

# Can Vision Language Models Understand Mimed Actions?

Hyundong Cho<sup>1</sup> Spencer Lin<sup>2</sup> Tejas Srinivasan<sup>3</sup> Michael Saxon<sup>4</sup>  
Deuksin Kwon<sup>2</sup> Natali T. Chavez<sup>5</sup> Jonathan May<sup>1</sup>

Information Sciences Institute<sup>1</sup>, Institute for Creative Technologies<sup>2</sup>, Department of Computer Science<sup>3</sup>  
University of Southern California  
University of California, Santa Barbara<sup>4</sup> Aristotle University of Thessaloniki<sup>5</sup>  
hd.justincho@gmail.com

## Abstract

Nonverbal communication (NVC) plays an integral role in human language, but studying NVC in general is challenging because of its broad scope and high variance in interpretation among individuals and cultures. However, mime—the theatrical technique of suggesting intent using only gesture, expression, and movement—is a subset of NVC that consists of explicit and embodied actions with much lower human interpretation variance. We argue that a solid understanding of mimed actions is a crucial prerequisite for vision-language models capable of interpreting and commanding more subtle aspects of NVC. Hence, we propose Mime Identification Multimodal Evaluation (MIME), a novel video-based question answering benchmark comprising of 86 mimed actions. Constructed with motion capture data, MIME consists of variations of each action with perturbations applied to the character, background, and viewpoint for evaluating recognition robustness. We find that both open-weight and API-based vision-language models perform significantly worse than humans on MIME, motivating the need for increased research for instilling more robust understanding of human gestures.

## 1 Introduction

Nonverbal communication (NVC) — the use of nonverbal cues such as gestures, facial expressions, and body language to convey messages — is an instrumental part of human language (Mehrabian, 1972; Poyatos, 1983; Stickley, 2011). NVC not only serves as a crucial substitute to communication when verbal modes are limited (Friedman, 1979; Mast, 2007; Park et al., 2022; Shafique et al., 2023; Karmakar and Sinha, 2024), but also makes interaction engaging and natural (Kendon, 1967; Duncan Jr, 1969; Ha et al., 2012; Xu et al., 2022), and may even betray true intent that contradicts what is verbally expressed (Mehrabian, 1972; Mc-



Figure 1: Simplified illustration of a sample in MIME shown with a few frames from a video of a 3D male character miming a basketball shot in a living room. Humans achieve almost perfect accuracy on identifying mimed actions regardless of evaluation format, adversarial perturbations, and the absence of salient context (e.g., basketball, court, basketball outfit), while VLMs struggle without salient context.

Neill, 1992; Eaves and Leathers, 2015). Therefore, AI systems need to establish a thorough understanding of NVC for them to become more accessible and effective assistants to humans (Argyle and Trower, 1979; Troshani et al., 2021).

Unfortunately, this is an overwhelming undertaking considering the broad scope of NVC (Mehrabian, 1972; Eaves and Leathers, 2015), variability in how individuals interpret and exhibit nonverbal cues (Kita, 2009; Matsumoto and Hwang, 2013), and the limited capabilities of current vision-language models (VLMs) (Radford et al., 2021; Xu et al., 2021; Chen et al., 2024; Abdin et al., 2024; Bai et al., 2023; Gemini, 2024; Tang et al., 2025). Despite impressive achievements of VLMs on action recognition benchmarks (Kong and Fu, 2022; Wang et al., 2023; Qu et al., 2024), we find that they cannot even reliably identify a subset of NVC

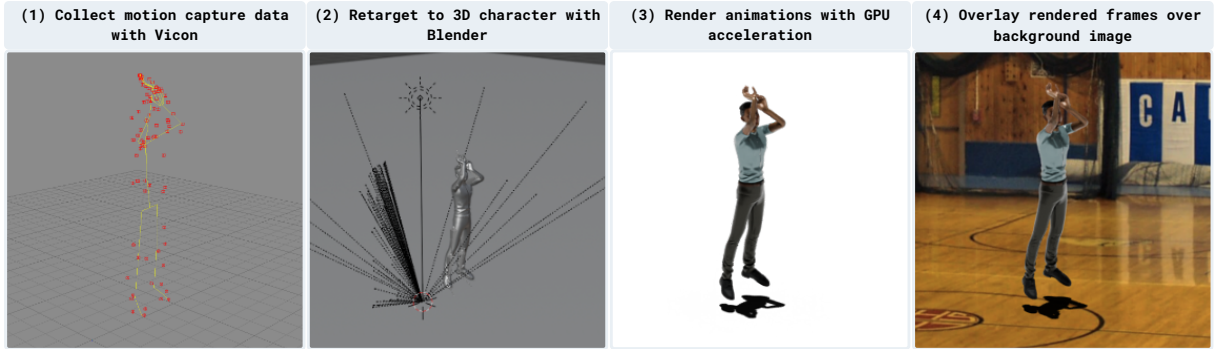


Figure 2: An overview of the pipeline for constructing MIME. (1) We first collect motion capture data of a mimed action on a Vicon stage. (2) Then, a 3D character is retargeted to our motion capture data in Blender, a computer graphics software. (3) Next, we render frames of the animation with a transparent background. (4) With frames rendered with transparent backgrounds, we can easily overlay them over images of our choice.

that human adults without apraxia<sup>1</sup> comprehend with ease (O’Reilly, 1995): mime, the theatrical technique of suggesting intent using only gesture, expression, and movement. Compared to other general gestures, many mimed actions are consistently identified among humans, in part due to their direct ties to physical movement and surfaces (O’Reilly, 1995; Alexanderson et al., 2017; van Nispen et al., 2017; Little and Firestone, 2021). Therefore, we propose studying whether VLMs can reliably recognize mimed actions as a foundational prerequisite towards the sophisticated comprehension of the full spectrum of NVC.

To this end, we address the following research questions: (i) *Can VLMs reliably recognize mimed actions?* and (ii) *If not, can we improve a VLM’s performance on identifying mimed actions?* For the first research question, we construct **Mime Identification Multimodal Evaluation (MIME)**,<sup>2</sup> a novel video-based question answering benchmark comprising of 86 mimed actions. We create MIME using motion capture data and computer graphics software, which enables us to create variations of each action with perturbations applied to the character, background, and viewpoint for evaluating recognition robustness (see Figure 1 for a sample of MIME and corresponding human and VLM predictions). On MIME, humans achieve almost 100% accuracy, regardless of adversarial perturbations and evaluation format. However, VLMs, open-weight models and API-based black-box models alike, only achieve at most 52.3% accuracy in a multiple choice format where contextual informa-

tion is provided by the answer choices and at most 19.8% with a free-from short answers format. Accuracy is even lower for videos with adversarial perturbations, for which all evaluated models achieve less than 10%. On the other hand, their performance is significantly boosted when provided a background that is contextually relevant (e.g., basketball court for mime of basketball shot).

