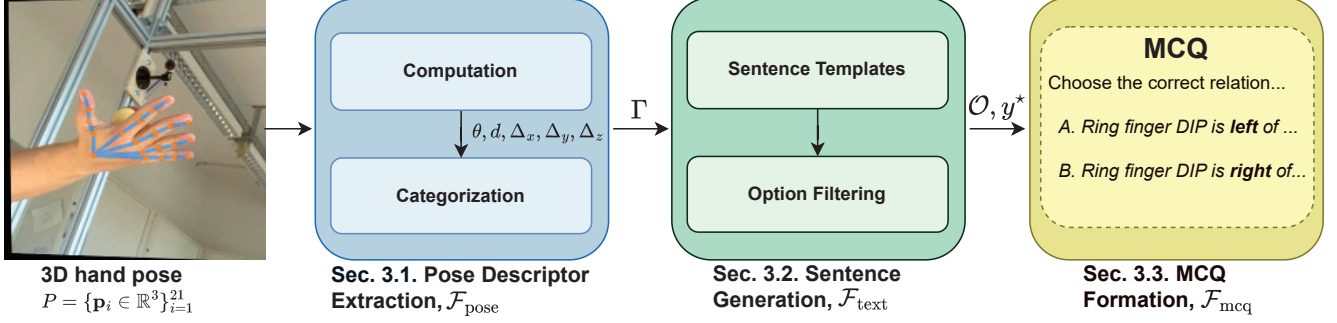


Figure 3. **Overview of the HandVQA pipeline.** The pipeline converts normalized 3D hand joints into interpretable VQA pairs through three deterministic stages: (1) $\mathcal{F}_{\text{pose}}$ computes continuous pose descriptors: angles (θ), distances (d), and relative positions ($\Delta_x, \Delta_y, \Delta_z$) and categorizes them into discrete pose descriptors (Γ); (2) $\mathcal{F}_{\text{text}}$ fills deterministic sentence templates using Γ and filters correct and incorrect options to form candidate answer options ($\mathcal{O}$) and the correct label y^* ; (3) $\mathcal{F}_{\text{mcq}}$ constructs multiple-choice questions (MCQ) by pairing each image with $\mathcal{O}$ and its correct label y^* .

questions multiple-choice and relatively constrained—allowing clean assessment of spatial reasoning accuracy. It thus adds to diagnostic tools assessing whether progress in general VQA translates to true fine-grained understanding.

Applications Requiring Precise Hand Pose Understanding.

Precise hand articulation understanding is critical across domains. In robotics, accurate perception and manipulation of hand gestures is essential. Duan et al. [16] 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 [59], while several works show that better 3D hand pose estimation improves activity recognition [7, 12, 23, 46, 47, 49]. More specifically, Ohkawa et al. [46] argue that hand pose is a compact representation of action. In augmented reality interfaces, devices like HoloLens or Meta Quest utilize hand tracking for gesture-based commands [46]. 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 [63]. Thus, many real-world applications demand detailed, error-free hand pose understanding. HandVQA offers a direct path to both benchmark and improve these capabilities. While many of the aforementioned works depend on reliable hand pose recognition as a prerequisite, our work demonstrates that the 3D-grounded spatial reasoning learned from HandVQA is a transferable skill. We validate this in our zero-shot experiments (Sec. 4.3), where models trained with HandVQA achieve significantly improved performance on these exact types of downstream tasks: static Gesture Recognition and dynamic, video-based Hand-Object

Interaction. By forcing models to reason about joint-level spatial details, HandVQA not only evaluates but actively improves multimodal perception and grounding for these critical applications.

3. Pipeline to generate HandVQA benchmark

This section presents the design of the automatic VQA generation pipeline, which deterministically converts normalized 3D hand joint coordinates into natural language question-answer (QA) pairs for an image, I given the normalized 3D hand pose $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{21}$, and transforms geometric representations into linguistically interpretable forms. The three stages—pose descriptor extraction, sentence generation, and MCQ formation are described below.

