Supplementary Materials

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

Contents

Figure 4. The map of the hand skeleton used in our HandVQA benchmark generation pipeline.

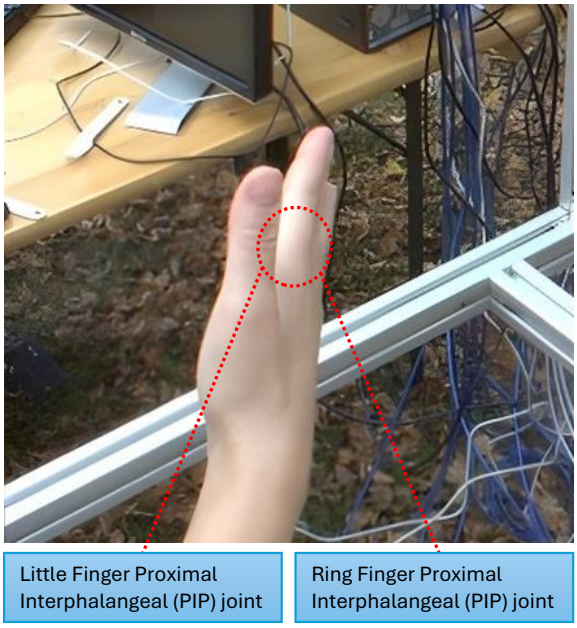


Figure 5. Possible location of the 'aligned' Little Finger Proximal Interphalangeal (PIP) joint and Ring Finger Proximal Interphalangeal (PIP) joint underneath the index and middle finger. The relationship along the x-axis for the two PIP joints is ambiguous, making it necessary to drop the relative position X information of the two joints.

6. Further HandVQA Pipeline Details

In this study, we employ the following abbreviations: carpometacarpal (CMC), metacarpophalangeal (MCP), interphalangeal (IP), proximal interphalangeal (PIP), and distal interphalangeal (DIP). Figure 4 illustrates the names and locations of the joints on the hand skeleton used in the generation pipeline for our HandVQA benchmark. Furthermore, Table 5 provides a comprehensive list of 107 joints and joint pairs for which pose descriptors are calculated within the benchmark generation pipeline.

6.1. Input to the HandVQA benchmark generation pipeline.

In this section, we formally explain the input representation used by the HandVQA benchmark generation pipeline and define how raw 3D joint coordinates are normalized.
Let the raw 3D hand joint coordinates be denoted by

$$P^{\text{raw}} = \{\mathbf{p}_i^{\text{raw}} \in \mathbb{R}^3 \,|\, i = 1,\ldots,21\}, \tag{4}$$

where each joint is given by $\mathbf{p}_i^{\text{raw}} = (x_i, y_i, z_i)$. The pipeline operates on a normalized version of these joints, producing

$$P = \{\mathbf{p}_i \in \mathbb{R}^3 \,|\, i = 1,\ldots,21\},$$

which is used throughout the HandVQA benchmark.

6.1.1. Datasets with hand meshes

For datasets providing full 3D hand meshes (*e.g.* FreiHAND [13], InterHand2.6M [9]), let

$$V^{\text{raw}} = \{\mathbf{v}_m^{\text{raw}} \in \mathbb{R}^3 \,|\, m = 1,\ldots,M\} \tag{5}$$

be the set of mesh vertices. We first center both vertices and joints using the mesh centroid

$$\mathbf{c} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{v}_m^{\text{raw}},$$

$$\tilde{\mathbf{v}}_m = \mathbf{v}_m^{\text{raw}} - \mathbf{c}, \qquad \tilde{\mathbf{p}}_i = \mathbf{p}_i^{\text{raw}} - \mathbf{c}.$$

Let

$$v_a^{\min} = \min_m \tilde{v}_{m,a}, \qquad v_a^{\max} = \max_m \tilde{v}_{m,a},$$

for each axis $a \in \{x, y, z\}$, and define the isotropic scale factor

$$s = \frac{1}{\max_a \left(v_a^{\max} - v_a^{\min}\right)}. \tag{6}$$

6.1.2. Datasets without hand meshes

Datasets such as FPHA [2] does not provide hand meshes, so we compute the normalization using only the joint coordinates. The centroid and centered joints are

$$\mathbf{c} = \frac{1}{21}\sum_{i=1}^{21} \mathbf{p}_i^{\text{raw}}, \qquad \tilde{\mathbf{p}}_i = \mathbf{p}_i^{\text{raw}} - \mathbf{c}.$$

Table 5. 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

For each axis $a \in \{x,y,z\}$, let

$$p_a^{\min} = \min_i \tilde{p}_{i,a}, \qquad p_a^{\max} = \max_i \tilde{p}_{i,a},$$

and define

$$s = \frac{1}{\max_a \left(p_a^{\max} - p_a^{\min}\right)}. \tag{7}$$

The normalized joints are then

$$\mathbf{p}_i = s\tilde{\mathbf{p}}_i. \tag{8}$$

This centering and isotropic scaling yields a normalized pose

$$P = \{\mathbf{p}_i \in \mathbb{R}^3 \mid i = 1,...,21\}$$

that serves as the input to the pose descriptor extraction module $\mathcal{F}_{\text{pose}}$ in the HandVQA pipeline.

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

In Figure 5, 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 5 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.

6.3. Sampling Details of Pose Descriptors and MCQs

We describe here the sampling strategy used to generate a tractable yet comprehensive set of multiple-choice questions (MCQs) for each image. From the discrete pose descriptors produced by $\mathcal{F}_{\text{pose}}$, only descriptors corresponding to joints or joint pairs included in Table 5 are considered for MCQ construction in HandVQA for scalibility. We define them as $\mathcal{T}$.

Angle descriptors. All anatomically valid finger-joint bending angles are used. Let

$$\mathcal{J}_{\text{angle}} = \{j_1, \ldots, j_{15}\}$$

be the set of the 15 joints for which a bending angle is defined. For each $j \in \mathcal{J}_{\text{angle}}$, $\mathcal{F}_{\text{text}}$ generates four candidate sentences, one for each category label, and the correct sentence is selected.

Distance and relative-position descriptors. For distance and relative position along each axis (x, y, z), the total set of possible unordered joint pairs for each pose descriptor is

$$\binom{21}{2} = 210.$$

However, many of these pairs are either anatomically implausible, rarely interacting with eachother in our daily activities, or redundant for fine-grained description. To ensure scalability, we restrict attention to the subset

$$\mathcal{J}_{\text{pair}} \subseteq \{(i,k) \mid 1 \leq i < k \leq 21\}$$

consisting only of joint pairs on *adjacent fingers*. The exception is the thumb, which is permitted to compare with all other fingers because of its distinct opposable role in interacting with other fingers. This restriction yields a significantly smaller and semantically meaningful subset of pairs (shown in Table 5) for which $\mathcal{F}_{\text{text}}$ generates four candidate sentences, and the correct option.

Sampling. To maintain a manageable dataset size while ensuring descriptor diversity, we sample a fixed number of MCQs per image. Specifically, for each descriptor type *Angle*, *Distance*, and *Relative Position X, Y, Z* we uniformly sample 5 distinct descriptor instances from the eligible joint or joint-pairs defined in $\mathcal{T}$ (demonstrated in Table 5). Each sampled element yields exactly one MCQ consisting of the prompt sentence and the option set $\mathcal{O}$ with a single correct answer. Therefore, each image results in 25 MCQs covering all five pose descriptor families.

Extensibility. Although the released benchmark samples from a reduced and anatomically meaningful subset $\mathcal{J}_{\text{pair}}$, it is possible to generate arbitrarily many MCQs per image within the limits of valid hand anatomy and application-specific needs by simple modifications.

7. Dataset Statistics

The details of each dataset and statistics of the benchmark are discussed in this section.

7.1. Dataset Construction

We use three hand datasets to construct our HandVQA benchmark: FreiHAND [13], InterHand2.6M [9], and FPHA [2]. Further details of them are discussed below,

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.

7.2. Balanced Coverage of Spatial Reasoning Tasks

The Figure 7 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 [2], FreiHAND [13], and InterHand2.6M [9].

In both the training (left) and evaluation (right) plots (as shown in Fig. 7), 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.