To answer the second research question, we conduct a preliminary exploration into whether existing methods can bridge this shortcoming. Specifically, we experiment with Chain of Thought (Wei et al., 2022), few-shot in-context learning, and fine-tuning with a subset of MIME. We find that the only method that consistently improves model performance over zero-shot is few-shot in-context learning for API-based black-box models, but their results remain significantly worse than human performance. A manual inspection into the descriptions of the mimed actions generated by using Chain of Thought with Gemini 1.5 Flash reveal that the majority of failure cases is due to incorrect observations of the demonstrated gestures (80%) and a smaller portion is from incorrectly interpreting correctly generated descriptions (15%). In conclusion, our findings with MIME motivate research that instills a more robust understanding of human gestures in VLMs for establishing an essential foundation for NVC comprehension.

## 2 MIME

In this section, we describe the data collection pipeline for MIME. An overview is shown in Figure 2. MIME is a video-based question answering benchmark that comprises of animations of 86 mimed actions, each with ten variants that are shown in Figure 3, resulting in a total of 860 eval-

<sup>1</sup>A neurological disorder that disrupts the ability to plan and execute purposeful movements, despite having the physical ability to do so.

<sup>2</sup>Data and code for MIME is available <https://justin-cho.com/mime>.

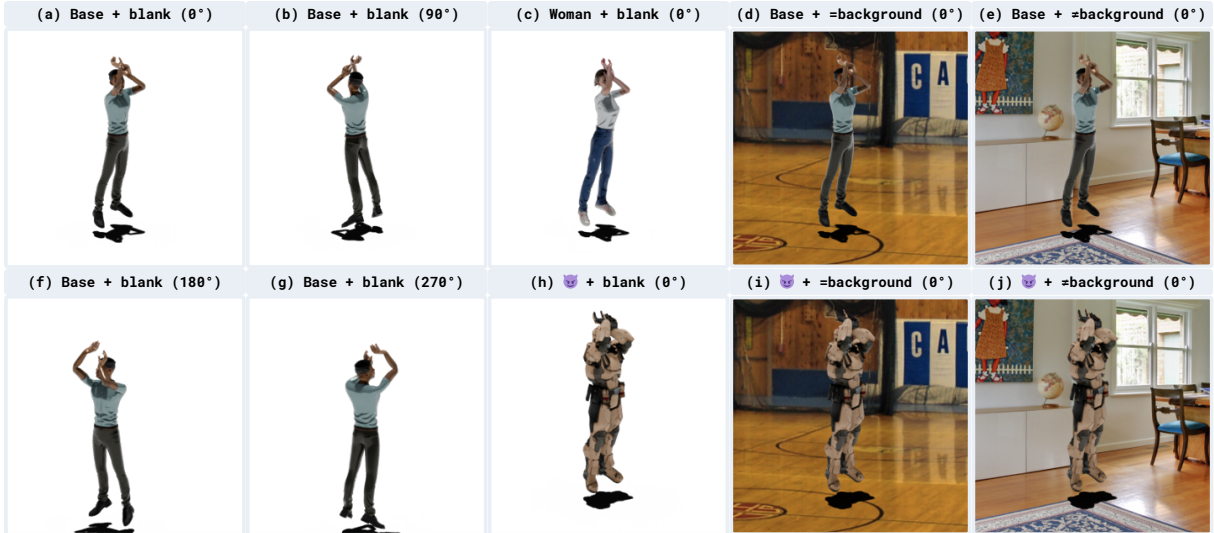


Figure 3: Overview of variations of each action in MIME. Our setup of using motion capture and computer graphics software allows us to flexibly permute different configurations for each action to ablate the robustness of a VLM’s understanding of mimed actions. (a, b, f, g) are examples of the same animation but with changes to the camera angle where different body parts become occluded depending on the angle. (c) and (h) only change the character from (a). (c) is a female human character while (h) is an adversarial character 🤖 in a sci-fi spacesuit. (d) and (i) are variants of (a) and (h) respectively with aligned backgrounds (=background, e.g., basketball court for basketball-related action) while (e) and (j) have adversarial backgrounds ( $\neq$ background, e.g., living room).

uation samples. The videos are rendered with 3D graphics software by combining digital assets with motion capture data of actors miming various actions. This setup is advantageous to alternative methods<sup>3</sup> for conducting a systematic study of recognition robustness with regards to various components that comprise an action as each action can be post-processed and remixed with different backgrounds, characters, and camera angles.

## 2.1 Collecting Motion Capture Data

First, we brainstorm 75 mimed action candidates for which salient context is missing. For example, playing a violin is a valid candidate because it is acted out without a violin and swimming is also valid because it is acted out without being in water, and both mimed actions are understood by human subjects. On the other hand, we exclude gestures such as hand-waving or thumbs-up as no salient context is missing in their enactment.

Next, we have two actors (one male nonprofessional actor and one female professional actor) act out these action candidates with three takes each. Each take introduces some variance of the same acts if there are multiple ways to perform them (e.g., swimming can be done with front stroke, back

<sup>3</sup>We discuss challenges with alternative methods, such as using live action footage and video generation models in Appendix F.

stroke, etc. and pushing can be done with various intensity) and if they are clearly distinct, multiple takes of the same action are kept. For more complex actions such as shot putting, the actors reference YouTube videos of professional athletes.

Only the motion capture data for which at least two out of three authors assign the same label to the final rendered output without seeing the action name are included in MIME. This process results in 47 action types and 86 mimed action samples. Additional technical details of our motion capture process is described in Appendix A.1.

## 2.2 Creating Blender Files

Motion capture data is imported into Blender and combined with digital assets to render frames with a transparent background so that they can be easily overlaid over our background of choice later without redundant rendering.

To efficiently combine various characters with a large number of motion capture data together, we write a Python-based macro that automates the process of creating blender files to be rendered. The result of the macro is shown in (2) of Figure 2. The detailed steps that our script automates are elaborated in Appendix A.2.

## 2.3 Rendering with Variations

**Characters** We use free 3D characters from Mixamo.<sup>4</sup> For the base setting of MIME, we use a male human character with casual clothes. To evaluate for mime recognition robustness with regards to the character, we also render with an adversarial character that is wearing a sci-fi spacesuit (shown in (h, i, j) in Figure 3. While we may choose even more adversarial characters that look less human to create a more challenging variant, we find that not all motion capture data is compatible for characters with largely diverging body proportions as the mimed action can become unrecognizable due to different body parts overlapping one another.

To test for a VLM’s robustness to the character’s gender, we also render with a female human character with casual clothes. The female character that we use is illustrated in (c) in Figure 3.

