Can Vision Language Models Understand Mimed Actions?

Hyundong (Justin) Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon,, Deuksin Kwon, Natali T. Chavez, Jonathan May

ACL 2025 Findings

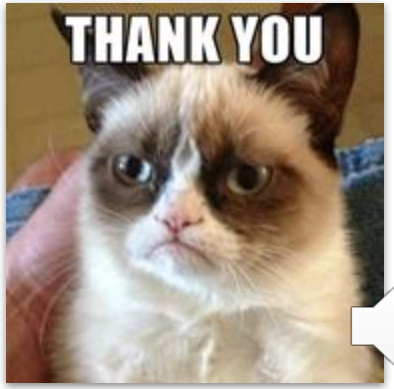
Nonverbal Communication

Use of nonverbal cues such as gestures, facial expressions, and body language to convey messages.

Nonverbal Communication (NVC)

Important because

- ✓ It’s an alternative means of communication when verbal modes are limited
- ✓ It makes interactions more engaging and natural
- ✓ It can convey true intentions that contradicts what is verbally expressed

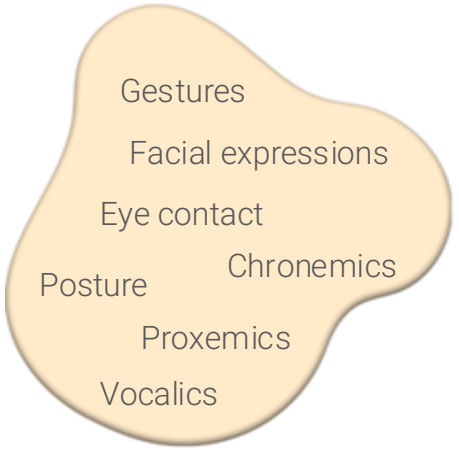


Nonverbal Communication (NVC)

So understanding NVC is important for AI systems to become more effective and accessible assistants.

But!

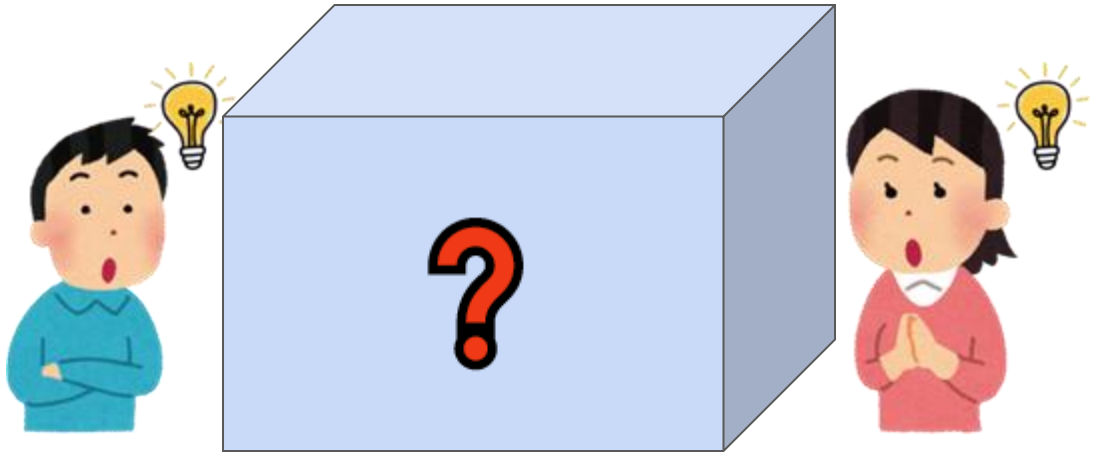
- ✗ Scope of NVC is large
- ✗ High variability in how it is interpreted among individuals and cultures



Nonverbal Communication (NVC)

How can we make a meaningful step towards NVC understanding?

- Clearly defined scope
- Tractable with low variability



Mime!