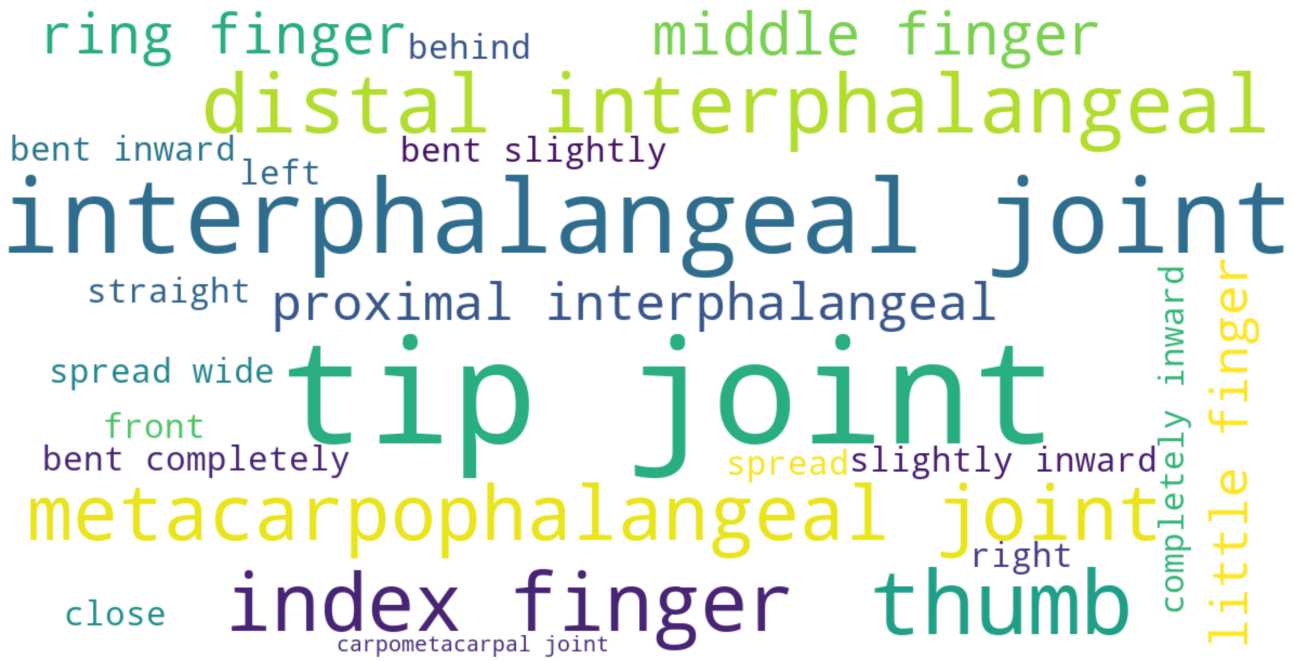


Figure 6. **Word cloud representation of the most frequently used terms in the caption options extracted from our dataset.** Prominent anatomical terms like "tip joint", "interphalangeal joint", and "metacarpophalangeal joint" highlight the fine-grained spatial and anatomical focus of the hand-centric question-answer pairs.

7.3. Word Cloud Analysis of Pose Descriptors

Figure 6 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.

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 5, 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 5.
- **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 and meaningful joint relationships.

Table 6. Label frequency grouped by pose descriptors.

Pose Descriptors	Label frequency (Count)
Angle	bent slightly inward (130,993), bent inward (107,143), straight (74,007), bent completely inward (21,622)
Distance	spread wide from (222,429), spread from (109,353), close to (1,983)
Rel. Pos. (X)	at the left of (198,988), at the right of (133,109)
Rel. Pos. (Y)	above (215,725), below (112,984)
Rel. Pos. (Z)	in front of (177,216), behind (152,481)

7.4. Detailed Category Label Statistics

We further analyze the distribution of category labels within each pose descriptor family to better understand the underlying data characteristics. Table 6 summarizes the frequency of all discrete labels used across angle, distance, and relative position descriptors.

For angle descriptors, the distribution is skewed toward mid-range articulation states, with *bent slightly inward* (130,993) and *bent inward* (107,143) dominating the dataset, while extreme configurations such as *bent completely inward* (21,622) are comparatively underrepresented. The *straight* category (74,007) occupies an intermediate proportion.

For distance descriptors, the dataset is heavily biased toward larger separations between joints. The label *spread*

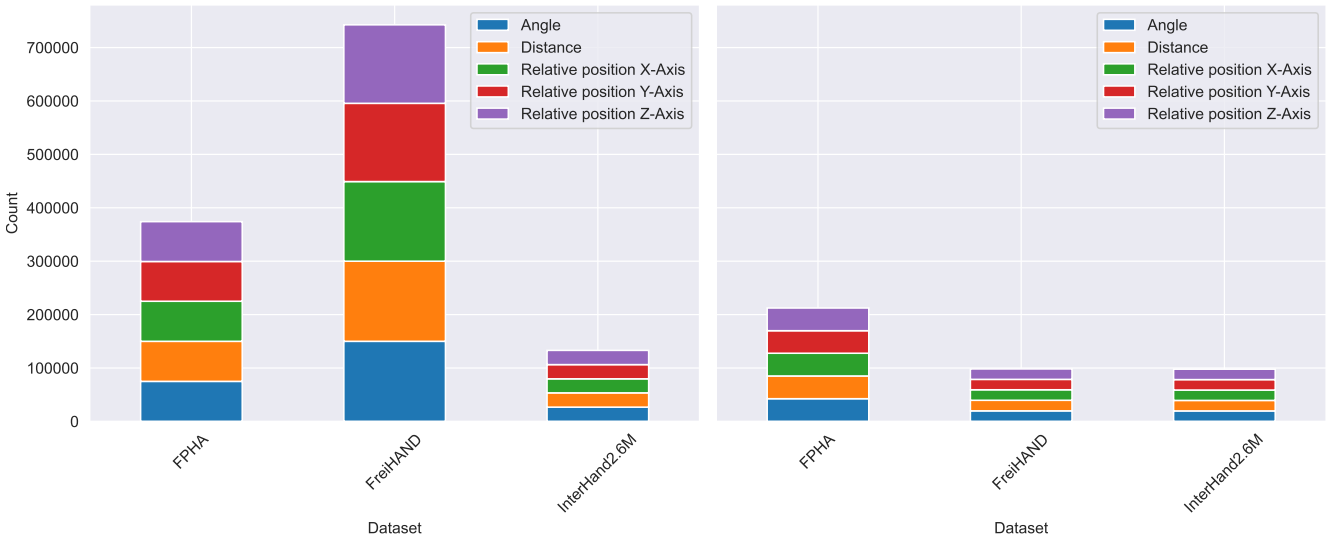


Figure 7. **Breakdown of question types across the training (left) and evaluation (right) splits for each dataset in the HandVQA benchmark.** Each dataset contains a balanced distribution of all five spatial reasoning tasks: angle, distance, and relative positions along the X, Y, and Z axes. This uniformity supports fair evaluation across all pose subtasks.

wide from (222,429) is the most frequent, followed by *spread from* (109,353), whereas *close to* (1,983).

In the case of relative position descriptors, the distributions are relatively balanced but still exhibit mild asymmetries. Along the X-axis, *at the left of* (198,988) appears more frequently than *at the right of* (133,109). Similarly, for the Y-axis, *above* (215,725) is more common than *below* (112,984). Along the Z-axis, the distribution between *in front of* (177,216) and *behind* (152,481) is comparatively more balanced, though still slightly skewed.

8. Training Details and Further Analysis on Experiments

This section provides a comprehensive overview of our training configuration alongside a series of additional analyses that expand upon the main experimental findings. We examine model behavior through confusion matrices, investigate cross-dataset transferability, study the effect of dataset-specific finetuning, compare VLM outputs with human judgments, and describe the construction of zero-shot evaluation datasets.

8.1. Training Setup

We finetune all VLMs using LoRA[4] 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 [12] package to finetune all VLMs.

8.2. Behavior of VLMs Across Spatial Descriptors

Figures 8, 9, 10, 11, 12, and 13 present confusion matrices for both base and fine-tuned VLMs, including DeepSeek [8], LLaVA [7], and Qwen-VL [1]. The confusion matrices are constructed from the evaluation sets of all three datasets combined, enabling a direct comparison of model behavior before and after fine-tuning.