**Backgrounds** We use images from Creative Commons licensed images from Wikimedia<sup>5</sup> as aligned and misaligned backgrounds (e.g., (d, i) and (e, j) in Figure 3, respectively). We do our best to find images for which the background provides a large open space in the middle so that the full action sequence does not look awkward and the character does not appear disproportionately large or small.<sup>6</sup>

**Angle** To test robustness to viewpoints of the observed mimed action, we also render videos with various angles by rotating the camera with the character at the center. We select angles of 90°, 180°, and 270° rotations applied to the base setting. These are shown in (b, f, g) in Figure 3.

## 3 Experimental Setup

### 3.1 Evaluation Format

Prior work examine mime recognition under two different question answering conditions, the choice condition and naming condition (Osiurak et al., 2012; van Nispen et al., 2017), as the results between the two can differ significantly. The choice condition provides answer choices, which in effect

supplies contextual information, while the latter requires answering directly without any choices and is therefore more challenging and leads to lower agreement. Therefore, we construct MIME so that it evaluates VLMs with both of these conditions. We elaborate on the setup for each condition in the following.

**Choice condition: multiple choice (MC)** This is the best setting for computing accuracy as it can be done with exact match, but performance is dependent on how confusing the distractors are. Our multiple choice setup has four options to choose from and the distractors are selected by randomly sampling from other action labels that are included in MIME after removing the top 10 that have highest cosine similarities when compared with sentence embeddings (Reimers and Gurevych, 2019).<sup>7</sup> While this may make the multiple choice setup easier, it simplifies evaluation by preventing instances where there are multiple valid answers.

**Naming condition: Free-form short answers (FF)** In order to test model performance when it is not provided any context from the multiple choice options, we also assess their performance with a free-form short answer format. To assess the reference-based accuracy of our freeform answers, we adopt a single sentence-embedding cosine-similarity-based metric, effectively a relaxation of BertScore (Zhang et al., 2019), which is popular in VLM question answering-based evaluation of text-image similarity (Hu et al., 2023; Saxon et al., 2024). We use a sentence transformers model, the same one used for selecting distractors in the multiple choice format, to produce sentence-level embeddings of the generated free-form answers and gold labels, and use a heuristically-selected cosine similarity threshold of 0.5 to mark an answer as correct. While we find these to return a few false positives (e.g., baseball swing given credit for baseball pitch) and false negatives (e.g., pulling not given credit for dragging), we find these to be a small subset that does not significantly shift the overall performance of a model.

### 3.2 Models

We evaluate a comprehensive set of open- and closed-source VLMs with MIME to get a general understanding of whether VLMs can identify mimed activities.

<sup>4</sup><https://www.mixamo.com>

<sup>5</sup><https://commons.wikimedia.org/>

<sup>6</sup>While most images fulfill this criteria, there are a few for which it was not feasible to scale or crop properly so that the character ends up disproportionately large, such as the example shown in Figure 6 in Appendix D. However, we find this not to be an issue for humans to correctly identify the mimed action, and therefore consider reasonable evaluation samples and keep them in MIME.

<sup>7</sup>[sentence-transformers/all-MiniLM-L6-v2](https://sentence-transformers/all-MiniLM-L6-v2)

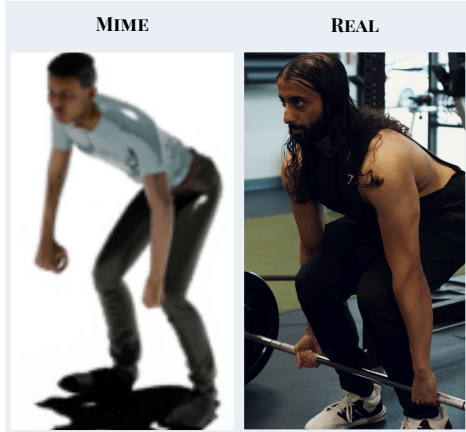


Figure 4: A frame from videos of deadlifting from MIME (left) and REAL (right). In MIME, salient context is missing (e.g., barbell and gym clothing).

For open-source models, we evaluate on (i) Qwen 2.5 VL Instruction (Bai et al., 2025), both 3B and 7B versions, (ii) InternVL 2.5 8B Instruct (Chen et al., 2024), (iii) Phi 3.5 VL Instruction, which is a 4.2B model released by Microsoft (Abdin et al., 2024). For closed-source models, we evaluate on (iv) Gemini 1.5 Flash from Google (Gemini, 2024) and (v) GPT-4o Mini from OpenAI.<sup>8</sup> For our first set of results, we use a zero-shot setting where the models are asked to directly predict the answer based on the video without any examples or reasoning steps. Our zero-shot prompt for multiple choice and free-form formats are shown in Appendix B.

### 3.3 Human Evaluation

We measure human performance on MIME to ensure that MIME is a tractable benchmark that humans perform well on and also confirm prior research that mimed action has low interpretation variability among humans (O’Reilly, 1995; Alexander et al., 2017; van Nispen et al., 2017; Little and Firestone, 2021). We recruit 60 internal participants from the University of Southern California’s Viterbi School of Engineering. They cover a wide demographic with eight unique nationalities, ages ranging from early 20s to mid 40s, and a 6:4 ratio of men to women. Although most are located in the same city, we believe their diverse international backgrounds provide a reasonable representation of general human performance. Each sample across all variations in MIME is completed by three participants. We share further detail on our human evaluation setup in Appendix D.

<sup>8</sup>[gpt-4o-mini-2024-07-18, https://platform.openai.com/docs/models/gpt-4o-mini](https://platform.openai.com/docs/models/gpt-4o-mini)

### 3.4 REAL

We ground the performance on MIME by measuring the performance on recognizing actions from live action footage (i.e., video created through traditional filmmaking techniques, capturing real actors, props, sets, and locations.) of the same set of actions in MIME. We collect a set of royalty- and copyright-free videos of such footage sourced from Pexels<sup>9</sup> and call it REAL. An example of a video from REAL and its corresponding sample in MIME is shown side by side in Figure 4. REAL functions as a control dataset that estimates a VLMs understanding of the actions that are mimed in MIME when all reasonable salient context is present. Therefore, the gap between performance on REAL and MIME serves as a proxy in the lack of generalizability in the understanding of the action to the understanding of its mimed counterpart. Note that while MIME contains 86 total mimed actions with multiple variations of the same activity, we only find one for each in REAL, and therefore REAL consists of 47 videos.

## 4 Results

### 4.1 MIME vs REAL

**Humans understand actions and their mimed counterparts equally well, while VLMs struggle significantly for the latter.** First, we share our results with the models mentioned in Section 3.2 on the base setting of MIME ((a) in Figure 3) and REAL in Figure 5.

Results on REAL clearly indicate that all VLMs are able to identify actions when all of the salient context is present (e.g., doing a deadlift in a gym with a barbell while wearing gym attire), achieving almost perfect scores for the MC while showing only a minor drop for the FF. This is on par with human performance.

