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 Tiruneh[1],
Muneeb A. Khan[1], Muhammad Salman Ali[1], Binod Bhattarai[2,3,4†], Seungryul Baek[1†]

* Equal contribution † Joint supervision

UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
1495 UNIVERSITY OF ABERDEEN
UCL
fogsphere

Talk Roadmap

1. Motivation
2. Benchmark
3. Results
4. Qualitative Evidence
5. Conclusion

Motivation

VLMs can answer broad visual questions,
but still fail on **fine hand geometry**.

Key Problem

Hand understanding depends on subtle 21-joint articulation: bend, distance, and 3D ordering.

Why Existing Benchmarks Are Not Enough

Existing VQA benchmarks often mix
visual geometry with **language priors**.

- A model may guess the plausible caption without reading joint geometry.
- We need questions whose answers are computed from 3D hand pose.

HandVQA in One Sentence

HandVQA is a controlled VQA benchmark
for diagnosing **joint-level spatial reasoning**
in vision-language models.

Central Idea

Every answer is grounded in normalized 3D hand keypoint geometry.

Benchmark Scale

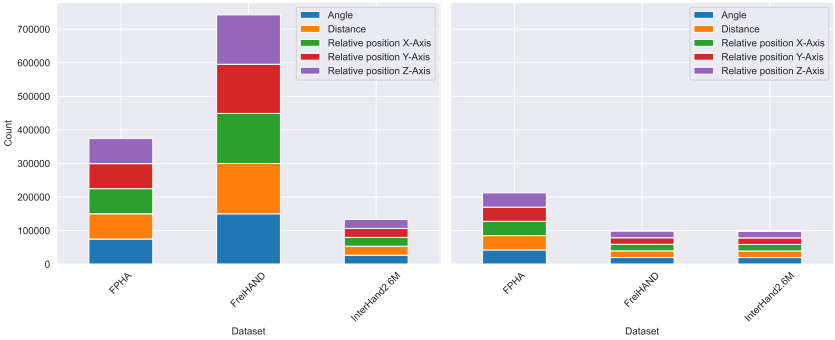
1.6M+

controlled multiple-choice questions

Data Sources

FreiHAND InterHand2.6M FPHA

Five Spatial Descriptors



HandVQA isolates angle, distance, and relative position along X/Y/Z.

Question Format

Hand Image

Joint Names

1. Thumb tip *(TIP) joint*
2. Middle finger distal interphalangeal (DIP) joint
3. Ring finger metacarpophalangeal (MCP) joint
4. Ring finger distal interphalangeal (DIP) joint
5. Ring finger tip (TIP) joint
6. Little finger metacarpophalangeal (MCP) joint
7. Little finger distal interphalangeal (DIP) joint
8. Little finger tip (TIP) joint

Angle

From the multiple choice answers *given in* the options below choose the sentence that correctly describes the relationship in the image:
Options:
A. The little finger is bent completely inward at the distal interphalangeal joint.
B. The little finger is bent inward at the distal interphalangeal joint.
C. The little finger is bent slightly inward at the distal interphalangeal joint.
D. The little finger is straight at the distal interphalangeal joint.

Distance

From the multiple choice answers *given in the* options below choose the sentence that correctly describes the relationship in the image:
Options:
A. The distal interphalangeal joint of the middle finger is spread from the distal interphalangeal joint of the ring finger.
B. The distal interphalangeal joint of the middle finger is close to the distal interphalangeal joint of the ring finger.
C. The distal interphalangeal joint of the middle finger is spread wide from the distal interphalangeal joint of the ring finger.

Relative Position X

From the multiple choice answers *given* in the options below choose the sentence that correctly describes the relationship in the image:
Options:
A. The tip joint of the thumb is at the right of the metacarpophalangeal joint of the little finger.
B. The tip joint of the thumb is at the left of the metacarpophalangeal joint of the little finger.