8.2.1. Angle

In Figure 8, we present the confusion matrix for the angle pose descriptor across four 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, 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 accurately dif-

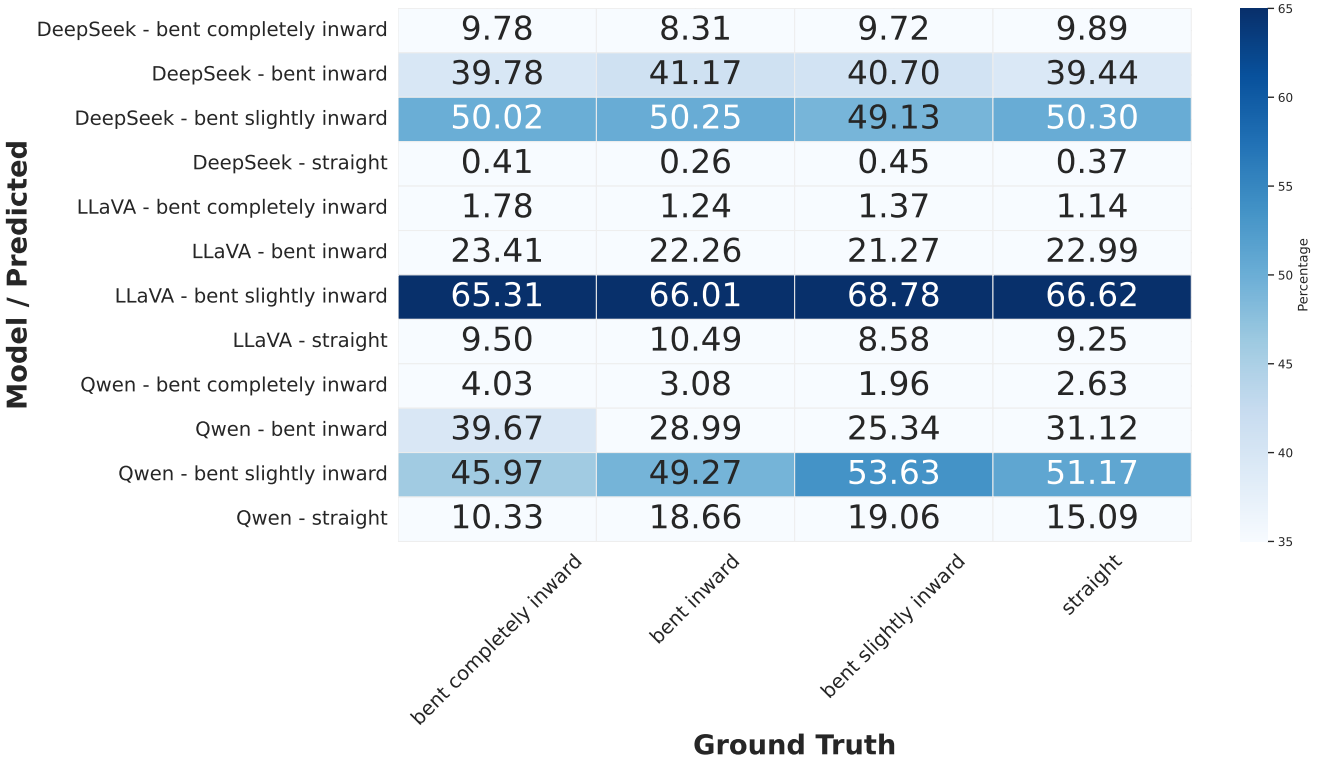


Figure 8. **Angle confusion matrix across three VLMs.** All models frequently predict *"bent slightly inward"* regardless of ground truth, revealing a strong prediction bias. Each model also follows a consistent preference ordering in its outputs, indicating difficulty in distinguishing fine-grained joint angles.

ferentiate 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.

After fine-tuning, this bias is substantially reduced as shown in Figure 9. The prediction distribution becomes more balanced, with a stronger concentration along the diagonal of the confusion matrix. This indicates improved discrimination of fine-grained joint angles. However, confusion between adjacent angle categories remains, suggesting that subtle variations in articulation are still challenging.

8.2.2. Distance

In Figure 10a, 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 the ground truth distance label. This tendency is especially pronounced in LLaVA, 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.

Following fine-tuning, the prediction distributions become significantly more balanced, with improved alignment along the diagonal (Figure 10b). Models demonstrate better discrimination between different levels of inter-joint distance. Nevertheless, residual confusion persists between "spread from" and "spread wide from", indicating that fine-grained distance reasoning remains difficult.

8.2.3. Relative Positions (X, Y and Z-axis)

The confusion matrices for the relative position tasks along the X, Y, and Z axes (Figures 11a, 12a, and 13a) 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.

For the Y-axis, LLaVA overpredicts "above", while others

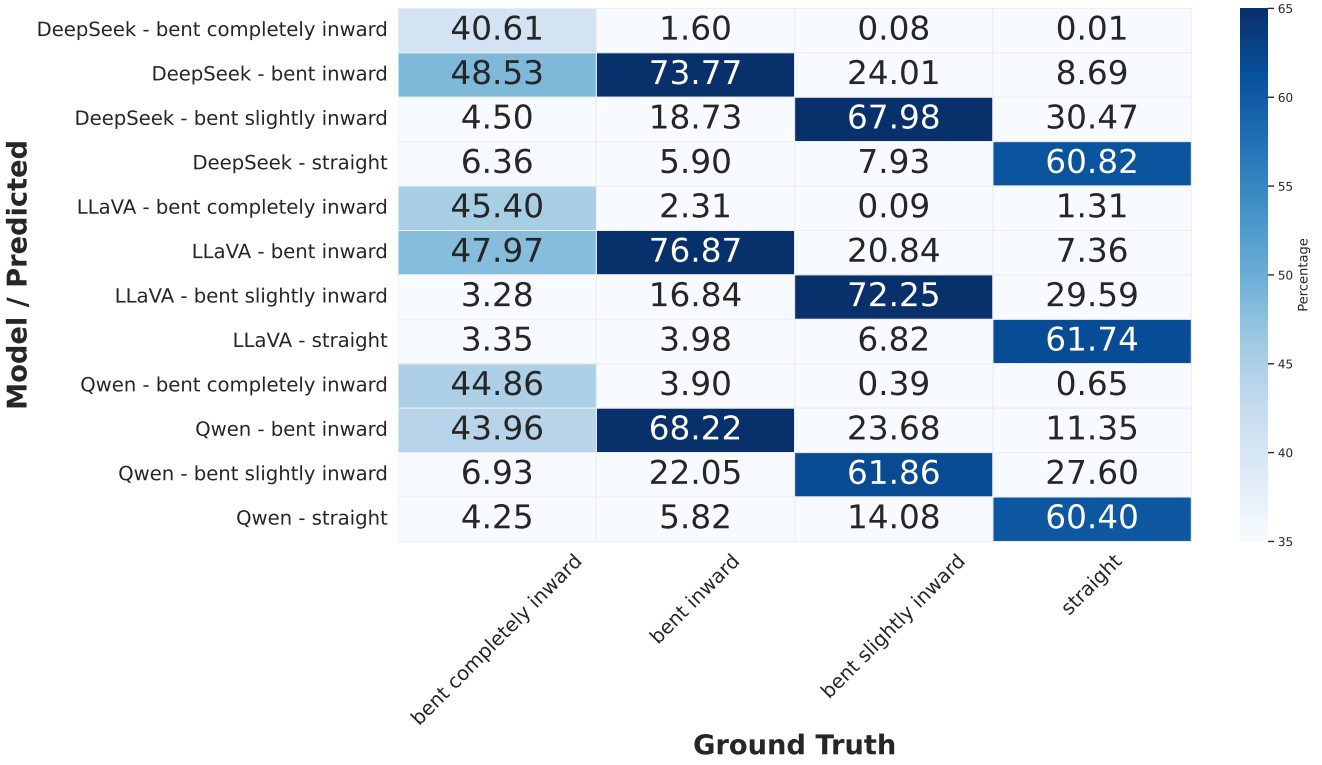
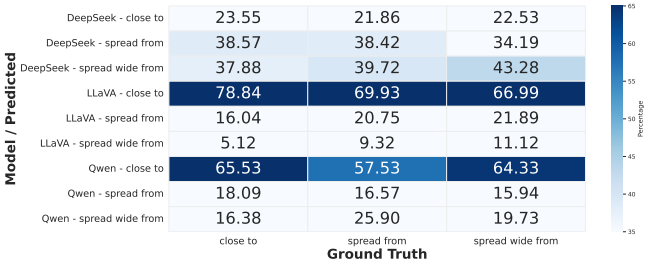
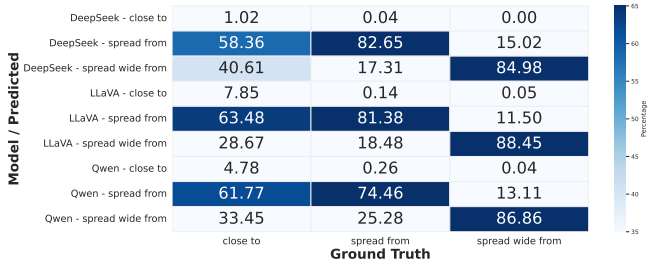


Figure 9. **Angle confusion matrix across three fine-tuned VLMs.** Fine-tuning significantly reduces the dominant bias toward *"bent slightly inward"* observed in the base models, resulting in stronger alignment along the diagonal. The models exhibit improved discrimination across angle categories, particularly for *"bent inward"* and *"straight"*, although some residual confusion remains between adjacent angle classes, indicating that fine-grained distinctions are still challenging.