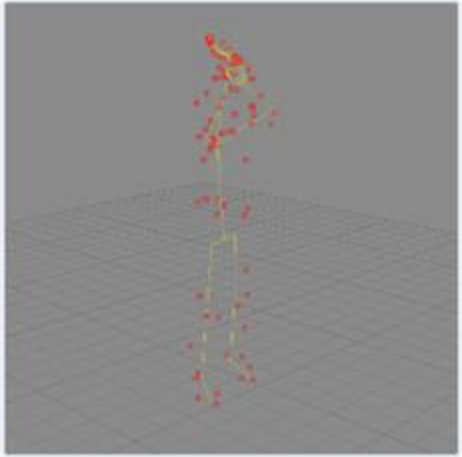
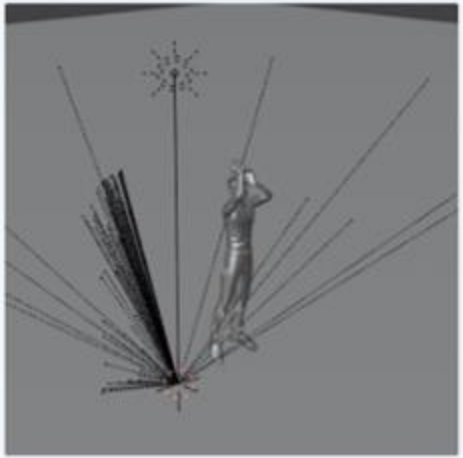
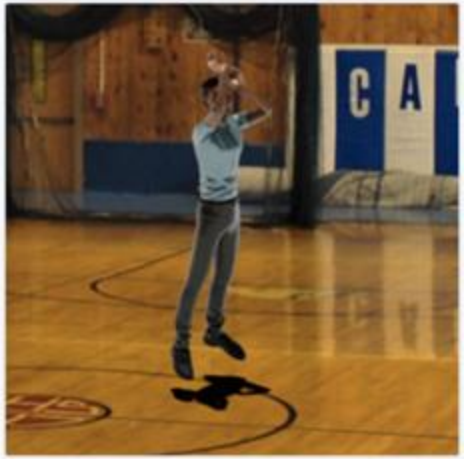
Theatrical technique of suggesting intent with gesture, expression, and movement

- ✓ Requires robust understanding of human gestures → a crucial prerequisite of understanding NVC
- ✓ Low interpretation variability → tractable

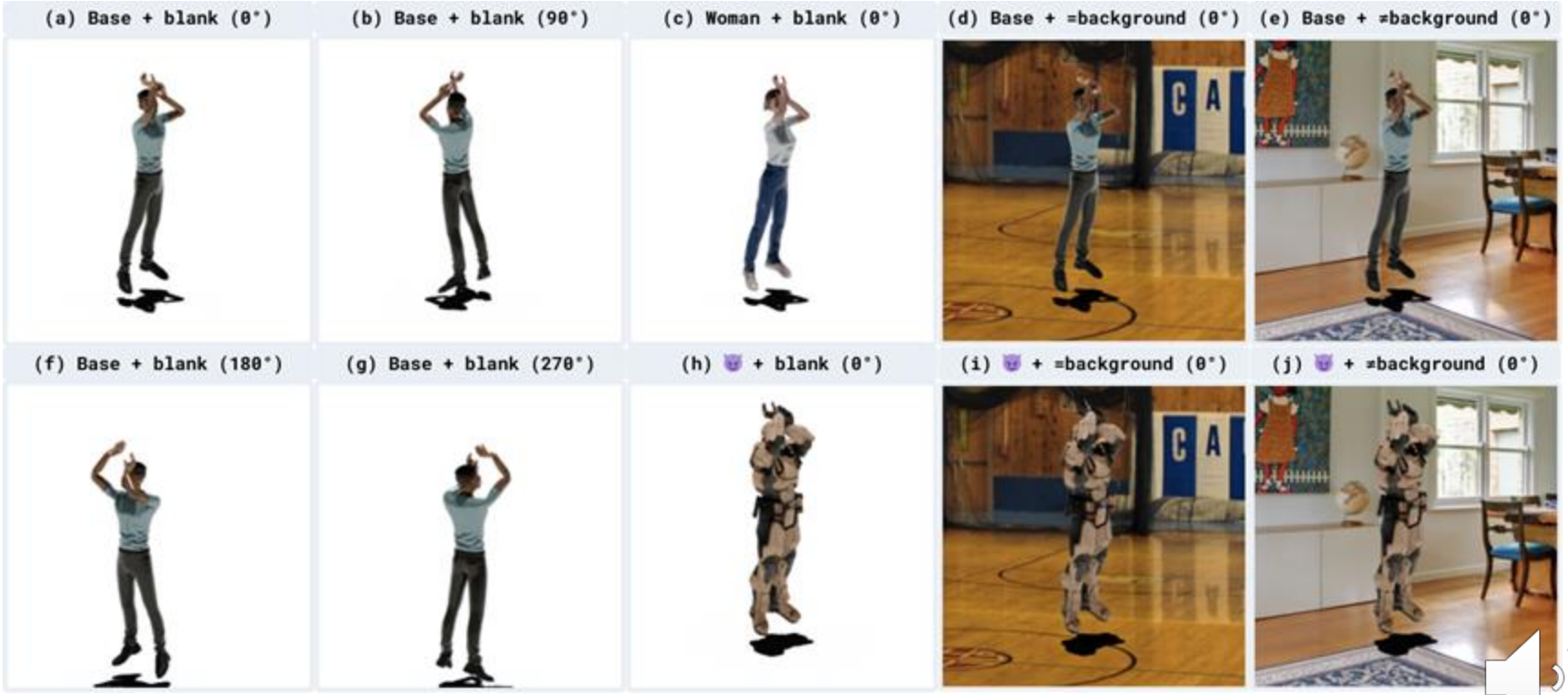
Can AI systems (Vision Language Models) reliably recognize mimed actions?