Relative Position Y

From the multiple choice answers *given* in the options below choose the sentence that correctly describes the relationship in the image:
Options:
A. The tip joint of the thumb is below the metacarpophalangeal joint of the ring finger.
B. The tip joint of the thumb is above the metacarpophalangeal joint of the ring finger.

Relative Position Z

From the multiple choice answers *given* in the options below choose the sentence that correctly describes the relationship in the image:
Options:
A. The tip joint of the ring finger is in front of the tip joint of the thumb.
B. The tip joint of the ring finger is behind the tip joint of the thumb.

One image supports many questions, each targeting one spatial relation.

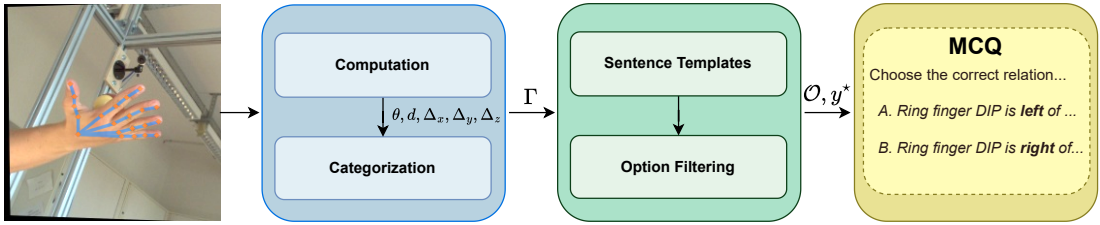
Ground Truth Comes from Geometry

Correct answers are computed from 3D coordinates,
not manual captions or text heuristics.

Example Descriptors

θ for angle d for distance $\Delta_x, \Delta_y, \Delta_z$ for direction

QA Generation Pipeline



Pose descriptors become anatomy-aware MCQs.

Angle Descriptor

Angle asks: **how bent is a finger joint?**

Categories			
Bent completely inward	Bent inward	Bent slightly inward	Straight

Distance Descriptor

Distance asks: **how far apart are two joints?**

Categories

Close to Spread from Spread wide from

Relative Position Descriptor

Relative position asks: **which joint is left/right, above/below, or in front/behind?**

Design Choice

Ambiguous “aligned” cases are removed to avoid visually unanswerable questions.

Main Result

Base VLMs are weak on fine hand geometry,
but HandVQA fine-tuning produces large spatial gains.

Most Visible Failure

Distance and directional relations are especially brittle before fine-tuning.

Result: Angle and Distance I

Table: Angle and Distance results for **DeepSeek Janus Pro 7B** and **LLaVA Mistral 7B**. **Gold**, **Silver**, and **Bronze** denote the top three per metric.

Model	Tuned	Eval	Angle		Distance	
			Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
Base model (no tuning)						
DeepSeek Janus Pro 7B	-	InterHand2.6M	34.10	0.883	45.55	0.657
DeepSeek Janus Pro 7B	-	FreiHAND	35.31	0.830	44.15	0.668
DeepSeek Janus Pro 7B	-	FPHA	26.46	0.991	39.02	0.819
Finetuned models						
DeepSeek Janus Pro 7B	InterHand2.6M	InterHand2.6M	68.00	0.334	88.02	0.122
DeepSeek Janus Pro 7B	FreiHAND	FreiHAND	61.30	0.402	85.23	0.151
DeepSeek Janus Pro 7B	FPHA	FPHA	66.08	0.438	81.60	0.184
Base model (no tuning)						
LLaVA Mistral 7B	-	InterHand2.6M	40.08	0.739	16.20	1.293
LLaVA Mistral 7B	-	FreiHAND	42.48	0.678	13.18	1.342
LLaVA Mistral 7B	-	FPHA	23.38	1.011	13.57	1.353
Finetuned models						
LLaVA Mistral 7B	InterHand2.6M	InterHand2.6M	**74.35**	**0.263**	**90.79**	**0.094**
LLaVA Mistral 7B	FreiHAND	FreiHAND	**62.91**	**0.382**	**86.19**	**0.141**
LLaVA Mistral 7B	FPHA	FPHA	**68.37**	**0.401**	**83.99**	**0.161**

Result: Angle and Distance II

Table: Angle and Distance results for **Qwen-2.5 VL 7B Instruct**. **Gold**, **Silver**, and **Bronze** denote the top three per metric.