(a) Base Models.

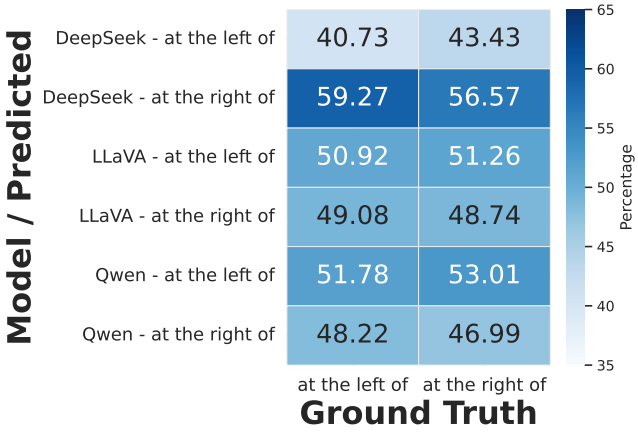


(b) Finetuned Models.

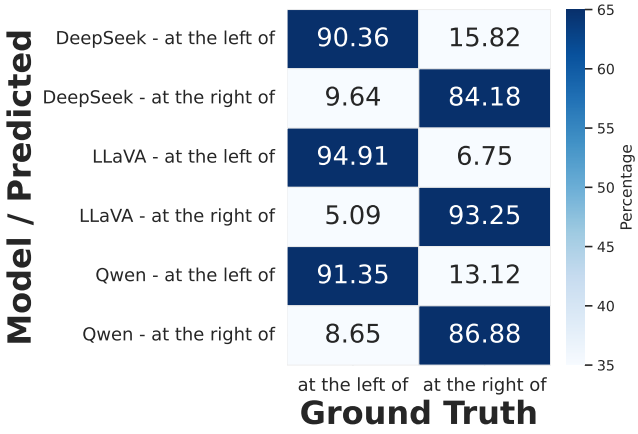
Figure 10. **Distance confusion matrix comparison between base and fine-tuned VLMs.** The base models exhibit a strong bias toward predicting *"close to"* across multiple ground-truth categories, leading to a skewed prediction distribution. In contrast, fine-tuning produces a more balanced distribution with increased alignment along the diagonal, indicating improved discrimination between distance categories. However, residual confusion persists between *"spread from"* and *"spread wide from"*, suggesting continued difficulty in distinguishing fine-grained spatial separations.

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."

In contrast, fine-tuning leads to a pronounced concentration along the diagonal across all axes, indicating substantial improvement in directional understanding (Figures 11b, 12b, and 13b). The models learn to reliably distinguish binary spatial relations such as left–right, above–below, and front–behind. While minor confusion remains, particularly in depth (Z-axis), the overall spatial grounding is significantly enhanced.

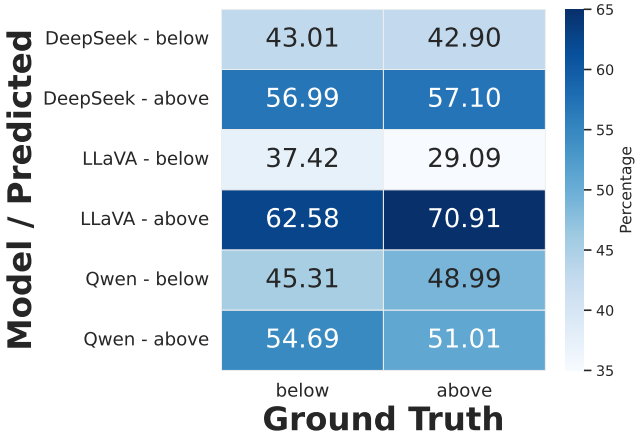


(a) Base models.

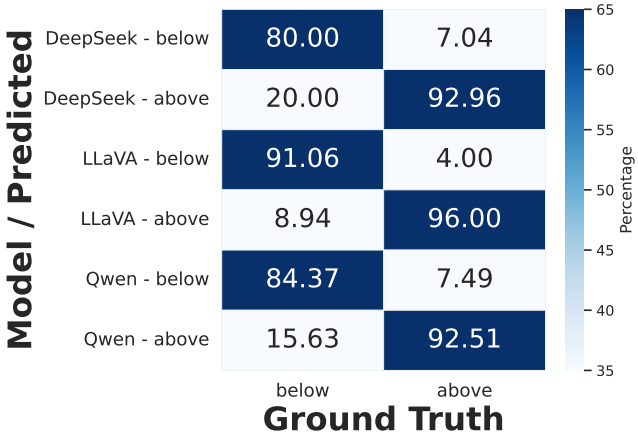


(b) Finetuned models.

Figure 11. **Relative Position X confusion matrix comparison between base and fine-tuned VLMs.** The base models exhibit near-random predictions between *"at the left of"* and *"at the right of"*, indicating a lack of reliable directional understanding. In contrast, fine-tuning leads to a strong concentration along the diagonal, demonstrating significantly improved left–right discrimination. This suggests that relative positional reasoning along the horizontal axis is effectively learned after fine-tuning, with minimal residual confusion.



(a) Base models.



(b) Finetuned models.

Figure 12. **Relative Position Y confusion matrix comparison between base and fine-tuned VLMs.** The base models exhibit inconsistent predictions between *"below"* and *"above"*, reflecting weak vertical positional understanding. In contrast, fine-tuning results in a strong alignment along the diagonal, indicating substantially improved discrimination of vertical relationships. While minor confusion remains, the fine-tuned models demonstrate a clear and consistent understanding of relative position along the vertical axis.

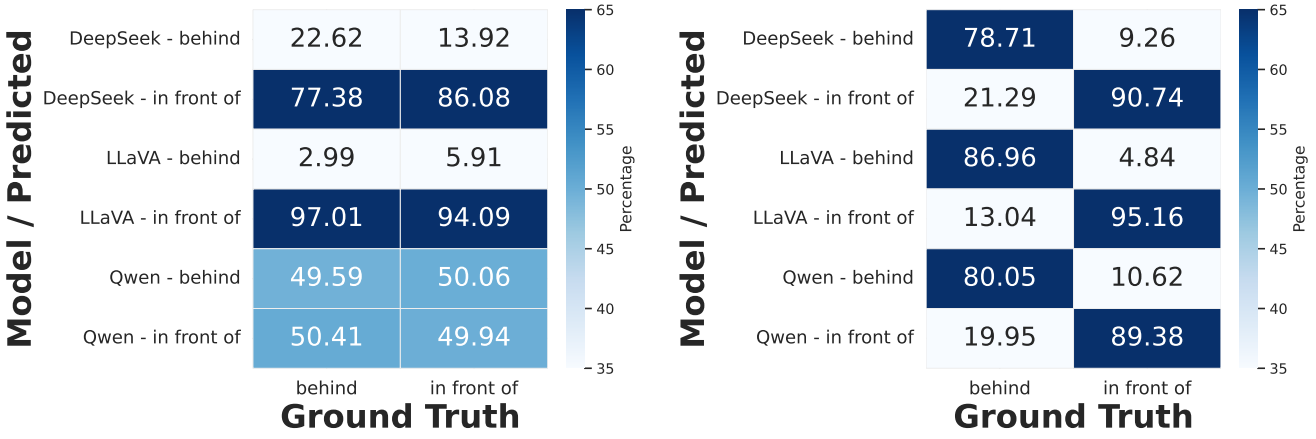
8.3. Further Experimental Comparisons

We further investigate how finetuning strategies and dataset characteristics influence cross-dataset generalization and overall spatial reasoning performance of VLMs.

8.3.1. Finetuned Models on Individual Datasets vs. Training on the Unified HandVQA Benchmark

To better understand the impact of large-scale, multi-dataset supervision, we compare LLaVA Mistral 7B models finetuned on each dataset individually (FreiHAND, InterHand2.6M, FPHA) against a single model trained on the full HandVQA benchmark, which unifies all three datasets into a large and diverse supervision signal. We focus on LLaVA Mistral 7B because it was the strongest base model in our benchmark and exhibited the most stable finetuning behavior. Overall, the HandVQA-tuned model substantially outperforms the base model by a large margin across all datasets and all task types: Angle, Distance, and Relative Position reasoning (Tables 7 and 8). This confirms that large-scale multimodal supervision with structured spatial queries transfers effectively and imparts generalizable spatial reasoning abilities that individual datasets alone cannot provide.