MIME: Mime Identification Multimodal Evaluation

MIME Construction Pipeline

(1) Collect motion capture data with Vicon	(2) Retarget to 3D character with Blender	(3) Render animations with GPU acceleration	(4) Overlay rendered frames over background image
			

MIME Variants



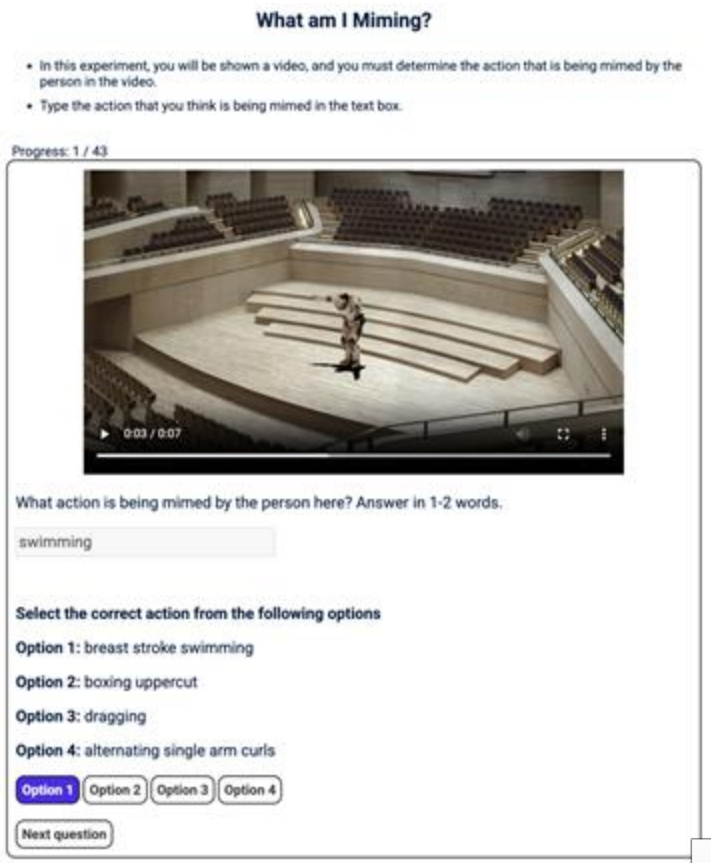
MIME Facts

86 videos per variation

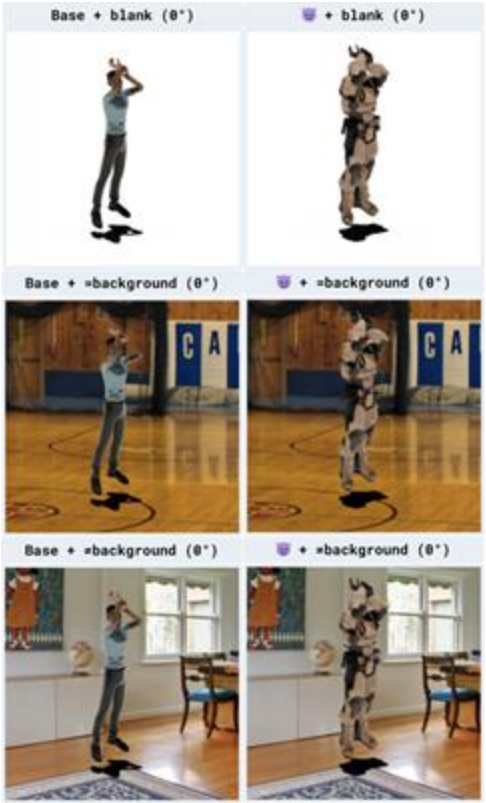
Number of variations	**10**
Total number of Videos	**860**
	#
Action Types	**47**
Videos	**86**
Sport activities	43
Musical activities	11
Daily activities	27
Miscellaneous	5
Motion Capture Actors	2
Nonprofessional Asian male	1
Professional European female	1
Test Format	2
Free-form short answer	1
Multiple choice	1