However, on MIME, VLM performance drops sharply, while human performance remains consistent, with only a 0.4% drop in MC while there is a boost for FF by 12.3%. Upon manual inspection, we find that this is not because human performance is worse with live action footage, but rather because humans are more descriptive in their responses for FF for REAL and this produces more false negatives. Gemini 1.5 Flash shows the strongest performance, but even its accuracy is slightly over 50% in MC and less than 20% in FF.

<sup>9</sup><https://www.pexels.com/>

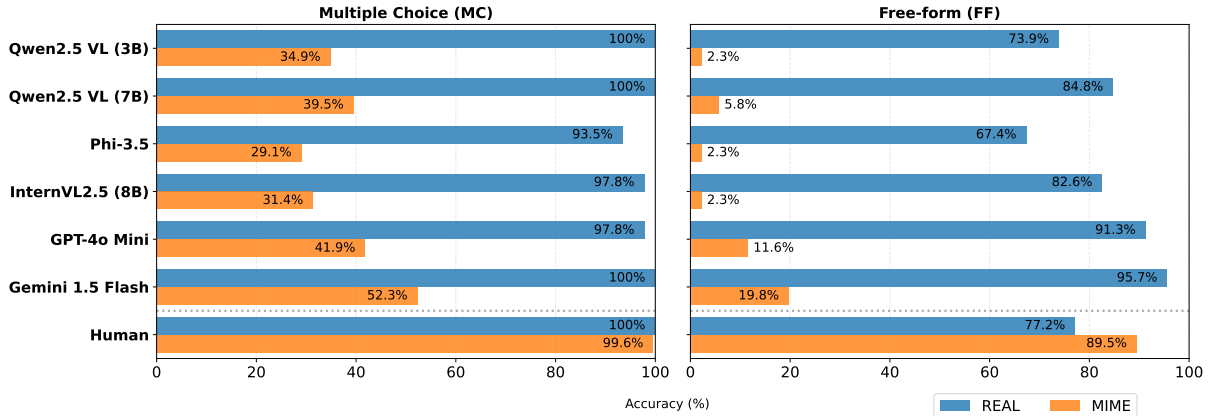


Figure 5: Performance comparison on the base setting of MIME and on the REAL dataset. Humans show equally strong performance on both MIME and REAL. VLMs struggle with MIME while achieving comparative performance on REAL, which suggests they lack a robust understanding of human actions.

Model	Base + blank		Base + =back.		Base + ≠back.		🐱 + blank		🐱 + =back.		🐱 + ≠back.	
	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF
Qwen 2.5 VL (3B)	34.9	2.3	61.6	30.2	27.9	0.0	30.2	1.2	60.5	29.1	24.4	0.0
Qwen 2.5 VL (7B)	39.5	5.8	<u>68.6</u>	38.4	32.6	1.2	34.9	0.0	64.0	30.2	30.2	0.0
Phi 3.5	29.1	2.3	<b>73.3</b>	27.9	31.4	<u>8.1</u>	<b>44.2</b>	0.0	<u>72.1</u>	27.9	<b>36.1</b>	<b>5.8</b>
InternVL2.5 8B	31.4	2.3	57.0	26.7	22.1	2.3	25.6	2.3	59.3	20.9	30.2	2.3
GPT-4o Mini	<u>41.9</u>	<u>11.6</u>	66.3	<u>39.5</u>	<b>37.2</b>	3.5	33.7	<b>8.1</b>	67.4	<u>33.7</u>	<b>36.1</b>	2.3
Gemini 1.5 Flash	<b>52.3</b>	<b>19.8</b>	<u>68.6</u>	<b>51.2</b>	<b>37.2</b>	<b>12.8</b>	<b>44.2</b>	<b>8.1</b>	<b>75.6</b>	<b>46.5</b>	<b>36.1</b>	<u>3.5</u>
Human	99.6	89.5	98.5	89.2	99.2	93.4	98.5	93.8	99.2	94.1	99.2	95.0

Table 1: Evaluation results on MIME for various perturbations. We use the same notations as Figure 3, with back. used as a shorthand for background. Refer to Figure 3 to view samples of each variation. Humans are robust to all variations, but VLMs drop performance for adversarial perturbations and get a significant boost when exposed to signals from the background that are aligned with the action.

## 4.2 Character and Background Variations

**Humans demonstrate similar performance across all variations, while VLMs benefit from contextual hints and suffer from adversarial perturbations.** The main advantage of MIME is the flexibility to swap out components of the animations in order to conduct ablation studies that shed light on the nature of the VLMs shortcomings.

We apply the perturbations shown in Figure 3 to test how performance is affected when the character and backgrounds are changed. Results from these perturbations with zero-shot are shown in Table 1.

The most noticeable result from this table is that the aligned background significantly boosts performance, even when the character is adversarial. With the direct opposite effect, changing the background to an adversarial one seriously harms performance for most models, but interestingly less so for the open-weight models. Interestingly, humans are extremely robust to all of the given perturbations, maintaining almost perfect scores on all MC settings while scoring at least 89.5% in the FF set-

tings. These results indicate that while humans are able to ignore irrelevant information and focus on the actions themselves, VLMs rely on other hints about the action present in the scene. These results are in line with the results on REAL.

## 4.3 Angle and Gender Variations

**VLMs demonstrate higher variance across angle and gender variations than humans.** Next, we share results with various angle perturbations to observe whether VLMs are viewpoint-agnostic for identifying mime. We see that humans clearly are consistent in this setting as well, as shown by the small variance in scores in the last row of Table 2. For the most part, MIME is challenging such that performance remains low regardless of the angle and there is no clearly preferred angle shared by VLMs. However, for MC, the variance in accuracy is much larger for VLMs than humans, another indication of a lack of robustness in VLMs in comparison to humans.

Lastly, although our dataset has been verified as

Model	Eval	Rotation Angle				Avg.	Std. ↓
		0°	90°	180°	270°		
Qwen 2.5 VL (3B)	MC	34.9	34.9	32.6	32.6	33.7	1.2
	FF	2.3	1.2	0.0	1.2	1.2	0.8
Qwen 2.5 VL (7B)	MC	39.5	39.5	50.0	43.0	43.0	4.3
	FF	5.8	7.0	3.5	8.1	6.1	1.7
Phi 3.5	MC	29.1	31.4	33.7	33.7	32.0	1.9
	FF	2.3	5.8	3.5	3.5	3.8	1.3
InternVL2.5 (8B)	MC	31.4	36.0	33.7	37.2	34.6	2.2
	FF	2.3	7.0	7.0	4.7	5.2	1.9
GPT-4o Mini	MC	41.9	47.7	43.0	47.7	45.1	2.6
	FF	11.6	15.1	13.9	13.9	13.7	1.3
Gemini 1.5 Flash	MC	52.3	47.7	52.3	53.5	51.5	2.2
	FF	19.8	18.6	17.4	23.3	19.8	2.2
Human	MC	99.6	98.8	98.8	98.7	99.0	0.4
	FF	89.5	95.0	90.7	85.1	90.1	3.5

Table 2: Performance on MIME for varying angles. For MC, relative to human performance, model performance varies largely depending on the viewpoint angle.

easily identifiable for humans by evaluators that span a balanced distribution across genders, we are interested in whether VLMs have any underlying gender biases that may affect their performance. Therefore, we only change the character to a female character and compare results. These results are shown in Table 3. As is the case in the angle variations, we also observe a lack of robustness in VLMs from the larger performance differences in the VLMs compared to that of humans. On a positive note, we do not observe a consistent preference for a particular gender by the VLMs.