However, we also observe that the HandVQA-trained model does not always surpass the models finetuned on



(a) Base models.

(b) Finetuned models.

Figure 13. **Relative Position Z confusion matrix comparison between base and fine-tuned VLMs.** The base models exhibit strong bias toward predicting *"in front of"*, leading to highly skewed outputs and poor discrimination of depth relationships. In contrast, fine-tuning significantly improves alignment with the diagonal, indicating enhanced understanding of front–back spatial relations. While minor confusion persists, the fine-tuned models demonstrate substantially improved depth-aware reasoning compared to the base models.

Table 7. Angle and Distance results of LLaVA Mistral 7B finetuned models on Individual Datasets vs. finetuned on the Unified HandVQA Benchmark.

Model	Tuned	Eval	Angle		Distance	
			Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
Finetuned Models on Individual Datasets						
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
Finetuned Model on HandVQA						
LLaVA Mistral 7B	HandVQA	InterHand2.6M	72.26	0.283	90.27	0.099
LLaVA Mistral 7B	HandVQA	FreiHAND	64.35	0.367	86.71	0.136
LLaVA Mistral 7B	HandVQA	FPHA	67.95	0.411	84.64	0.154

Table 8. Relative Position accuracies of LLaVA Mistral 7B finetuned models on Individual Datasets vs. finetuned on the Unified HandVQA Benchmark.

Model	Tuned	Eval	Rel. Pos. X	Rel. Pos. Y	Rel. Pos. Z
			Accuracy ↑	Accuracy ↑	Accuracy ↑
Finetuned Models on Individual Datasets					
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
Finetuned Model on HandVQA					
LLaVA Mistral 7B	HandVQA	InterHand2.6M	96.15	98.66	96.74
LLaVA Mistral 7B	HandVQA	FreiHAND	93.82	94.34	90.24
LLaVA Mistral 7B	HandVQA	FPHA	95.12	93.26	90.30

each dataset individually, especially in tasks where dataset specific characteristics dominate performance. For example, in FreiHAND and FPHA, the individually finetuned models achieve slightly higher accuracy or lower MAE on certain attributes. We attribute this phenomenon to the intrinsic limitations of LoRA finetuning. LoRA constrains parameter

updates to a low rank decomposition, meaning that the model must compress all newly learned information across three large and heterogeneous datasets into a small number of trainable low rank matrices. As the diversity of training signals grows, these matrices must encode more variations, object contexts, hand configurations, and camera setups. This compression inevitably introduces information bottlenecks, leading to mild degradation on certain dataset specific details compared to models finetuned exclusively on a single, homogeneous dataset.

8.3.2. Base Models vs. Cross-Dataset Performance

To evaluate cross-dataset transfer, we compare base VLMs against models finetuned on the FreiHAND dataset and test them on two out-of-distribution datasets: InterHand2.6M (allocentric, multiview) and FPHA (egocentric, first-person). FreiHAND is predominantly allocentric, and therefore provides a controlled setup for studying how allocentric supervision transfers to both similar (allocentric) and different (egocentric) camera viewpoints.

As shown in Tables 9 and 10 across all descriptors Angle, Distance, and Relative Position FreiHAND finetuned models show substantial improvement when evaluated on InterHand2.6M. This trend is consistent across all three architectures (DeepSeek Janus Pro 7B, LLaVA Mistral 7B, and Qwen2.5 VL 7B Instr.), confirming that allocentric-to-allocentric transfer is highly effective. The improved performance suggests that the spatial reasoning learned from FreiHAND generalizes well to another multiview, third-person dataset that shares similar camera geometry and viewpoint distribution. However, the same finetuned models show less consistent gains when evaluated on the egocentric FPHA dataset. While Distance and Relative Position Y exhibit large improvements—likely because these descriptors depend more on coarse spatial relations rather than precise articulation—other attributes such as Angle and Relative Position Z remain challenging. These findings indicate that allocentric training alone does not fully prepare the model for the viewpoint distortions and hand-camera proximities inherent to egocentric perspectives.

8.3.3. Comparison with Pure Vision Models

To assess the extent to which explicit multimodal supervision contributes to fine-grained spatial reasoning, we compare vision-language models against a strong pure vision baseline, HaMeR [10]. HaMeR directly predicts 3D hand meshes from images, using our pipeline that was later converted into text description and evaluated against ground truth. The comparison is reported in Table 11, covering Angle and Distance prediction tasks on the FreiHAND and InterHand2.6M test sets. The FPHA dataset is excluded, as it is not part of HaMeR's training data.

We first observe that the base LLaVA model performs significantly worse than HaMeR across all settings. For instance, on FreiHAND, LLaVA achieves only 42.48% accuracy for Angle and 13.18% for Distance, compared to 59.53% and 88.86% respectively, for HaMeR. A similar trend holds on InterHand2.6M. This gap highlights that, without task-specific supervision, VLMs lack the ability to extract precise geometric cues required for accurate hand pose reasoning.

After finetuning on the HandVQA benchmark, LLaVA shows substantial improvements across all metrics, outperforming HaMeR on several tasks. In particular, LLaVA *ft* achieves higher Angle accuracy on FreiHAND (64.35% vs. 59.53%) and significantly improves Distance prediction on InterHand2.6M (90.27% vs. 88.11%). These results demonstrate that structured multimodal supervision can compensate for the lack of explicit geometric inductive bias in VLMs and enable them to learn fine-grained spatial relationships directly from data.

8.3.4. Model Confidence and Uncertainty Analysis

We further analyze the reliability of model predictions by studying the relationship between predicted confidence and empirical accuracy.

We initially explored estimating confidence via an additional classification head over discrete angle bins, jointly trained with the VLM. However, we observed inconsistent behavior between the auxiliary head outputs and the textual predictions, indicating a lack of alignment between internal representations and generated answers. Due to this inconsistency, we instead adopt a prompting-based approach for confidence estimation, following prior work on verbalized uncertainty in VLMs [3].

Specifically, we prompt the LLaVA Mistral 7B to output likelihoods over discrete answer categories and evaluate calibration using reliability diagrams and confidence density histograms. This setup enables direct comparison between predicted confidence and empirical accuracy.

As shown in Fig. 14, the base model exhibits a skewed confidence distribution, where a significant portion of incorrect predictions are associated with high confidence. This is further confirmed in the reliability diagram (Fig. 15), where predictions deviate substantially from the diagonal, indicating strong overconfidence. In other words, the model assigns high confidence even when accuracy is low.

After finetuning, the prediction distribution shifts noticeably (Fig. 16). While incorrect high-confidence predictions are reduced, the overall confidence mass moves toward lower and mid-range values. The corresponding reliability diagram (Fig. 17) shows that confidence aligns more closely with accuracy, indicating improved calibration compared to the base model.

However, a distinct behavior emerges: despite higher accuracy, the finetuned model exhibits systematically lower confidence. As observed in Fig. 17, most points lie below the diagonal, suggesting a tendency toward underconfidence. That is, even correct predictions are often assigned conservative confidence scores.

Table 9. Angle and Distance results of Base Models vs Out-of-Distribution Finetuned Models.

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	–	FPHA	26.46	0.991	39.02	0.819
Finetuned Models						
DeepSeek Janus Pro 7B	FreiHAND	InterHand2.6M	56.15	0.456	82.70	0.175
DeepSeek Janus Pro 7B	FreiHAND	FPHA	25.52	1.028	79.73	0.203
Base model (no tuning)						
LLaVA Mistral 7B	–	InterHand2.6M	40.08	0.739	16.20	1.293
LLaVA Mistral 7B	–	FPHA	23.38	1.011	13.57	1.353
Finetuned Models						
LLaVA Mistral 7B	FreiHAND	InterHand2.6M	56.81	0.447	83.26	0.170
LLaVA Mistral 7B	FreiHAND	FPHA	26.45	1.015	79.94	0.201
Base model (no tuning)						
Qwen 2.5 VL 7B Instr.	–	InterHand2.6M	37.92	0.779	19.58	1.247
Qwen 2.5 VL 7B Instr.	–	FPHA	24.22	1.055	18.03	1.306
Finetuned Models						
Qwen 2.5 VL 7B Instr.	FreiHAND	InterHand2.6M	50.85	0.536	80.67	0.196
Qwen 2.5 VL 7B Instr.	FreiHAND	FPHA	24.65	1.083	78.94	0.211

Table 10. Relative-Position Results of Base Models vs Out-of-Distribution Finetuned Models.