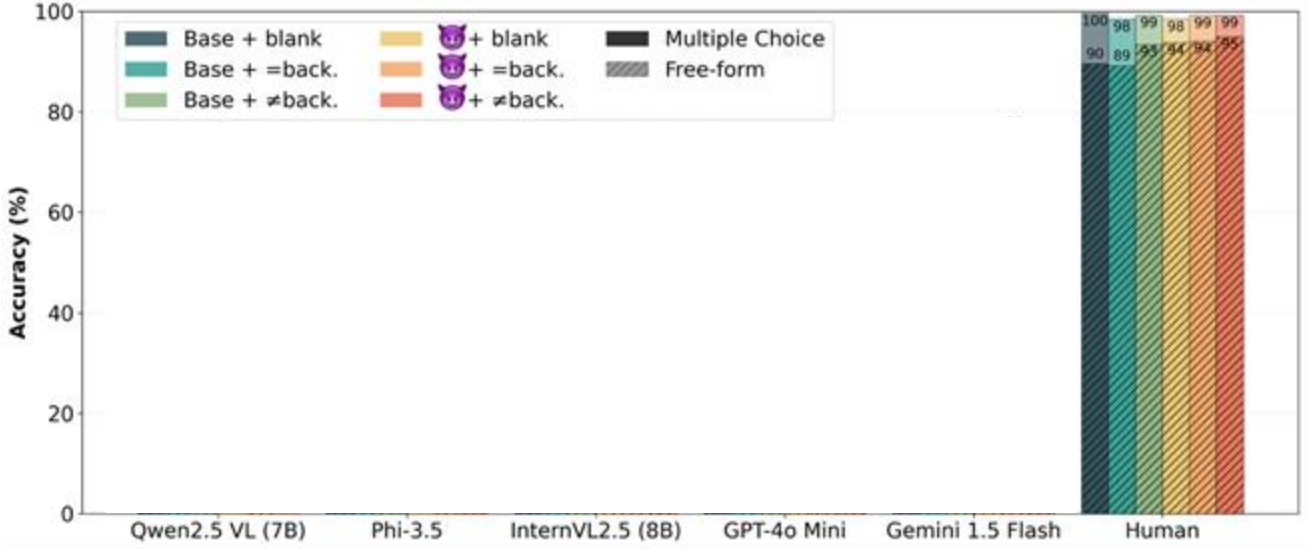
Human Performance

- 60 participants
- Each complete half of one variation (43 samples)



MIME Results





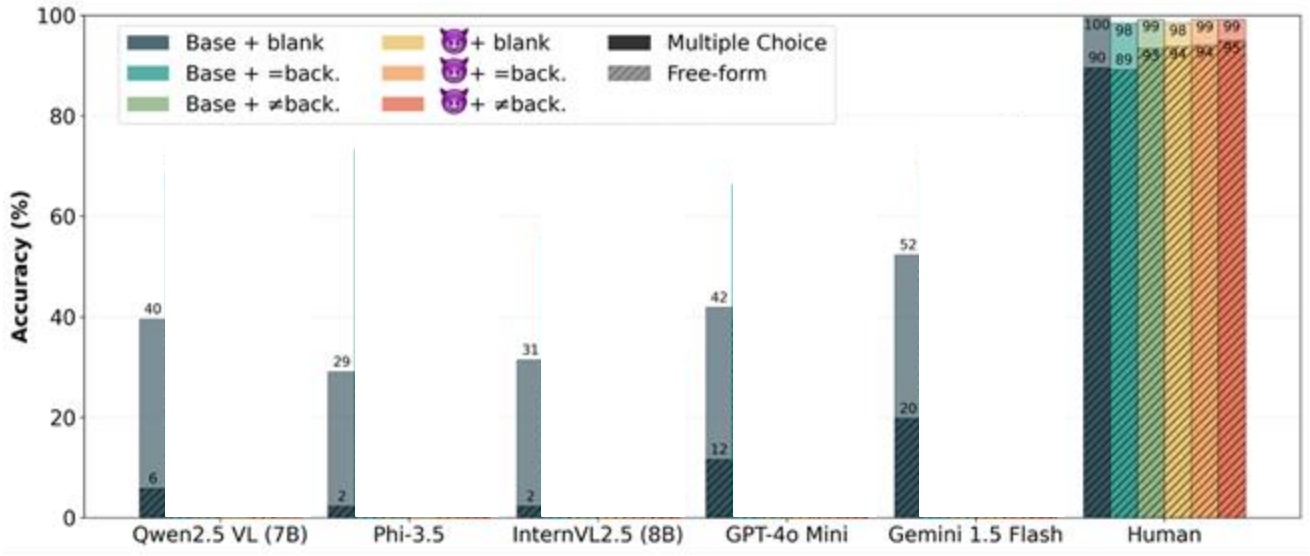
→ **Takeaway:** Humans achieve almost perfect accuracy regardless of variations and evaluation format.