3.1. Pose Descriptor Extraction ($\mathcal{F}_{\text{pose}}$)

Given normalized 3D hand joint coordinates $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{21}$, $\mathcal{F}_{\text{pose}}$ computes continuous pose descriptors and categorizes them to describe the geometric and spatial relationships among hand joints. Formally, we define the continuous pose descriptor set as

$$\Psi = \left\{ \theta_j \mid j \in \mathcal{J}_{\text{angle}} \right\} \cup \left\{ d_{(i,k)} \mid (i,k) \in \mathcal{J}_{\text{pair}} \right\} \cup \left\{ \Delta_a(i,k) \mid (i,k) \in \mathcal{J}_{\text{pair}}, a \in \{x,y,z\} \right\}. \quad (1)$$

where $\mathcal{J}_{\text{angle}}$ and $\mathcal{J}_{\text{pair}}$ denote the predefined joint index sets used for angle and pairwise computations, respectively. Here, indices i, j, k refer to joint identifiers in the hand skeleton, and a denotes one of the Cartesian axes $\{x, y, z\}$.

- **Angle (θ_j):** Measures the bending degree at joint j , formed by its adjacent joints $a(j)$ and $b(j)$ along the same finger:

$$\theta_j = \arccos \frac{(\mathbf{p}_{a(j)} - \mathbf{p}_j) \cdot (\mathbf{p}_{b(j)} - \mathbf{p}_j)}{\|\mathbf{p}_{a(j)} - \mathbf{p}_j\| \|\mathbf{p}_{b(j)} - \mathbf{p}_j\|}. \quad (2)$$

Angles are categorized into four bins based on fixed thresholds: *bent completely inward*, *bent inward*, *bent slightly inward*, and *straight*.

Table 1. Pose descriptor categorization with conditions and illustration.

Illustration	Pose Descriptor	Category Label	Condition
	angle	bent completely inward	$\theta < 105^\circ$
		bent inward	$105^\circ \leq \theta < 150^\circ$
		bent slightly inward	$150^\circ \leq \theta < 170^\circ$
		straight	$\theta \geq 170^\circ$
	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	$\Delta_x < -0.15$
		aligned	$-0.15 \leq \Delta_x < 0.15$
		at the right of	$\Delta_x \geq 0.15$
	rel. pos. (Y)	below	$\Delta_y < -0.15$
		aligned	$-0.15 \leq \Delta_y < 0.15$
		above	$\Delta_y \geq 0.15$
	rel. pos. (Z)	behind	$\Delta_z < -0.15$
		aligned	$-0.15 \leq \Delta_z < 0.15$
		in front of	$\Delta_z \geq 0.15$

- **Distance** ($d_{(i,k)}$): Quantifies the Euclidean distance between two joints i, k : $d_{(i,k)} = \|\mathbf{p}_i - \mathbf{p}_k\|_2$. These distances are discretized into: *close to*, *spread from*, and *spread wide from*.
- **Relative Position** ($\Delta_a(i,k)$): Represents the signed offset of joint i relative to k along each axis: $\Delta_a(i,k) = \langle \mathbf{p}_i - \mathbf{p}_k, \mathbf{a} \rangle$. The relative positions are categorized as *left/aligned/right*, *below/aligned/above* and *behind/aligned/in front of*. Instances categorized as *aligned* are excluded due to visual ambiguity. Further details in supplemental section 6.

All threshold definitions are summarized in Table 1. This process converts each continuous pose descriptor $\psi_n \in \Psi$ into a discrete pose descriptor label γ_n using the category label from Table 1, forming the discrete pose descriptors

$$\Gamma = \{\gamma_n \mid \psi_n \in \Psi, n = 1, 2, \dots, N\}, \quad (3)$$

where N denotes the total number of pose descriptors computed from a hand pose. This step ensures every continuous pose descriptor has a unique linguistic interpretation.

3.2. Sentence Generation ($\mathcal{F}_{\text{text}}$)

Given discrete pose descriptors Γ , the sentence generation step first fills deterministic templates with joint names and category labels to create natural language sentences $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$. Each descriptor type uses a fixed syntactic structure. For instance, the template for a distance descriptor is:

“The {joint A} joint of the {finger A} is {category label} the {joint B} joint of the {finger B}.”