## 5 Improving on MIME

Given the poor performance of VLMs on MIME, we are interested in whether simple methods can surface VLMs’ potential to understand mimed actions. In this section, we discuss our attempts to improve their performance via such methods.

### 5.1 Methods

The methods that we explore are the following: (i) **Chain-of-Thought** (CoT) is a method of producing a reasoning chain before making a final judgement. We ask the model to describe what it sees in detail and then provide its prediction (Wei et al., 2022). (ii) **Few-shot in-context learning** (Few-shot): For models that support few-shot in-context learning, we select three samples from the base configuration of MIME with minimal overlap in enacted actions (shooting a soccer ball, fishing, playing violin) and provide them as in-context examples that the models can leverage to improve their predictions on the remaining samples. (iv) **Fine-**

Model	Method	MC			FF		
		♂	♀	Δ ↓	♂	♀	Δ ↓
Qwen2.5 VL (3B)	Zero-shot	34.9	29.1	5.8	2.3	1.2	1.1
	CoT	43.0	37.2	5.8	0.0	2.3	2.3
Qwen2.5 VL (7B)	Zero-shot	39.5	41.9	2.4	5.8	9.3	3.5
	CoT	41.9	46.5	4.6	8.1	10.5	2.4
Phi-3.5	Zero-shot	29.1	34.9	5.8	2.3	2.3	0.0
	CoT	41.9	33.7	8.2	4.7	2.3	2.4
InternVL2.5 (8B)	Zero-shot	31.4	33.7	2.3	2.3	5.8	3.5
	CoT	25.6	24.4	1.2	1.2	5.8	4.6
GPT-4o Mini	Zero-shot	41.9	44.2	2.3	11.6	12.8	1.2
	CoT	43.0	53.5	10.5	16.3	10.5	5.8
	Few-shot	74.4	65.1	9.3	9.3	10.5	1.2
Gemini 1.5 Flash	Zero-shot	52.3	47.7	4.6	19.8	20.9	1.1
	CoT	54.6	52.3	2.3	22.1	19.8	2.3
	Few-shot	57.0	59.3	2.3	13.9	22.1	8.2
Human	-	99.6	98.5	1.1	89.5	90.3	0.8

Table 3: Performance comparison on MIME for gender variations. Δ is shown in blue if  $\sigma - \varphi \geq 0$  and in orange otherwise. Similar to angle variation results, results for VLMs vary largely depending on the gender without a consistent performance advantage of a certain gender, while human performance is consistent.

**tuning:** Lastly, we experiment with fine-tuning to see if fine-tuning on a small amount of data containing mimed actions can help models generalize to unseen ones. Since MIME only contains 86 samples in total per configuration, we fine-tune (FT) our model using a 5-fold validation approach with a 36/14/36 train/validation/test split. The details of these splits are present in Appendix C. Fine-tuning is conducted separately for each task type (free-form, and multiple choice). Due to limited compute, we limit our fine-tuning experiments to the base configuration and 🐱 + blank background configuration ((a) and (h) in Figure 3) and the Qwen 2.5 VL (3B, 7B) models and Phi 3.5 (4.2B). Refer to Appendix B for the few-shot and CoT prompts and Appendix C for further details of our fine-tuning setup.

### 5.2 Improvement Results

The main results of these preliminary methods are shown in Table 4. We observe that, apart from the API-based black box models, most methods do not lead to consistent and significant improvements over the results from zero-shot. One noticeable improvement is that of GPT-4o Mini when it is given few-shot examples, where results on most variations are boosted to over 50% for MC. While a smaller boost, we see a similar trend for Gemini 1.5 Flash. However, the performance for most cases still remain very low for FF, indicating that they continue to struggle without contextual information.

Model	Method	Base & blank		Base & =back.		Base & ≠back.		🐱 & blank		🐱 & =back.		🐱 & ≠back.	
		MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF
Qwen 2.5 VL (3B)	Zero-shot	34.9	2.3	61.6	30.2	27.9	0.0	30.2	1.2	60.5	29.1	24.4	0.0
	CoT	43.0	0.0	57.0	25.6	27.9	0.0	29.1	0.0	58.1	22.1	25.6	0.0
	FT†	31.6	0.0	-	-	-	-	22.0	0.0	-	-	-	-
Qwen 2.5 VL (7B)	Zero-shot	39.5	5.8	68.6	38.4	32.6	1.2	34.9	0.0	64.0	30.2	30.2	0.0
	CoT	41.9	8.1	62.8	37.2	31.4	3.5	27.9	0.0	61.6	17.4	26.7	1.2
	FT†	36.8	0.0	-	-	-	-	25.0	0.0	-	-	-	-
Phi 3.5 (4.2B)	Zero-shot	29.1	2.3	73.3	27.9	31.4	8.1	44.2	0.0	72.1	27.9	36.1	5.8
	CoT	41.9	4.7	64.0	30.2	24.4	1.2	31.4	1.2	59.3	30.2	33.7	2.3
	FT†	26.3	0.0	-	-	-	-	22.0	0.0	-	-	-	-
InternVL2.5 (8B)	Zero-shot	31.4	2.3	57.0	26.7	22.1	2.3	25.6	2.3	59.3	20.9	30.2	2.3
	CoT	25.6	1.2	60.5	23.3	32.6	2.3	26.7	1.2	52.3	15.1	23.3	0.0
GPT-4o Mini	Zero-shot	41.9	11.6	66.3	39.5	37.2	3.5	33.7	8.1	67.4	33.7	36.1	2.3
	CoT	43.0	16.3	73.3	47.7	44.2	8.1	44.2	4.7	65.1	38.4	36.1	1.2
	Few-shot†	74.4	9.3	94.2	39.5	52.3	0.0	70.9	2.3	89.5	40.7	59.3	0.0
Gemini 1.5 Flash	Zero-shot	52.3	19.8	68.6	51.2	37.2	12.8	44.2	8.1	75.6	46.5	36.1	3.5
	CoT	54.7	22.1	69.8	48.8	40.7	11.6	48.8	9.3	74.4	51.2	41.9	7.0
	Few-shot†	57.0	14.0	72.1	41.9	46.5	10.5	48.8	4.7	77.9	39.5	44.2	0.0
Human	-	99.6	89.5	98.5	89.2	99.2	93.4	98.5	93.8	99.2	94.1	99.2	95.0

Table 4: Results for various methods to improve performance on MIME. The table follows the same format as Table 1. †Refer to §5.1 for details on the experimental setup for few-shot and fine-tuning results for which the number of evaluation samples is smaller. Accuracy is shown in blue for methods that is higher than the corresponding zero-shot score and in orange otherwise.