Models

- Open-weight models
  - Qwen 2.5 VL (3B & 7B)
  - Phi 3.5
  - InternVL 2.5 (8B)

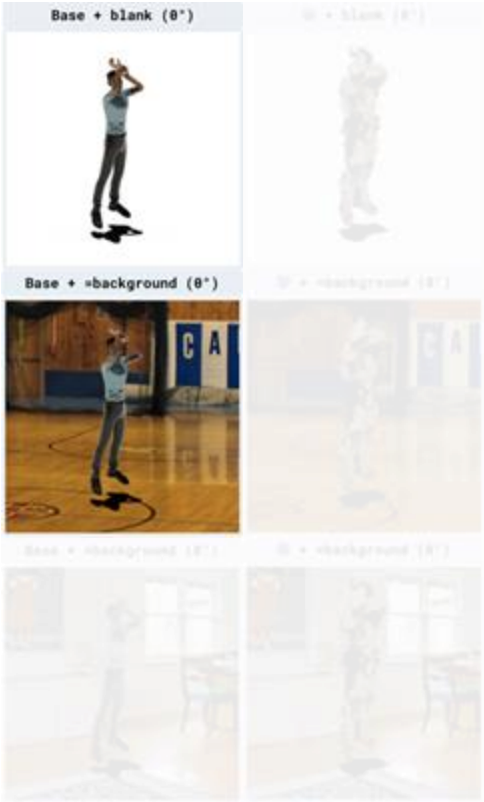
- Black box API models
  - GPT-4o Mini
  - Gemini 1.5 Flash

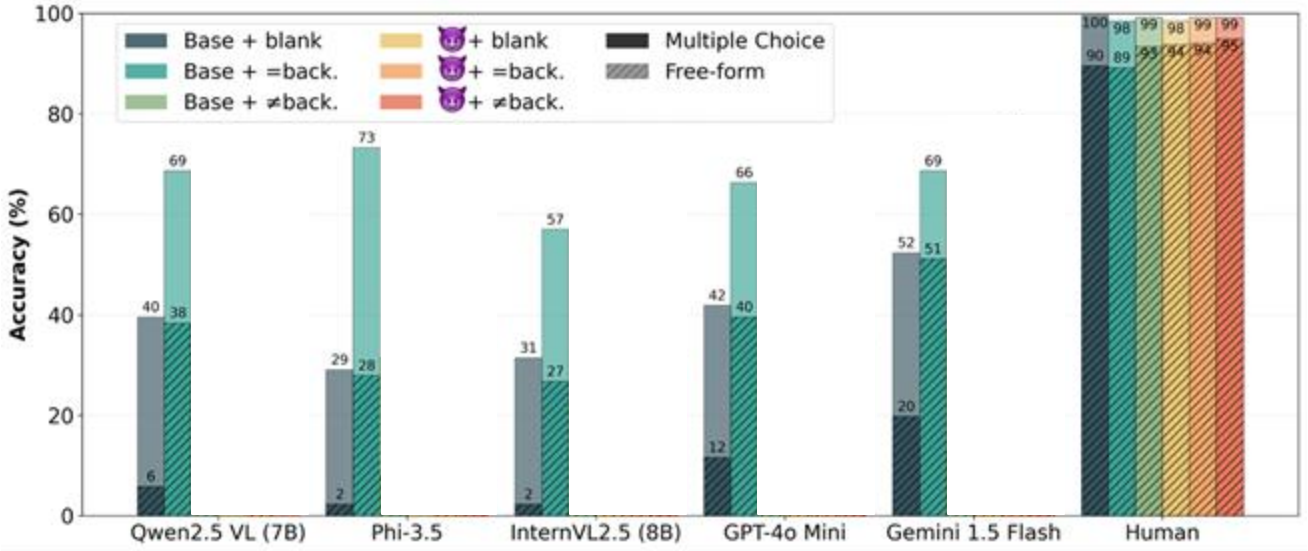
MIME Results



→ **Takeaway:** VLMs struggle even with the base setting, and a lot more with the free-form format.