If the category label is “close to,” the resulting sentence becomes: “The distal interphalangeal joint of the middle finger is close to the distal interphalangeal joint of the ring finger.” From the full set of generated sentences $\mathcal{S}$, $\mathcal{F}_{\text{text}}$ identifies the

single sentence that corresponds to the true category label and treats the remaining sentences for that joint pair as distractors in the option filtering step. This produces a pool of candidate answer options $\mathcal{O}$, along with the index of the correct label y^* , which are later used to form an MCQ.

3.3. MCQ Formation ($\mathcal{F}_{\text{mcq}}$)

Given the the set of candidate answer options $\mathcal{O}$, and the index of the correct label y^* , $\mathcal{F}_{\text{mcq}}$ constructs an MCQ prompt asking the model to identify the sentence that correctly describes the relationship between the specified joint(s). To ensure diverse coverage of geometric relationships, multiple joint or joint-pair instances are considered for each descriptor type. Prior to filtering out “aligned” cases, up to 107 possible MCQs can be generated per image. For scalability, we randomly sample five joint or joint-pair instances per descriptor type, yielding a total of 25 MCQs per image across the five pose descriptor types (angle, distance, and relative positions along X/Y/Z). See Fig. 2 for examples and Section 6 of the supplemental for sampling details. 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.

4. Experiments

4.1. Datasets, Models, and Evaluation Metrics

Dataset Construction. We use three hand datasets to construct our HandVQA benchmark: FreiHAND [67], InterHand2.6M [45], and FPHA [20]. 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. Further details on dataset construction are available in the supplemental section 7.

Models. We evaluate four state-of-the-art 7B vision-language models—LLaVA Mistral[34], DeepSeek Janus Pro [40], and Qwen 2.5 VL Instruct [3] on the HandVQA benchmark, using both base and LoRA [21] 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

Table 2. Angle and Distance Results for all three 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 $\uparrow$	MAE $\downarrow$	Accuracy $\uparrow$	MAE $\downarrow$
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)						
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

(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.

4.2. Results and Analysis

Tables 2 and 3 summarize the experimental results. In the following analysis, we examine the performance and behavior of the evaluated VLMs on our proposed HandVQA benchmark. Further detailed experimental analyses, including confidence analysis and cross-dataset comparisons, are provided in the supplementary (Sec. 8).

Scarcity of data the cause for VLMs’ poor performance. As per Table 2 and Table 3, 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 2 shows, base VLMs generally seem to perform poorly on distance pose descriptor with LLaVA 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 two 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 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 supplementary. While base VLMs struggles 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 finetuned on the FPHA dataset.

VLMs struggle to grasp angle even after fine-tuning. According to Table 2, 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 InterHand2.6M

Table 3. Relative Position Results for all three 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)					
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

dataset. A common trend observed in the confusion matrix in the supplemental 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 [18].

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 2, for angle, among base models, while base LLaVA on average performs best in terms of accuracy and base DeepSeek performs best in distance in terms of both MAE and accuracy. As can be seen in Table 3, 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 3, 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

Table 4. **Zero-shot generalization results.** HandVQA fine-tuning strengthens 3D spatial reasoning, yielding higher accuracy (%) on gesture and hand-object interaction recognition.

Setup	Interaction Recognition	Gesture Recognition
LLaVA Mistral 7B [34]	-	57.42
Qwen 2.5 VL 7B [3]	80.26	71.86
LLaVA Mistral 7B <i>finetuned</i>	-	69.58
Qwen 2.5 VL 7B <i>finetuned</i>	82.89	82.19

example, LLaVA achieves about 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 work, which explicitly isolates and evaluates these fine-grained spatial relations, enabling effective diagnosis and improvement of directional understanding in VLMs.

4.3. Zero-Shot Generalization to Novel Hand-Related Tasks

Beyond diagnosing failures, we hypothesize that the 3D-grounded spatial reasoning taught by HandVQA is a transferable skill that should generalize to other novel, hand-related downstream tasks. To test this transferable learning, we designed a zero-shot generalization experiment, with results shown in Table 4. We selected two distinct, novel tasks for evaluation: static image-based Gesture recognition and video-based hand-object Interaction recognition. For these experiments, we chose LLaVA Mistral 7B [34], best performing model in our benchmark and Qwen 2.5 VL 7B [3], given its support for temporal sequences (required for the video-based interaction task). Following common VLM finetuning practice [18, 37, 51, 56], we include a small, domain-agnostic instruction-alignment set alongside HandVQA (FreiHAND + InterHand2.6M + FPHA) to preserve general instruction-following capabilities during training.