Overall, our results demonstrate that there is ample room for improvement for VLMs to acquire an understanding of human gestures that is as robust as those of humans.

### 5.3 Failure Mode Analysis

In order to understand where improvement opportunities lie, we analyze the modes of failure by Gemini 1.5 Flash with CoT on FF to examine the reasoning they generate for making predictions without contextual hints provided by multiple choice options. The reasoning serves as a proxy of what the VLM observes and thus analyzing it can surface the point of failure that needs to be corrected. We want to know whether the models fail to correctly describe the shown action or whether they can accurately describe it but cannot interpret it as the intended mimed action, and also how much they are affected by the aligned and misaligned backgrounds. We manually categorize the modes of failure for predictions of the first three columns in Table 4 (i.e., using the base character with blank, aligned, and misaligned backgrounds).

We find that, with the blank background, Gemini 1.5 Flash generates a description of the shown mimed action that is only partially correct 54% of the time and completely incorrect 16% of the time (e.g., *They wind up their arm as if holding*

*a ball, then perform a throwing motion with their arm and hand extending forward* for arm curls). In 13% of instances, the description is correct, but it is interpreted incorrectly, leading to an incorrect prediction (e.g., predicts bowling a ball after generating *They start with a wind-up motion, bringing their arm back, then swing forward as if releasing a ball* for baseball pitch). When shown an aligned background, 43% of predictions that were incorrect with the blank background become correct predictions. With the misaligned background, 24% of all predictions are confused by irrelevant context provided by the background (e.g., predicts conducting an orchestra for climbing given a concert hall background), which leads to a drop in proportion of predictions that had completely or partially correct descriptions of the shown mimed action.<sup>10</sup>

## 6 Related Work

### 6.1 Nonverbal Communication

One major branch of NVC research leverages NVC to enhance predictions for a downstream task, such

<sup>10</sup>We share the full statistics and more examples of each mode of failure in Appendix E. Other VLMs also show similar failure patterns, with most instances of failure caused by partial or completely incorrect descriptions of the mimed actions. Overall, these results indicate that future research should prioritize training VLMs that can accurately describe the human gestures they observe.



as using posture, prosodic features, and facial expressions to predict dialogue acts (Sridhar et al., 2009; Boyer et al., 2011; Ha et al., 2012), gaze, head movement, and breath patterns to detect turn-taking and engagement behavior (Jokinen, 2010; Ishii et al., 2013, 2014, 2015, 2016a,b), visual information and motion capture data for emotion representations and predictions (Busso et al., 2008; Zhang et al., 2023). There are a few prior work that seeks to predict NVC with gestures, but they are constrained to those expressed with limited body parts, such as hands for hand gestures (Burke and Lasenby, 2015; Kapitanov et al., 2024) or sign language (Papastratis et al., 2021; Kezar et al., 2023). Others use verbal signals to predict or generate NVC, usually in the context of developing realistic virtual agents (Graf et al., 2002; Busso et al., 2007) or robots (Shamsuddin et al., 2011; Sakai et al., 2015; Cass et al., 2018), such as using dialogue acts and affective information to predict nods (Lee and Marsella, 2010; Ishii et al., 2018). In contrast, MIME examines whether machine learning models have a robust understanding of explicit full-body gestures, a fundamental prerequisite to comprehending more variable and subtle gestures in the full spectrum of NVC, by evaluating whether they can identify mimed actions — a subset of NVC with low interpretation variability.

## 6.2 Action Recognition

While mimed action understanding is an instrumental step towards general NVC understanding, it is also highly related to action understanding. Many datasets exist for evaluating whether machine learning models understand human actions (Kong and Fu, 2022; Sun et al., 2022), with early work focused on sporting actions (Kuehne et al., 2011; Soomro et al., 2012; Karpathy et al., 2014; Idrees et al., 2017) and recent work expanding to a larger scope and scale, including daily activities that are crowdsourced (Sigurdsson et al., 2016; Damen et al., 2018) and extracted from YouTube (Heilbron et al., 2015; Xu et al., 2016; Krishna et al., 2017; Kay et al., 2017; Sanabria et al., 2018; Zhou et al., 2018; Miech et al., 2019; Weinzaepfel and Rogez, 2021), Tumblr (Li et al., 2016), Flickr (Anne Hendricks et al., 2017), or movies (Torabi et al., 2015; Rohrbach et al., 2015, 2017). Miech et al. (2020) takes a step further to evaluate action recognition robustness with a dataset that contains rare activities (e.g., blending phone and cutting keyboard). While these datasets collectively cover more than

hundreds of different classes, none of them contain mimed actions. As such, even if VLMs perform well on these datasets, it is unclear whether they have a robust understanding of complex human body motions or if they are relying on spurious correlations provided by the salient context.

The CMU-MMAC Database (la Torre et al., 2008) is similar to MIME in that it contains animated videos of motion capture data, but the animations are extremely simple, featuring a black background with yellow stick figures, and do not include finger and thumb joints, which lowers fidelity to the captured actions. IEMOCAP (Busso et al., 2008) is also based on motion capture data but it is focused on emotion prediction in dyadic conversations and motion capture is only collected for the face, head, and hands. van Nispen et al. (2017) presents a dataset of footage of human participants pantomiming various objects, but it is limited to hand gestures. Lastly, the Mimetics dataset (Weinzaepfel and Rogez, 2021) is the most similar to MIME in that it contains live action footage videos of mimed actions extracted from YouTube with varying amounts of relevant context provided. However, it lacks the flexibility to systematically adjust the amount of relevant context provided in each video with precision for conducting the ablative analysis possible with MIME. Moreover, the Mimetics dataset may be included in the training of VLMs as the videos are from YouTube, and therefore MIME serves as a more reliable benchmark for mimed action understanding unaffected by data leakage.

## 7 Conclusion

We introduce MIME, a novel video-based question answering benchmark that consist of animations of 86 mimed actions created with motion capture data and 3D graphics software. MIME contains systematic perturbations in character, background, and viewpoint to assess the robustness of VLMs’ understanding of full-body gestures. While humans demonstrate almost perfect accuracy and remain highly robust to all modifications in MIME, both open-weight and API-based VLMs struggle, particularly in the free-form format where recognition accuracy approaches zero under adversarial perturbations. These results highlight the need for further research in enhancing VLMs with a robust understanding of human gestures to establish a crucial bedrock for NVC comprehension.

## Acknowledgment

We thank David Nelson and the Mixed Reality Lab at the Institute for Creative Technologies for providing access to their Vicon system, which was an indispensable equipment for collecting the motion capture data used in creating MIME.