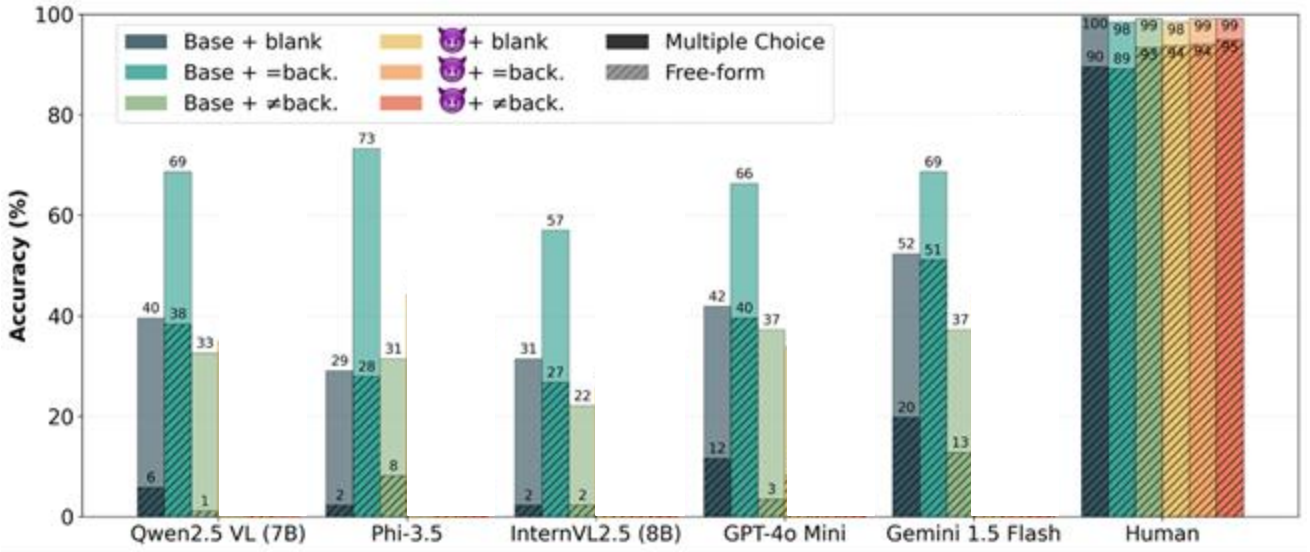
MIME Results





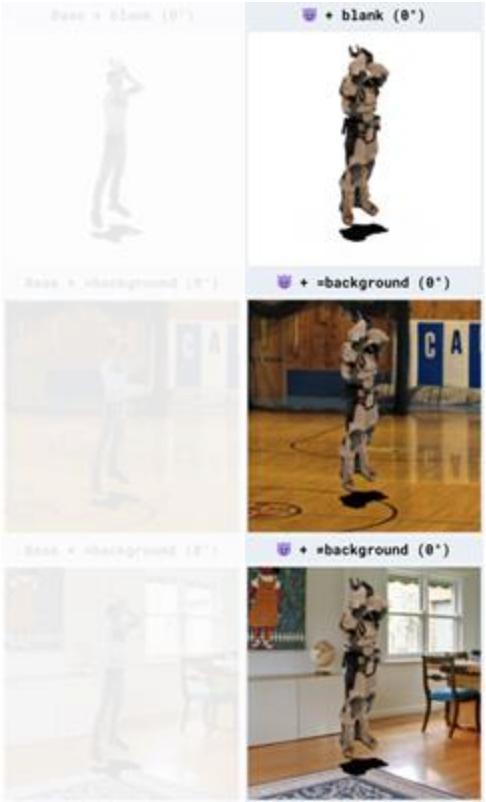
→ **Takeaway:** VLMs get a significant boost when salient context is provided from the background.