Gesture Recognition. For this zero-shot task, we utilized the HaGRID dataset [28], which provides full-body images annotated with 34 hand gestures (e.g., ‘taking picture’, ‘heart’, ‘peace’, etc.). To adapt this dataset into a VLM-evaluable format, we first convert each gesture label into a detailed natural-language description using Gemini [55]. Then, based on these descriptions, we generate multiple-choice questions (MCQs) with one correct option and three carefully designed distractor options per sample. Following this process, we constructed a final test set of 33,500 MCQs from the HaGRID test split (Further details is available in supplemental.). As shown in Table 4, the base LLaVA and Qwen models achieve 57.42% and 71.86% accuracy, respectively. After fine-tuning, their accuracies increase substantially to 69.58% (LLaVA) and 82.19% (Qwen), demonstrating strong positive transfer of 3D-grounded spatial reasoning to gesture understanding. Since both evaluated

models show consistent improvement, we expect this trend to extend to other VLMs as well, given that HandVQA targets fundamental spatial reasoning skills shared across architectures.

Hand-Object Interaction Recognition. This experiment aimed to determine if the 3D spatial knowledge of hand geometry learned from HandVQA could transfer to improve the understanding of complex, egocentric hand-object interactions. We used the H2O dataset [30], which provides diverse video-based interactions such as ‘taking out cappuccino from the coffee box’, ‘grabbing lotion’, etc. From the H2O test split, we constructed a multiple-choice question (MCQ) benchmark, sampling 4 images per video for each question. As shown in Table 4, our results validate this hypothesis. The base Qwen-VL model achieved 80.26% accuracy and the finetuned model achieved the highest accuracy of 82.89%. LLaVA was not evaluated in this setting due to its lack of temporal sequence support. This demonstrates that the spatial skills imparted by our benchmark are not limited to static poses but successfully transfer to enhance performance on dynamic, video-based interaction tasks.

5. Conclusion and Future Direction

HandVQA provides the first large-scale benchmark for diagnosing fine-grained 3D spatial reasoning about hands in VLMs. Our analyses reveal that current models, despite strong general perception and instruction-following abilities, systematically fail on geometric relations such as joint angles, distances, and relative positions. By introducing a fully automated 3D-grounded annotation pipeline, we make these failures explicit and measurable. Furthermore, our zero-shot experiments show that HandVQA acts as an effective source of transferable 3D knowledge. Fine-tuning on our benchmark substantially improves performance on unseen tasks, such as gesture recognition and hand-object interaction.

Future direction. As the first large-scale benchmark for fine-grained hand-centric spatial reasoning, HandVQA focuses on core capabilities and naturally opens several research avenues: (i) Its discretized geometric representation offers a controlled and deterministic evaluation setup, with fixed thresholds providing a simplified view of the continuous 3D space. Future work can develop adaptive or learned mappings that better capture perceptual distinctions and support richer geometric reasoning. (ii) The templated language ensures unambiguous supervision, yet expanding to more diverse phrasing, comparative expressions, and explanations could yield stronger linguistic grounding of 3D geometry. (iii) Since the benchmark currently targets static images, extending it to video would enable reasoning about motion cues and contact dynamics central to many interaction tasks. (iv) Integrating HandVQA with Vision Language Action (VLA) models offers a path toward improved grasp planning, dexterous control, and embodied manipulation. We hope HandVQA facilitates the development of multimodal models with stronger geometric understanding and more reliable physical reasoning.

Acknowledgments

This work is supported by NRF grant (No. RS-2025-00521013 50%, No. RS-2025-02216916 10%) and IITP grants (No. RS2020-II201336 Artificial intelligence graduate school program (UNIST) 5%; No. RS-2025-25442824 AI Star Fellowship Program(UNIST) 5%; No. RS-2022-II220264 Comprehensive video understanding and generation with knowledge-based deep logic neural network 10%, No. RS-2025-25442149 LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain 10%), funded by the Korean government (MSIT). This work is also supported by the InnoCORE program of the Ministry of Science and ICT(26-InnoCORE-01) 10%.

