Do you know what action this person is miming?

If so, you're smarter than vision language models!

Can Vision Language Models Understand Mimed Actions?

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

MIME

Mime Identification Multimodal Evaluation

Nonverbal Communication (NVC)

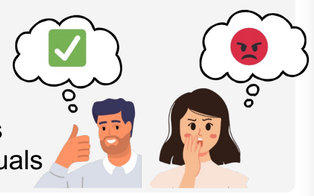
NVC: Use of nonverbal cues such as gestures, facial expressions, and body language to convey messages.

- ✓ Alternative means of communication when verbal modes are limited
- ✓ Makes communication more engaging & natural
- ✓ Convey true intent when NVC ≠ verbal expression

∴ 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 across individuals



Mimed Action Understanding

How can we make a meaningful step towards NVC understanding? Need a task with:

- Well-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

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

Not a photorealistic video dataset containing live action footage. Consists of 3D animation videos rendered with Blender using motion capture data of mimed actions and digital characters. See limitations in paper.

What action is this person miming?

Free-form

VLM 🤖**:** Catching invisible ball ❌

Human: Shooting a basketball ✅

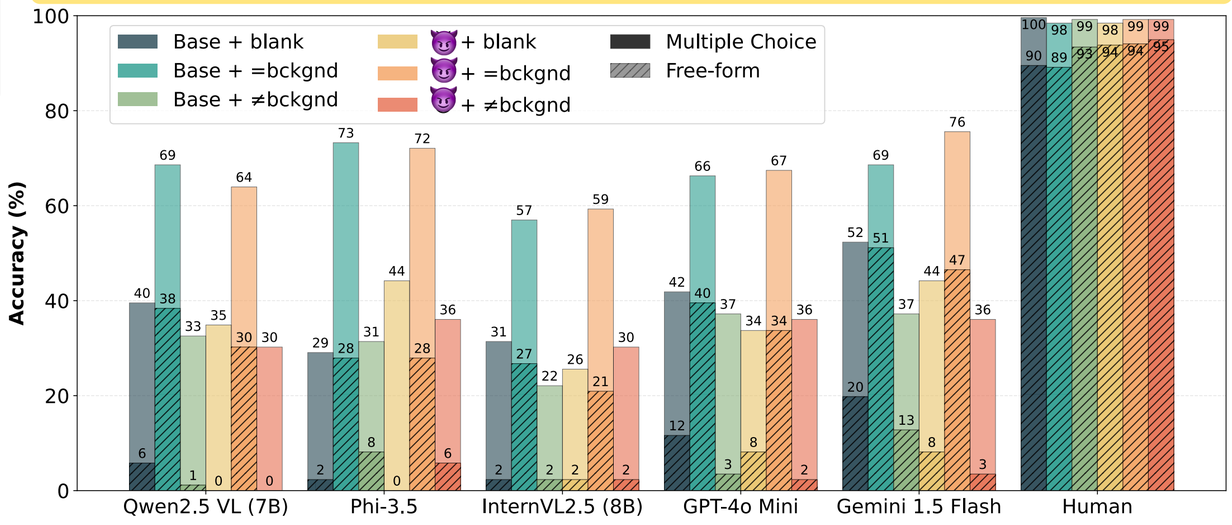
Multiple Choice

A: Playing harp
B: Using phone
C: Basketball shot
D: Playing guitar

VLM 🤖**:** B ❌

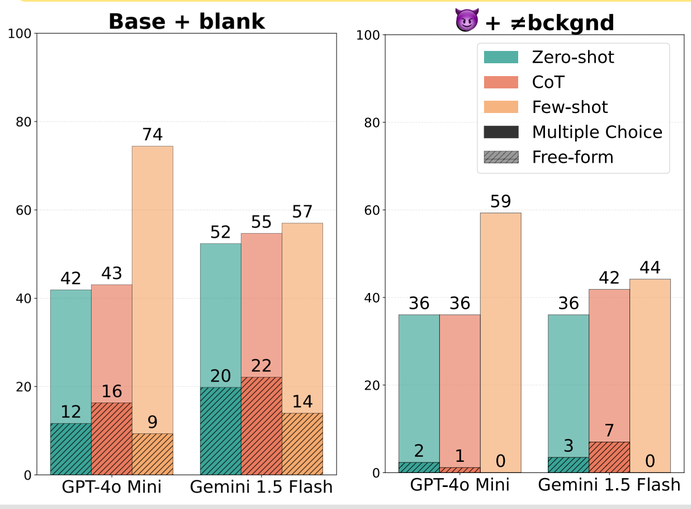
Human: C ✅

MIME Results



Takeaway: Humans are robust to all variations and evaluation formats, but while VLMs struggle. VLMs do significantly worse in free-form format compared to multiple choice, drop performance for adversarial variations (😈 and ≠bckgnd), and get a boost with an aligned background (=bckgnd).

Improving on MIME



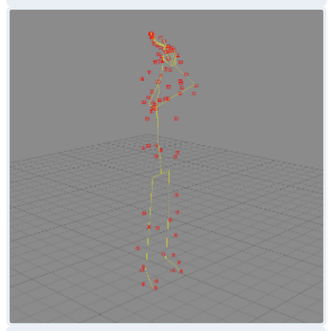
Takeaway: Chain-of-Thought (CoT) and few-shot lead to some improvements in multiple choice format but not free-form.

Most common failures revealed by CoT are partially or completely incorrect description of actions.

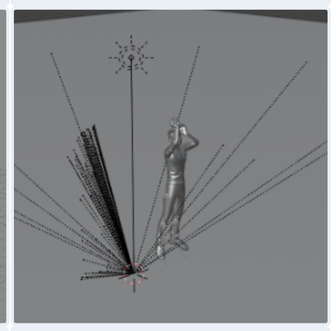
Prediction: Catching a ball
Reasoning:
They use their hands to make the motion of scooping up an unseen ball, bringing their hands up to their chest as if to secure the catch. (❌) Their body bends slightly at the knees (✅) and they shift their weight...

MIME Construction Pipeline

(1) Collect motion capture data with Vicon

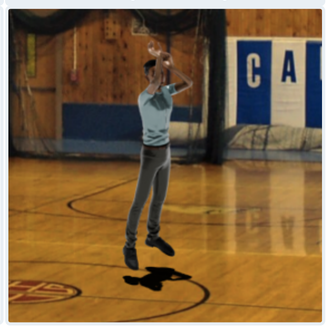


(2) Retarget to 3D character with Blender



(3) Render animations with GPU acceleration

(4) Overlay rendered frames over background image



MIME Variants

Base + blank (0°)

Base + blank (90°)

Base + blank (180°)

Base + blank (270°)

Woman + blank (0°)

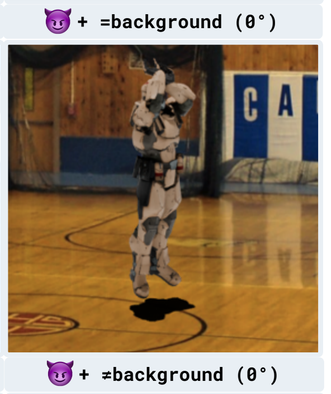
😈 + blank (0°)

Base + =background (0°)

😈 + =background (0°)

Base + ≠background (0°)

😈 + ≠background (0°)



USC Viterbi School of Engineering
USC Institute for Creative Technologies
USC Information Sciences Institute
CUTE LAB NAME
UCSB University of California Santa Barbara
Aristotle University of Thessaloniki

Contact: hd.justincho@gmail.com
Project page: justin-cho.com/mime