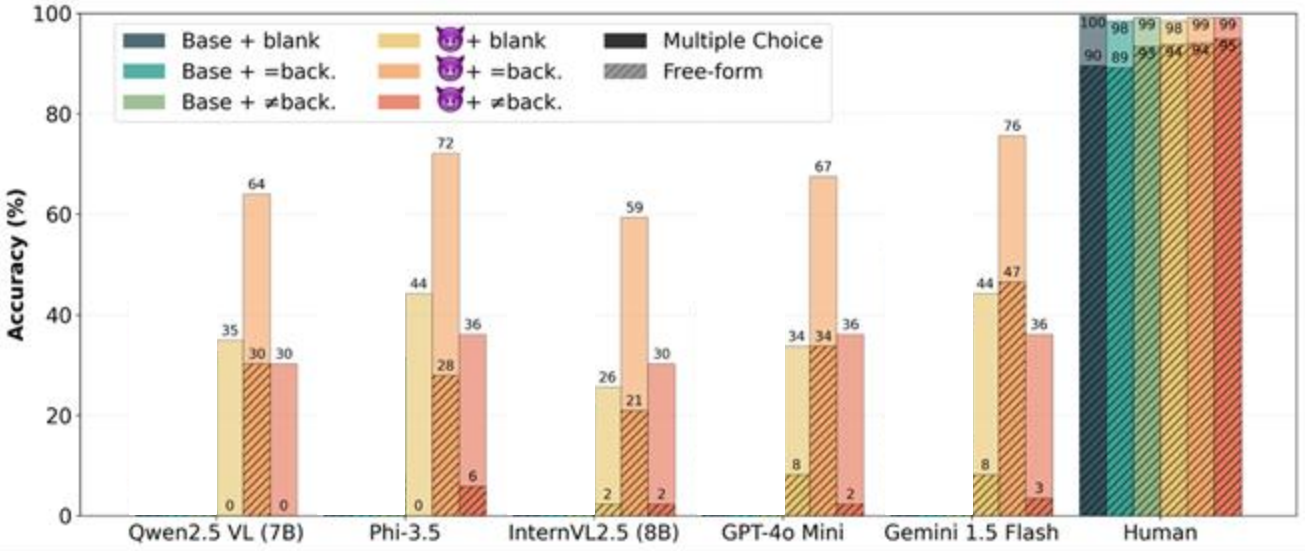
MIME Results



→ **Takeaway:** VLMs get confused when an adversarial background is provided.

MIME Results





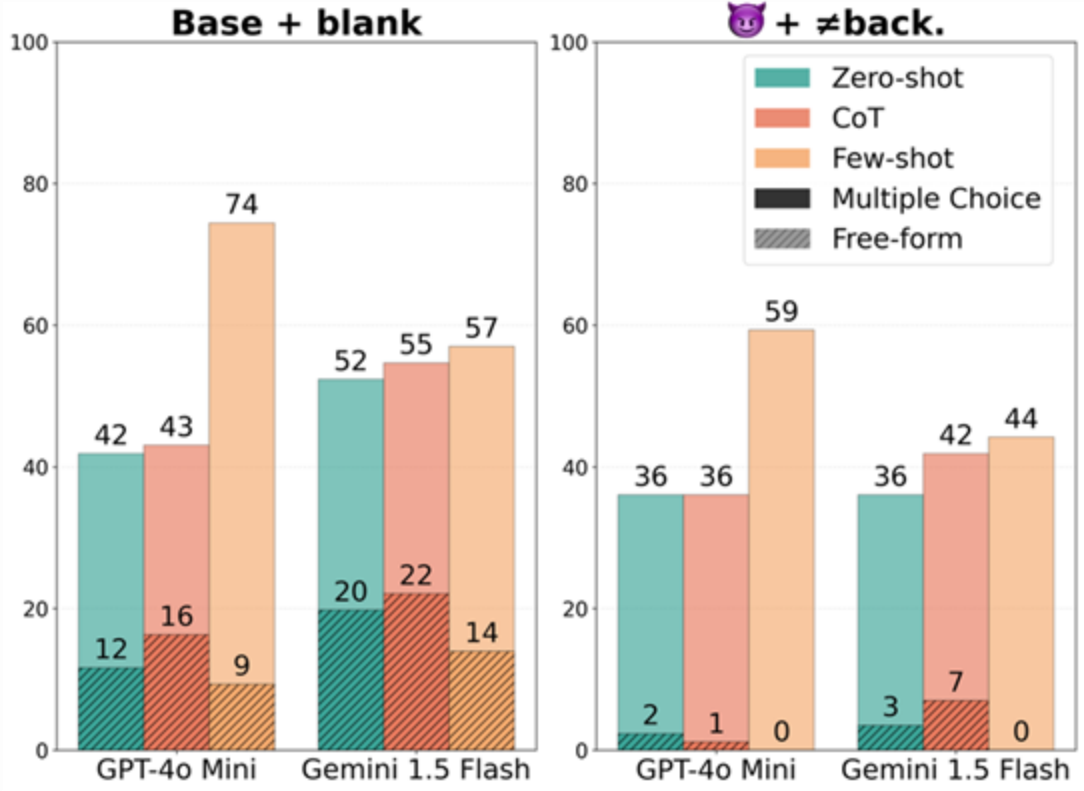
→ **Takeaway:** Adversarial character also leads to a minor drop in accuracy for VLMs.