Model	Tuned	Eval	Rel. Pos. X	Rel. Pos. Y	Rel. Pos. Z
			Accuracy ↑	Accuracy ↑	Accuracy ↑
Base model (no tuning)					
DeepSeek Janus Pro 7B	–	InterHand2.6M	50.41	52.46	51.16
DeepSeek Janus Pro 7B	–	FPHA	43.02	52.64	61.73
Finetuned Models					
DeepSeek Janus Pro 7B	FreiHAND	InterHand2.6M	77.32	89.80	75.33
DeepSeek Janus Pro 7B	FreiHAND	FPHA	50.58	74.96	48.45
Base model (no tuning)					
LLaVA Mistral 7B	–	InterHand2.6M	49.72	66.26	40.87
LLaVA Mistral 7B	–	FPHA	50.27	56.33	56.73
Finetuned Models					
LLaVA Mistral 7B	FreiHAND	InterHand2.6M	84.53	92.73	84.49
LLaVA Mistral 7B	FreiHAND	FPHA	50.27	78.04	56.65
Base model (no tuning)					
Qwen 2.5 VL 7B Instr.	–	InterHand2.6M	48.98	49.78	49.33
Qwen 2.5 VL 7B Instr.	–	FPHA	50.98	48.53	49.79
Finetuned Models					
Qwen 2.5 VL 7B Instr.	FreiHAND	InterHand2.6M	77.61	86.27	66.98
Qwen 2.5 VL 7B Instr.	FreiHAND	FPHA	55.82	70.34	60.07

Overall, these results indicate that base VLMs suffer from overconfident errors, while finetuning reduces such failures but introduces underconfidence. This highlights that improved accuracy does not directly translate to well-calibrated uncertainty, and that reliable confidence estimation remains an open challenge even after task-specific supervision.

8.3.5. Failure Mode Analysis

We additionally analyze failure cases of the finetuned model by separating samples into *Easy* and *Hard* subsets based

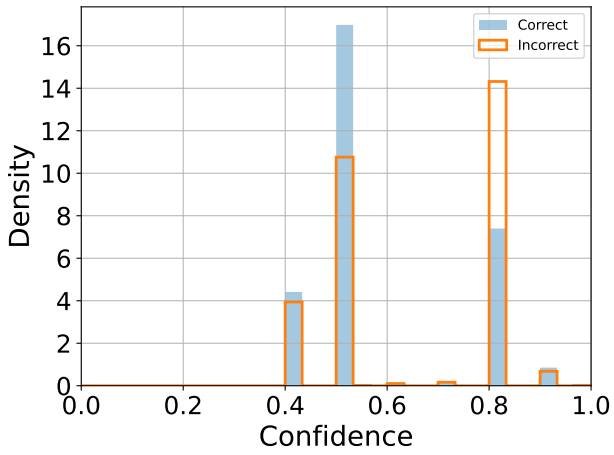


Figure 14. **Base Model Prediction Distribution (LLaVA-Mistral-7B).** Distribution of predicted confidence for correct and incorrect responses in the base LLaVA-Mistral-7B model. The confidence distribution is skewed, with a substantial portion of incorrect predictions assigned high confidence, revealing pronounced overconfident errors.

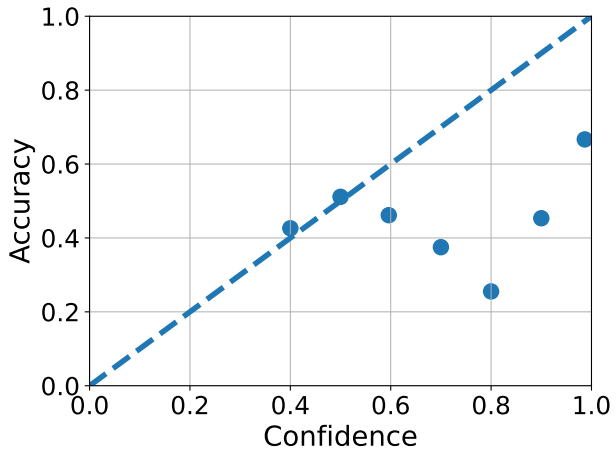


Figure 15. **Base Model Reliability Diagram (LLaVA-Mistral-7B).** Reliability diagram comparing predicted confidence and accuracy. The model is poorly calibrated and overconfident. Although accuracy is higher at confidence > 0.8, such predictions are rare (Fig. 14), limiting their impact.

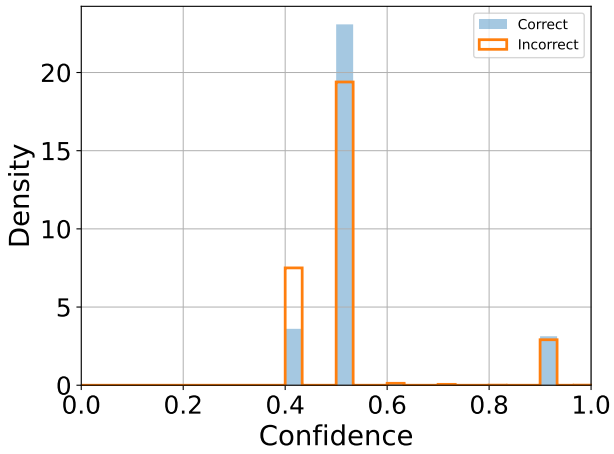


Figure 16. **Finetuned Model Prediction Distribution (LLaVA-Mistral-7B).** Distribution of predicted confidence for correct and incorrect responses after finetuning LLaVA-Mistral-7B. High-confidence incorrect predictions are reduced, and the confidence mass shifts toward lower and mid-range values, reflecting more conservative predictions.

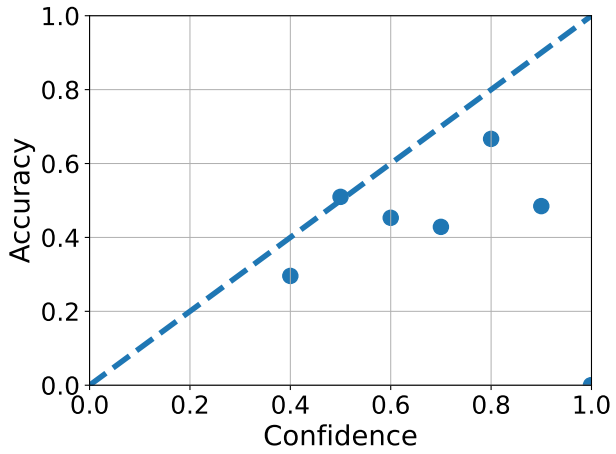


Figure 17. **Finetuned Model Reliability Diagram (LLaVA-Mistral-7B).** Reliability diagram of the finetuned LLaVA-Mistral-7B model. Confidence aligns more closely with empirical accuracy, indicating improved calibration. However, most points lie below the diagonal, suggesting systematic underconfidence despite higher accuracy.

on ambiguity factors such as occlusion and interaction complexity. Quantitative results are summarized in Table 12, and representative qualitative examples are shown in Fig. 18.

As shown in Table 12, LLaVA Mistral 7B *finetuned* achieves strong performance on Easy samples across all descriptors (e.g., 75.92% for Angle and 95.01% for Distance), but exhibits a consistent drop on Hard samples (55.1% for Angle and 78.01% for Distance). Similar degradation is observed across all relative position axes. This gap indicates that performance is primarily affected by visual ambiguity rather than uniformly limited across all samples.

Fig. 18 provides qualitative examples illustrating this behavior. In Easy cases (top row), the model correctly predicts both single-hand and hand-object interactions, where the spatial configuration is clearly visible and minimally occluded. In contrast, Hard cases (bottom

Table 11. Pure vision model (HaMeR [10]) vs. LLaVA Mistral 7B.

Model	FreiHAND test		InterHand2.6M test	
	Angle Accuracy % (MAE)	Distance Accuracy % (MAE)	Angle Accuracy % (MAE)	Distance Accuracy % (MAE)
HaMeR	59.53 (0.428)	**88.86 (0.113)**	**78.82 (0.218)**	88.11 (0.119)
LLaVA Mistral 7B	42.48 (0.678)	13.18 (1.342)	40.08 (0.739)	16.20 (1.293)
LLaVA Mistral 7B *finetuned*	**64.35 (0.367)**	86.71 (0.136)	72.26 (0.283)	**90.27 (0.099)**

Table 12. LLaVA Mistral 7B *finetuned* on FreiHAND: Easy (225 QA), Hard (223 QA).