Model	Tuned	Eval	Angle		Distance	
			Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
Base model (no tuning)						
Qwen-2.5 VL 7B	-	InterHand2.6M	37.92	0.779	19.58	1.247
Qwen-2.5 VL 7B	-	FreiHAND	38.70	0.746	20.48	1.208
Qwen-2.5 VL 7B	-	FPHA	24.22	1.055	18.03	1.306
Finetuned models						
Qwen-2.5 VL 7B	InterHand2.6M	InterHand2.6M	67.08	0.341	88.56	0.116
Qwen-2.5 VL 7B	FreiHAND	FreiHAND	54.55	0.483	82.16	0.182
Qwen-2.5 VL 7B	FPHA	FPHA	62.94	0.481	80.88	0.192

Result: Directional Relations I

Table: Relative-position accuracy for **DeepSeek Janus Pro 7B** and **LLaVA Mistral 7B**. **Gold**, **Silver**, and **Bronze** denote the top three per dataset.

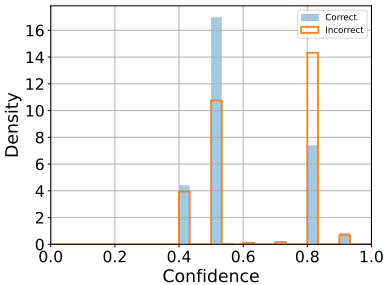
Model	Tuned	Eval	Rel. Pos. X Accuracy ↑	Rel. Pos. Y Accuracy ↑	Rel. Pos. Z Accuracy ↑
Base model (no tuning)					
DeepSeek Janus Pro 7B	-	InterHand2.6M	50.41	52.46	51.16
DeepSeek Janus Pro 7B	-	FreiHAND	49.80	51.55	50.03
DeepSeek Janus Pro 7B	-	FPHA	43.02	52.64	61.73
Finetuned Models					
DeepSeek Janus 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**

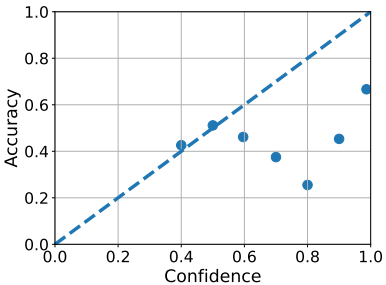
Result: Directional Relations II

Table: Relative-position accuracy for **Qwen-2.5 VL 7B Instruct**. **Gold**, **Silver**, and **Bronze** denote the top three per dataset.

Model	Tuned	Eval	Rel. Pos. X Accuracy ↑	Rel. Pos. Y Accuracy ↑	Rel. Pos. Z Accuracy ↑
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

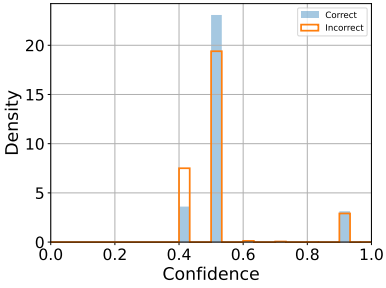
Confidence: Base Models

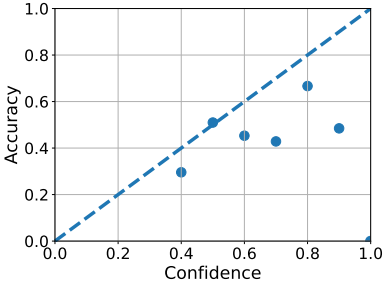




Base models are often **confidently wrong**.