Can we improve a VLM's performance on MIME?

- Chain-of-Thought (CoT)
  - i.e. Focus only on the action, describe the action that you see, and then predict

- Few-shot In-Context Learning (Few-shot)
  - Three examples

Can we improve a VLM's performance on MIME?



- CoT and few-shot leads to improvements in multiple-choice format
- But not for free-form!

Why Do VLMs fail?

Examine Gemini 1.5 Flash's CoTs as a proxy of what VLMs are observing.

Base + blank (0°)

Mode	%	Mimed Action	Prediction	Generated CoT Description
○	16.3	Drinking out of a cup	Answering a phone call	They bring their hand to their ear as if holding a phone, and move their mouth as if talking. They also have a slightly downward gaze, as if focusing on the phone.

○: Completely incorrect description

Why Do VLMs fail?

Examine Gemini 1.5 Flash's CoTs as a proxy of what VLMs are observing.

Base + blank (0°)

Mode	%	Mimed Action	Prediction	Generated CoT Description
○	16.3	Drinking out of a cup	Answering a phone call	They bring their hand to their ear as if holding a phone, and move their mouth as if talking. They also have a slightly downward gaze, as if focusing on the phone.
◑	53.5	Driving	Riding a tiny board	They are miming the action of holding a small object, such as a remote control, and are making small, precise movements with their hands as if they are controlling or manipulating something. Their body posture and stance remain relatively static. The actions are subtle, suggesting a delicate or precise control of the imagined object.

○ : Completely incorrect description
◑ : Mix of correct and incorrect description

Why Do VLMs fail?

Examine Gemini 1.5 Flash's CoTs as a proxy of what VLMs are observing.

Base + blank (0°)

Mode	%	Mimed Action	Prediction	Generated CoT Description
○	16.3	Drinking out of a cup	Answering a phone call	They bring their hand to their ear as if holding a phone, and move their mouth as if talking. They also have a slightly downward gaze, as if focusing on the phone.
◑	53.5	Driving	Riding a tiny board	They are miming the action of holding a small object, such as a remote control, and are making small, precise movements with their hands as if they are controlling or manipulating something. Their body posture and stance remain relatively static. The actions are subtle, suggesting a delicate or precise control of the imagined object.
●	12.8	Pushing	Falling down	They bend their knees, lean forward, and spread their arms out as if they are losing their balance and falling onto the ground. Their posture and movements strongly suggest a loss of equilibrium resulting in a fall. The shadow on the ground further emphasizes this interpretation.

○ : Completely incorrect description
◑ : Mix of correct and incorrect description
● : Fully correct description but incorrect interpretation

Why Do VLMs fail?

Relevant context biases descriptions to become correct.

Base + =background (0°)

Mode	%	Mimed Action	Prediction	Generated CoT Description
💡	37.2	Throwing dart	Throwing darts	They go through the motions of holding an imaginary dart, aiming, and then releasing it with a follow-through.

💡 : Predictions helped by context

Why Do VLMs fail?

Adversarial context leads to incorrect action descriptions.

Mode	%	Mimed Action	Prediction	Generated CoT Description
😀	24.4	Climbing	Conducting orchestra	They are standing in a relatively formal posture, and their arm movements are sweeping and expressive, suggesting they are leading a musical piece. The upward and downward motions of their arms mimic the conducting of different musical sections or instruments. The movements are deliberate and rhythmic, further supporting the interpretation of conducting.

😵🐟 : Predictions confused by context

Summary

- ✓ **MIME**: a video-based question answering benchmark for mimed action understanding, created with motion capture data and 3D graphics software
- ✓ Humans achieve almost perfect accuracy regardless of variation and evaluation format, while VLMs struggle even in the base setting.
- ✓ VLMs demonstrate only a superficial understanding by achieving higher performance when given relevant context and lower performance with adversarial context.
- ✓ Chain-of-Thought descriptions show that most failures are attributed to VLMs not reliably generate correct descriptions of human actions.

Thank you!

Project page: justin-cho.com/mime

Contact: hd.justincho@gmail.com