Difficulty	Angle Accuracy % (MAE)	Distance Accuracy % (MAE)	R. Pos.(X) Accuracy %	R. Pos.(Y) Accuracy %	R. Pos.(Z) Accuracy %
Easy	**75.92 (0.255)**	**95.01 (0.08)**	**97.20**	**96.19**	**98.11**
Hard	55.1 (0.454)	78.01 (0.183)	82.11	81.72	78.92

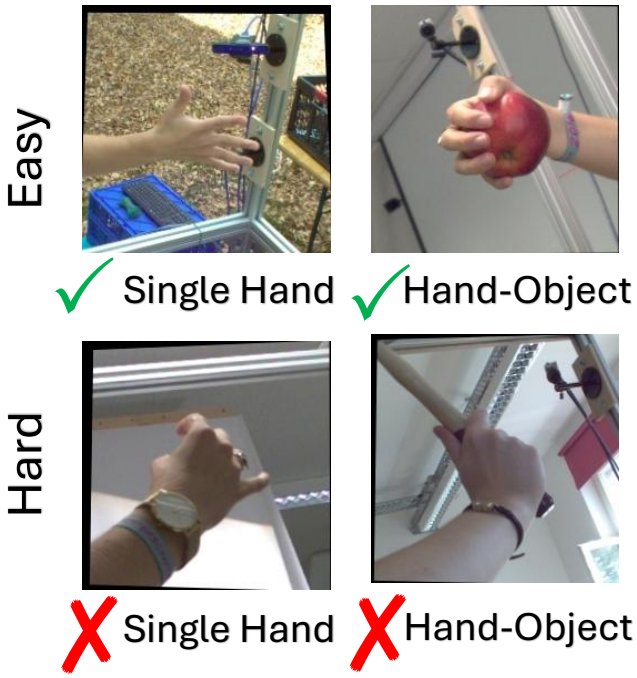


Figure 18. Easy vs Hard Examples Performance of LLaVA Mistral 7B *finetuned*. Correct(✓)/wrong(X) QA samples for "Which is the distance between thumb tip and index fingertip?".

row) show failure examples where occlusion, viewpoint, or interaction complexity makes the spatial relationship ambiguous, leading to incorrect predictions.

8.4. Human Evaluation

To understand how current VLMs compare to human spatial reasoning, we conducted a small scale human study on a subset of HandVQA. As shown in Table 13, humans achieve 80.94% accuracy, significantly outperforming all evaluated VLMs, which score between 41–46%. This substantial gap

Table 13. Human vs VLMs Accuracy on a small subset of HandVQA.

Models	Overall Accuracy
LLaVA Mistral 7B	41.96%
DeepSeek Janus Pro 7B	45.82%
Qwen 2.5 VL 7B	41.97%
Humans	80.94%

highlights the difficulty of fine-grained hand pose reasoning, even for the strongest models such as LLaVA Mistral 7B and DeepSeek Janus Pro 7B. While VLMs can interpret coarse spatial relations, they frequently struggle with subtle joint-level distinctions such as small angular differences or depth ordering that humans can reliably discern. These results underscore the importance of specialized datasets like HandVQA for pushing VLMs toward human level performance in fine-grained 3D hand understanding.

8.5. Zero-shot Evaluation Dataset Construction

For our zero-shot evaluation, we consider two tasks: image-based gesture recognition and temporal sequence-based hand–object interaction recognition. The gesture recognition task is built from the HaGRID dataset [5], while the interaction recognition task uses the H2O dataset [6].

For the H2O interaction task, we directly use the action annotations provided in the dataset and convert them into multiple-choice questions (MCQs), each containing one correct answer and three distractors. In contrast, HaGRID provides only single-word gesture labels, which are insufficiently descriptive for zero-shot evaluation. To address this, we expand each gesture label into two semantically rich natural-language descriptions using Gemini [11], as illustrated in Table 14. These expanded descriptions provide the linguistic diversity necessary for evaluating zero-shot gesture reasoning. After generating gesture descriptions,

we construct MCQs for each image, again with one correct description and three incorrect alternatives. To avoid semantic ambiguity, we group visually and semantically similar gestures such as `two_up` and `two_up_inverted` into the same category. Consequently, when one appears as the correct option, the other is never used as an incorrect distractor. This ensures that evaluation difficulty arises from genuine reasoning challenges rather than annotation artifacts or label confusion.

9. Qualitative Results

Figure 19 and Figure 20 present qualitative zero-shot results on the HaGRID and H2O datasets, respectively. In the HaGRID gesture recognition task, the fine-tuned model consistently selects the correct semantic description, despite never having been trained on gesture sentences or gesture-specific supervision. The base model, by contrast, frequently misidentifies even visually obvious gestures. A similar trend appears in the H2O interaction recognition task: when reasoning over short temporal sequences, the fine-tuned model accurately identifies subtle object manipulations while the base model often predicts incorrect actions, demonstrating that joint-level training on HandVQA yields transferable spatial reasoning even in multi-frame, object-centric scenarios.

Figure 21, Figure 22, and Figure 23 provide qualitative comparisons on FreiHAND, InterHand2.6M, and FPHA. Each figure shows HandVQA-style MCQs spanning all five pose descriptors, together with predictions from both the base LLaVA model and its HandVQA-fine-tuned counterpart. Across all datasets, the base model most of the time chooses incorrect answers, including in cases involving clear geometry or simple articulation patterns. The fine-tuned model, however, consistently resolves the correct spatial relation, highlighting its ability to interpret fine-grained joint positions, bending angles, distances, and relative spatial orientations.

Finally, Figure 24 examines generalization to in-the-wild images paired with questions phrased differently from the HandVQA templates and targeting higher-level finger-level geometry. The fine-tuned model reliably interprets these queries while the base model fails. These examples demonstrate that fine-grained joint-level supervision not only improves in-domain performance but also enables robust transfer to higher-level geometric reasoning.

10. License Details of Source Datasets

HandVQA is constructed entirely from three existing and publicly available 3D hand datasets: FreiHAND [13], InterHand2.6M [9], and FPHA [2]. Each dataset is properly cited in the main paper and used in accordance with its respective license and terms of use.

- **FreiHAND** is released strictly for research purposes only. Any commercial use is explicitly prohibited, and users are required to cite the original paper if the dataset or parts of it are used.
- **InterHand2.6M** is distributed under the CC-BY-NC 4.0 license, which permits use for non-commercial research with appropriate credit to the original authors.
- **FPHA** is available for free for academic research and non-commercial use.

We have adhered to the terms and conditions of each dataset as per their official distribution policies. This ensures that all licenses are respected in full, and no proprietary or restricted-use data is included in HandVQA.

Table 14. Gemini-generated descriptions for HaGRID gesture labels.

Base Label	Description 1	Description 2
one	A single index finger is extended, representing the number one.	The hand is held up with only the index finger pointing upwards.
two_up	The index and middle fingers are extended upwards (peace sign), representing two.	A “V” sign, often used for “two” or “peace”.
two_up_inverted	An inverted “V” sign, with the palm facing inward.	The gesture for “two” or “peace” but inverted.
three	Three fingers are extended to represent the number three.	The hand gesture indicates a quantity of three.
four	Four fingers are extended to show the number four.	The hand gesture indicates a quantity of four.
fist	The hand is closed tightly into a fist.	All fingers are curled inward with the thumb wrapped around them.
palm	The hand is open with the palm facing forward, showing five fingers.	An open palm, often representing the number five.
ok	The thumb and index finger form a circle to signify “OK”.	A hand gesture indicating that everything is alright.
peace	The index and middle fingers form a “V” to symbolize peace, with the palm facing out.	A hand gesture representing peace or victory.
peace_inverted	The “V” sign for peace is made with the palm facing inward.	An inverted peace sign.
rock	The index finger and little finger are extended in a “rock on” gesture.	The hand forms a horns sign, often associated with rock music.
hand_heart	Both hands are brought together to form the shape of a heart.	A symbol of love or affection made with the hands.
like	The thumb is pointed up in a “thumbs-up” gesture of approval.	A “like” or “good job” sign made with the thumb.
dislike	The thumb is pointed down in a “thumbs-down” gesture of disapproval.	A “dislike” sign made by pointing the thumb downwards.
stop	The hand is held up with the palm facing forward to signal “stop”.	A universal sign to halt or cease an action.
stop_inverted	An inverted version of the stop gesture.	A stop sign made with the back of the hand facing forward.
point	The index finger is extended to point at a person or object.	A gesture used to indicate a direction or draw attention to something.
grabbing	The hand is held up with fingers spread and curled, as if grabbing a large object.	A claw-like gesture used to represent grabbing.
grip	Holding a small object securely between the thumb and index finger.	A precision grip used to manipulate a small item.
call	The thumb and little finger are extended in a “call me” gesture.	The hand shape mimics holding a telephone receiver.
timeout	The hands form a “T” shape, signaling a pause or timeout.	A common gesture in sports to request a break.
no_gesture	The hand is in a neutral, resting state with no specific gesture.	No specific gesture is being performed by the hand.
holy	The hands are held together in a prayer-like or reverent gesture.	A gesture symbolizing prayer or respect.
little_finger	Only the little finger is extended, often for a promise or pinky swear.	The pinky finger is held up.
middle_finger	An offensive gesture with the middle finger extended.	The middle finger is raised while the others are in a fist.
mute	A gesture indicating a request for silence.	The hand covers the mouth or fingers are held to the lips to mean “be quiet”.
take_picture	The hand mimics the action of pressing a camera shutter button.	A gesture that looks like someone is taking a photograph.
three_gun	A gesture resembling a gun, often made with the thumb and first two fingers.	A playful hand gesture shaped like a firearm.
thumb_index	A single hand is held up with the thumb and index finger extended to form an “L” shape.	A one-handed gesture shaped like the letter “L”.
thumb_index2	Both hands simultaneously form an “L” shape with the thumb and index finger.	A two-handed gesture where each hand makes the shape of the letter “L”.
xsign	The arms are crossed over the chest to form an “X”.	A defensive or blocking gesture made by crossing the arms.