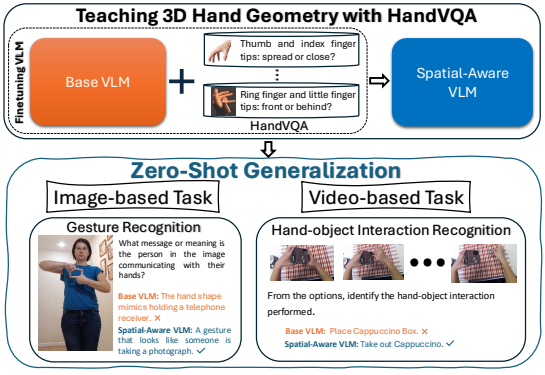
Confidence: After Fine-Tuning





Fine-tuning reduces high-confidence errors,
but tends to make predictions more conservative.

Zero-Shot Transfer



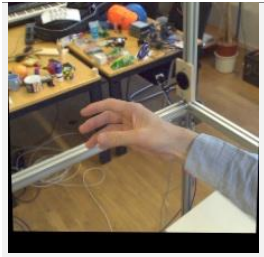
Learning 3D hand geometry transfers beyond HandVQA.

Transfer Gains

Model	Setting	Gesture	Interaction
LLaVA Mistral 7B	Base	57.42	-
LLaVA Mistral 7B	Tuned	**69.58**	-
Qwen 2.5 VL 7B Instruct	Base	71.86	80.26
Qwen 2.5 VL 7B Instruct	Tuned	**82.19**	**82.89**

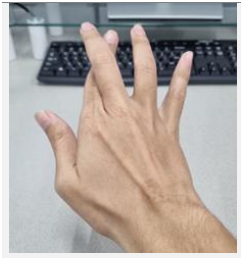
Gesture improves by up to **+10.33%**; interaction improves by **+2.63%** for Qwen 2.5 VL 7B.

Qualitative: FreiHAND

Image	Question	LLaVA (base)	LLaVA (fine-tuned on FreiHAND)
	From the multiple choice answers given in the options below choose the sentence that correctly describes the relationship in the image: A. The index finger is bent completely inward at the proximal interphalangeal joint. B. The index finger is bent inward at the proximal interphalangeal joint. C. The index finger is bent slightly inward at the proximal interphalangeal joint. D. The index finger is straight at the proximal interphalangeal joint.	B ✗	C ✓

Fine-tuning improves controlled single-hand reasoning.

Qualitative: Out-of-Distribution

Image	Question	Options	LLaVA (base)	LLaVA (fine-tuned on FreiHAND)	Why the question matters
	Are any fingers crossing each other?	A) Yes B) No	B ✗	A ✓	Detects self-occlusion patterns

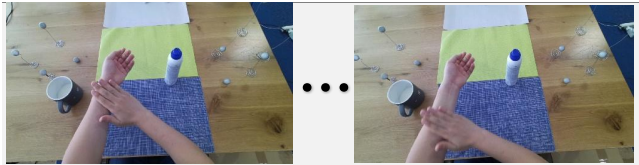
The tuned model is more reliable on natural OOD hand images.

Qualitative: Gesture Recognition

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 ✓

Better spatial reasoning supports gesture recognition.

Qualitative: Hand-Object Interaction

Sequence	*Question*	*base*	*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 ✓

Geometry-aware hand reasoning helps interaction recognition.

Conclusion

HandVQA reveals a concrete gap:
general VQA ability $\neq$ fine-grained hand spatial reasoning.

Takeaway

3D grounded training substantially improves angle, distance, direction, calibration, and downstream transfer.

Future Directions

HandVQA opens paths toward richer geometric and physical reasoning.

- **Geometry:** adaptive or learned mappings beyond fixed discretized thresholds.
- **Language:** diverse phrasing, comparative expressions, and explanations.
- **Video:** motion cues and contact dynamics for interaction tasks.
- **Embodiment:** connect HandVQA with VLA models for grasping, dexterous control, and manipulation.

Project Page

HandVQA resources

Questions?

Paper, code, dataset, and updates:
`https://kcsayem.github.io/handvqa/`