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. 3
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2, 3
- [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. 2, 3, 5, 8
- [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. 1, 3, 4
- [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. 3
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 3
- [7] Junuk Cha, Jiyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *CVPR*, 2024. 4
- [8] Beita 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. 3
- [9] 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. 3
- [10] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *ICML*, 2025. 2
- [11] 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. 3
- [12] Hoseong Cho, Chanwoo Kim, Jiyeon Kim, Seongyeon Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *CVPR*, 2023. 4
- [13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1), 2023. 3
- [14] 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. 1
- [15] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: 3D Human Poses from Natural Language. In *ECCV*, 2022. 2
- [16] 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. 3, 4
- [17] Hongyan Dui, Huiting Xu, Luyu Zhang, and Jia Wang. Cost-based preventive maintenance of industrial robot system. *Reliability Engineering & System Safety*, 2023. 1
- [18] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 7, 8
- [19] Fanny Ficuciello, Alberto Villani, Tommaso Lisini Baldi, and Domenico Prattichizzo. A human gesture mapping method to control a multi-functional hand for robot-assisted laparoscopic surgery: The musha case. *Frontiers in Robotics and AI*, 2021. 1

- [20] 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. 2, 5
- [21] 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. 2, 5
- [22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3
- [23] Elkhani 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. 4
- [24] 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. 3
- [25] 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. 3
- [26] 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. 3
- [27] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023. 2, 3
- [28] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid – hand gesture recognition image dataset. In *WACV*, 2024. 8
- [29] Jihyung Kil, Farideh Tavazoei, Dongyeop Kang, and Joo-Kyung Kim. II-MMR: Identifying and improving multi-modal multi-hop reasoning in visual question answering. In *ACL*, 2024. 3
- [30] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 8
- [31] Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *ICCV*, 2025. 2
- [32] 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. 3
- [33] 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. 3
- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 5, 8
- [35] 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. 3
- [36] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *ICLR*, 2025. 3
- [37] Wenzhuo Liu, Fei Zhu, Haiyang Guo, Longhui Wei, and Cheng-Lin Liu. Llava-c: Continual improved visual instruction tuning, 2025. 8
- [38] 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. 3
- [39] 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 *ALVR*, 2024. 3
- [40] 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. 2, 3, 5
- [41] Runzhuo Ma, Ashwin Ramaswamy, Jiashu Xu, Loc Trinh, Dani Kiyasseh, Timothy N Chu, Elyssa Y Wong, Ryan S Lee, Ivan Rodriguez, Gina DeMeo, et al. Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *NPJ Digital Medicine*, 2022. 1
- [42] Yuen 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. 1
- [43] Thomas Mensink, Jasper Uijlings, Lluís 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. 3
- [44] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024, 2024. 3
- [45] Gyeongsik Moon, Shou-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. 2, 5
- [46] 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. 3, 4
- [47] Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In *CVPR*, 2025. 4
- [48] Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *ICCV*, 2025. 2
- [49] Dominick Reilly, Rajat Subhra 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. 4
- [50] 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. 3
- [51] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Y. Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Mixture-of-experts meets instruction

tuning: A winning combination for large language models. In *ICLR*, 2024. 8

- [52] 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 *EMNLP*, 2024. 3
- [53] A G Starr, R J Wynne, and I Kennedy. Failure analysis of mature robots in automated production. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 213(8):813–824, 1999. 1
- [54] 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. 3
- [55] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2024. 8
- [56] Renxi Wang, Haonan Li, Minghao Wu, Yuxia Wang, Xudong Han, Chiyu Zhang, and Timothy Baldwin. Demystifying instruction mixing for fine-tuning large language models. In *ACL*, 2024. 8
- [57] 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. 3
- [58] 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. 3
- [59] 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. 1, 4
- [60] 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. 3
- [61] 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. 3
- [62] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *ICLR*, 2025. 3
- [63] 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. 3, 4
- [64] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 3
- [65] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *TPAMI*, 2024. 2
- [66] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Wanglu Wanglu. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *ACL*, 2025. 3
- [67] 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 *ICCV*, 2019. 2, 5