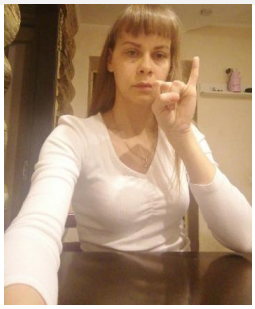
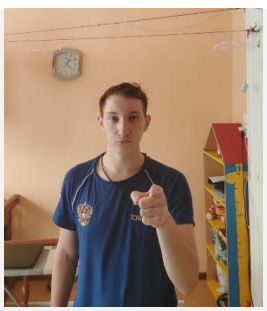
Image	*Question*	*Qwen base*	*Qwen finetuned*
	Considering the person's pose and hand shape, which description accurately identifies the gesture? **A.** Holding a small object securely between the thumb and index finger. **B.** The hands form a "T" shape, signaling a pause or timeout. **C.** All fingers are curled inward with the thumb wrapped around them. **D.** A single hand is held up with the thumb and index finger extended to form an 'L' shape.	D ✗	C ✓
	Choose the sentence that correctly describes the hand gesture in the image. **A.** An inverted version of the stop gesture. **B.** A gesture indicating a request for silence. **C.** The index finger and little finger are extended in a "rock on" gesture. **D.** The pinky finger is held up.	A ✗	D ✓
	Choose the sentence that correctly describes the hand gesture in the image. **A.** A common gesture in sports to request a break. **B.** A hand gesture indicating that everything is alright. **C.** The hand gesture indicates a quantity of four. **D.** The hand is held up with only the index finger pointing upwards.	D ✗	A ✓
	From the options below, identify the gesture being performed. **A.** The hand is closed tightly into a fist. **B.** The thumb is pointed up in a "thumbs-up" gesture of approval. **C.** The hand forms a horns sign, often associated with rock music. **D.** The index and middle fingers form a "V" to symbolize peace, with the palm facing out.	C ✗	D ✓
	What message or meaning is the person in the image communicating with their hands? **A.** A gesture used to indicate a direction or draw attention to something. **B.** The hand is open with the palm facing forward, showing five fingers. **C.** A defensive or blocking gesture made by crossing the arms. **D.** A symbol of love or affection made with the hands.	B ✗	A ✓

Figure 19. **Qualitative Results on Zero-shot Gesture Recognition on HaGRID dataset [5].**

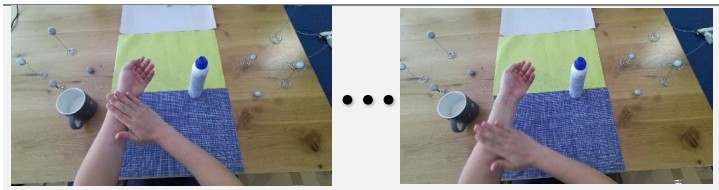
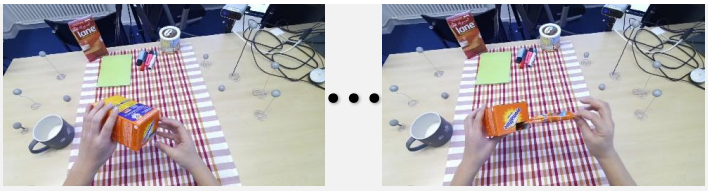
Sequence	*Question*	*Qwen base*	*Qwen finetuned*
	From the options below, identify the action being performed. **A.** apply spray **B.** put in espresso **C.** grab lotion **D.** close lotion	C ✗	A ✓
	From the options below, identify the action being performed. **A.** open milk **B.** take out espresso **C.** grab cappuccino **D.** place lotion	C ✗	B ✓
	From the options below, identify the action being performed. **A.** place spray **B.** grab cappuccino **C.** grab espresso **D.** grab spray	A ✗	D ✓
	From the options below, identify the action being performed. **A.** grab cappuccino **B.** place cocoa **C.** open milk **D.** take out cocoa	D ✗	B ✓
	From the options below, identify the action being performed. **A.** take out chips **B.** take out cocoa **C.** close lotion **D.** take out espresso	B ✗	A ✓
	From the options below, identify the action being performed. **A.** put in cappuccino **B.** close milk **C.** apply lotion **D.** take out cappuccino	A ✗	D ✓

Figure 20. **Qualitative Results on Zero-shot Hand-Object Interaction Recognition on H2O dataset [6].**

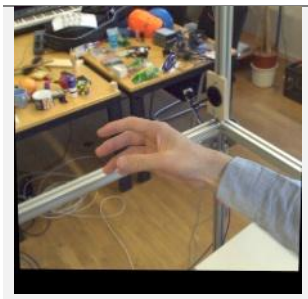
Image	*Question*	*LLaVA base*	*LLaVA finetuned*
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: **A.** The index finger is bent completely inward at the proximal interphalangeal joint. **B.** The index finger is 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 ✓

Figure 21. **Qualitative Comparison on FreiHAND [13].** Examples comparing LLaVA (base) and LLaVA fine-tuned on FreiHAND.

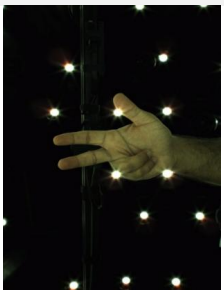
Image	*Question*	*LLaVA base*	*LLaVA finetuned*
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: **A.** The middle finger is bent completely inward at the distal interphalangeal joint. **B.** The middle finger is 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 ✓

Figure 22. **Qualitative Comparison on InterHand2.6M [9].** Examples comparing LLaVA (base) and LLaVA fine-tuned on InterHand2.6M.

Image	*Question*	*LLaVA base*	*LLaVA finetuned*
	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 23. **Qualitative Comparison on FPHA [2].** Examples comparing LLaVA (base) and LLaVA fine-tuned on FPHA.

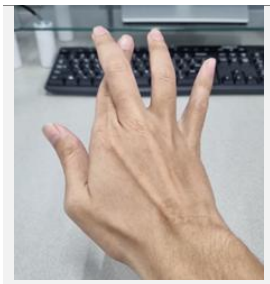
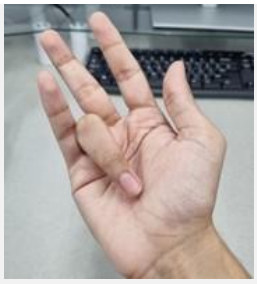
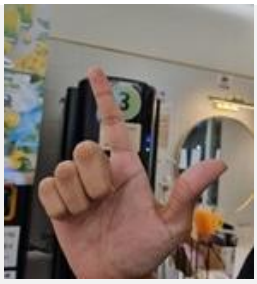
Image	Question	Options	LLaVA base	LLaVA finetuned	Why the question matters
	Are any fingers crossing each other?	A) Yes B) No	B ✗	A ✓	Detects self-occlusion patterns
	Which pair of fingers are spread widest at their tips?	A) Index–Middle B) Middle–Ring C) Ring–Little	A ✗	B ✓	Compares distance between 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

Figure 24. **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] 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. 6

[2] 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, 4, 15, 21

[3] Tobias Groot and Matias Valdenegro Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico, 2024. Association for Computational Linguistics. 11

[4] 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. 6

[5] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid – hand gesture recognition image dataset. In *WACV*, 2024. 14, 17

[6] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 14, 18

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 6

[8] 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. 6

[9] 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. 2, 4, 15, 20

[10] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 11, 14

[11] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2024. 14

[12] 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. 6

[13] 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, 4, 15, 19