## Limitations

**Lack of Photorealism** The main limitation of MIME is that it is not photorealistic as it contains animated videos of motion capture data. This lack of realism may introduce a domain shift for VLMs for which the majority of the training data is likely to be live action footage rather than animations, leading to an artificially discounted performance of VLMs on MIME. However, given that humans can successfully interpret mimed actions in MIME, we argue that models should also be able to achieve comparable performance to humans on MIME if they develop a robust understanding of human gestures by generalizing from what they learn through more photorealistic content.

The flip side of this concern is that performance on MIME do not translate to equivalent performance on mimed actions captured as live action footage. We believe that this concern is addressed to a reasonable extent in that MIME contains multiple perturbations of the same set of actions with varying characters, backgrounds, and viewpoints. It would be unlikely for a VLM to achieve high accuracy on all of these variants without a robust understanding of human actions that do not translate to understanding of these actions in live action footage. In addition, as the fidelity of digital assets improve and with the availability of more compute, we will be able to create versions of MIME that are increasingly photorealistic, which further mitigates this concern. Therefore, despite concerns arising from the lack of photorealism, we argue that MIME is the most advantageous for systematic analysis of robustness because of the ease of producing variants that enable ablation studies. Alternative methods, i.e., using live action footage or video generation models, are significantly limited in being able to modify equivalent mimed actions at the same level of flexibility and consistency. Refer to a detailed discussion on alternative methods in Appendix F.

**Representation Bias** Next, MIME only consists of animations that are based on motion capture data

from two actors. However, one of these actors is an Asian male non-professional actor and the other is an European female professional actor. Despite minimal overlap in demographics of these two actors, our human evaluation results show that there is not a higher recognition accuracy for samples that come from one actor over those of another.<sup>11</sup> In other words, there may be differences in how people enact actions as mimes, but they are not significant enough to affect recognition, at least for the pool of 60 participants that we recruited. We believe these results should address concerns of representation bias of the mimed actions in MIME.

**Data contamination in REAL** Another caveat of our results that compare VLM performance between MIME and REAL is that samples in REAL may have been included in the training of the VLMs that we evaluated, thus inflating their results on REAL due to data contamination. Also, an ideal systematic study of a model’s understanding of actions would have entailed studying REAL without contextual information provided by the background, but we could not pursue this path due to technical challenges preventing background removal in REAL. Unfortunately, the VLMs that we study in this paper do not share the full scope of their training data, and therefore we cannot confirm whether strong performance on REAL is due to data leakage or because they can reliably identify actual actions (as opposed to mimed actions) from live action footage.

**Fine-tuning result caveats** A noteworthy limitation in our improvement results from §5 is that our fine-tuning experiments do not provide conclusive evidence regarding the effectiveness of fine-tuning for improving model performance on MIME. Our fine-tuning experiments are an attempt at domain adaptation using a limited sample size, which likely leads to overfitting, preventing the model from achieving meaningful generalization. While this does not rule out the potential benefits of fine-tuning on larger and more diverse datasets, our findings suggest that additional research is necessary to explore optimal fine-tuning strategies for tasks as challenging as MIME.

---

<sup>11</sup>99.0% vs. 98.8% in multiple choice format and 92.6% vs. 90.6% for free-form format. A t-test indicates insignificant differences in these accuracies at  $p < 0.01$  ( $p = 0.71$  and  $p = 0.03$ , respectively).

## Ethical Considerations

While MIME serves as an important milestone for VLMs to reach on their path to commanding fluent NVC, strong performance on MIME should be interpreted with caution. First, it is important to note that the set of mimed actions explored in MIME is not exhaustive of actions that can be possibly mimed. It is a carefully curated subset that we find high agreement among human participants with diverse backgrounds and therefore propose as one of the lowest-hanging fruits in NVC recognition. In other words, strong performance on MIME should be considered a prerequisite being met for VLMs that can be further improved to understand and also generate the more nuanced forms of NVC. It should not be interpreted incorrectly as an indication that VLMs can command NVC fluently and thus is ready to be applied to downstream tasks that require such skills. Future work should expand the scope on MIME to include more nuanced gestures that potentially have relatively lower universal agreement but nonetheless have high intracultural agreement.

In addition, with concerns of test data leakage on the rise (Zhou et al., 2023; Xu et al., 2024; Jiang et al., 2024), performance improvement on MIME may not be indicative of robust understanding of human gestures if improvement on MIME is achieved simply by training more on a data distribution similar to MIME instead of improving the generalizability of VLMs. Therefore, it is important for VLM developers to be cautious and transparent about how they source visual data to prevent misleading performance gains on MIME. Luckily, our pipeline for creating MIME can be easily replicated for producing unseen permutations of mimed actions in MIME to further test the limits of generalizability if there is suspected data contamination.

## References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Simon Alexanderson, Carol O’Sullivan, Michael Neff, and Jonas Beskow. 2017. Mimebot—investigating the expressibility of non-verbal communication across agent embodiments. *ACM Transactions on Applied Perception*, 14(4).
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.
- Michael Argyle and Peter Trower. 1979. Person to person: ways of communicating. (*No Title*).
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Siboz Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Kristy Boyer, Joseph Grafsgaard, Eun Young Ha, Robert Phillips, and James Lester. 2011. [An affect-enriched dialogue act classification model for task-oriented](#)

- dialogue. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1190–1199, Portland, Oregon, USA. Association for Computational Linguistics.
- Michael Burke and Joan Lasenby. 2015. [Pantomimic gestures for human–robot interaction](#). *Trans. Rob.*, 31(5):1225–1237.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.
- Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth S. Narayanan. 2007. [Rigid head motion in expressive speech animation: Analysis and synthesis](#). *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1075–1086.
- Aaron G. Cass, Kristina Striegnitz, and Nick Webb. 2018. [A farewell to arms: Non-verbal communication for non-humanoid robots](#). In *IEEE/ACM International Conference on Human-Robot Interaction*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736.
- Starkey Duncan Jr. 1969. Nonverbal communication. *Psychological bulletin*, 72(2):118.
- Michael Eaves and Dale G Leathers. 2015. Successful nonverbal communication: Principles and applications.
- Howard S Friedman. 1979. Nonverbal communication between patients and medical practitioners. *Journal of Social Issues*, 35(1):82–99.
- Gemini. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). Preprint, arXiv:2403.05530.
- Hans Peter Graf, Eric Cosatto, Volker Strom, and Fu Jie Huang. 2002. [Visual prosody: facial movements accompanying speech](#). *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 396–401.
- Eun Young Ha, Joseph F. Grafsgaard, Christopher Mitchell, Kristy Elizabeth Boyer, and James C. Lester. 2012. [Combining verbal and nonverbal features to overcome the “information gap” in task-oriented dialogue](#). In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 247–256, Seoul, South Korea. Association for Computational Linguistics.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nieves. 2015. [Activitynet: A large-scale video benchmark for human activity understanding](#). *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.
- Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.
- Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorbunov, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23.
- Ryo Ishii, Ryuichiro Higashinaka, and Junji Tomita. 2018. [Predicting nods by using dialogue acts in dialogue](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. [Predicting next speaker based on head movement in multi-party meetings](#). *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2319–2323.
- Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. 2013. [Predicting next speaker and timing from gaze transition patterns in multi-party meetings](#). In *International Conference on Multimodal Interaction*.
- Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2014. [Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings](#). *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 694–698.
- Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016a. [Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings](#). *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 6:1 – 31.
- Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016b. [Using respiration to predict who will speak next and when in multiparty meetings](#). *ACM Trans. Interact. Intell. Syst.*, 6:20:1–20:20.

- Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*.
- Kristiina Jokinen. 2010. [Non-verbal signals for turn-taking and feedback](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Krainov, and Andrei Makhliarchuk. 2024. Hagrid – hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4572–4581.
- Piyali Karmakar and Manjira Sinha. 2024. [Aiding non-verbal communication: A bidirectional language agnostic framework for automating text to AAC generation](#). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 324–331, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLP AI).
- Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.
- Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023. [Improving sign recognition with phonology](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167.
- Yu Kong and Yun Fu. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.
- Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE.
- Fernando De la Torre, Jessica K. Hodgins, Adam W. Bargteil, Xavier Martin, J. Robert Macey, Alex Tusell Collado, and Pep Beltran. 2008. [Guide to the carnegie mellon university multimodal activity \(cmu-mmac\) database](#).
- Jina Lee and Stacy Marsella. 2010. [Predicting speaker head nods and the effects of affective information](#). *IEEE Transactions on Multimedia*, 12:552–562.
- Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650.
- Patrick C Little and Chaz Firestone. 2021. Physically implied surfaces. *Psychological Science*, 32(5):799–808.
- Marianne Schmid Mast. 2007. On the importance of nonverbal communication in the physician–patient interaction. *Patient education and counseling*, 67(3):315–318.
- David Matsumoto and Hyisung C Hwang. 2013. Cultural similarities and differences in emblematic gestures. *Journal of Nonverbal Behavior*, 37:1–27.
- David McNeill. 1992. Hand and mind: What gestures reveal about thought.
- Albert Mehrabian. 1972. Nonverbal communication.
- Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. [Rareact: A video dataset of unusual interactions](#). *ArXiv*, abs/2008.01018.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.
- Anne Watson O’Reilly. 1995. Using representations: Comprehension and production of actions with imagined objects. *Child development*, 66(4):999–1010.
- François Osiurak, Christophe Jarry, Nicolas Baltenneck, Bertrand Boudin, and Didier Le Gall. 2012. Make a gesture and i will tell you what you are miming: pantomime recognition in healthy subjects. *cortex*, 48(5):584–592.
- Ilias Papastratis, Kosmas Dimitropoulos, and Petros Daras. 2021. [Continuous sign language recognition through a context-aware generative adversarial network](#). *Sensors (Basel, Switzerland)*, 21.

- Chanjun Park, Yoonna Jang, Seolhwa Lee, Jaehyung Seo, Kisu Yang, and Heuseok Lim. 2022. [PicTalky: Augmentative and alternative communication for language developmental disabilities](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 17–27, Taipei, Taiwan. Association for Computational Linguistics.
- Fernando Poyatos. 1983. Language and nonverbal systems in the structure of face-to-face interaction. *Language & Communication*, 3(2):129–140.
- Haoxuan Qu, Yujun Cai, and Jun Liu. 2024. Lfms are good action recognizers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18395–18406.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.
- Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123:94–120.
- Kurima Sakai, Carlos Toshinori Ishi, Takashi Minato, and Hiroshi Ishiguro. 2015. [Online speech-driven head motion generating system and evaluation on a tele-operated robot](#). *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 529–534.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metzger. 2018. How2: A large-scale dataset for multimodal language understanding. In *NeurIPS*.
- Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. 2024. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (ts2). *arXiv preprint arXiv:2404.04251*.
- Zoya Shafique, Haiyan Wang, and Yingli Tian. 2023. Nonverbal communication cue recognition: A pathway to more accessible communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5666–5674.
- Syamimi Shamsuddin, Luthffi Idzhar Ismail, Hanafiah Yussof, Nur Ismarrubie Zahari, Saiful Bahari, Hafizan Hashim, and Ahmed Jaffar. 2011. Humanoid robot nao: Review of control and motion exploration. In *2011 IEEE international conference on Control System, Computing and Engineering*, pages 511–516. IEEE.
- Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan. 2009. [Combining lexical, syntactic and prosodic cues for improved online dialog act tagging](#). *Comput. Speech Lang.*, 23:407–422.
- Theodore Stickley. 2011. From soler to surety for effective non-verbal communication. *Nurse education in practice*, 11(6):395–398.
- Xingwu Sun, Yanfeng Chen, and Yiqing Huang. 2024. [Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent](#). *Preprint*, arXiv:2411.02265.
- Zehua Sun, Qihong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225.
- Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Ole Tange. 2024. [Gnu parallel 20240522 \('tbilisi'\)](#).
- Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*.
- Indrit Troshani, Sally Rao Hill, Claire Sherman, and Damien Arthur. 2021. Do we trust in ai? role of anthropomorphism and intelligence. *Journal of Computer Information Systems*, 61(5):481–491.

- Karin van Nispen, W Mieke E van de Sandt-Koenderman, and Emiel Krahmer. 2017. Production and comprehension of pantomimes used to depict objects. *Frontiers in Psychology*, 8:1095.
- Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. 2023. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Zijian Zhang Weijie Kong, Qi Tian. 2024. [Hunyuan-video: A systematic framework for large video generative models](#).
- Philippe Weinzaepfel and Grégory Rogez. 2021. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690.
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. [Video-CLIP: Contrastive pre-training for zero-shot video-text understanding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msrvt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.
- Yang Xu, Yang Cheng, and Riya Bhatia. 2022. [Gestures are used rationally: Information theoretic evidence from neural sequential models](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 134–140, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Sitao Zhang, Yimu Pan, and James Z Wang. 2023. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18993–19004.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don’t make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.
- Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

## Appendix

### A MIME Details

#### A.1 Motion Capture Technical Details

We collect motion capture with actors wearing motion capture suits configured in the Vicon 10 finger marker setup, in addition to the standard 53 body marker setup.<sup>12</sup> Motion capture is performed on a Vicon stage configured with Vero capture cameras driven by Vicon Shogun 1.11. An example of a single frame from the resulting motion capture data is shown in (1) of [Figure 2](#). Finally, the dataset is batch cleaned, post-processed, and exported via Shogun Post into FBX format for further processing in Blender.

#### A.2 Blender Macro Script

Our script imports the character and motion capture armature, adjusting their resting positions to be as aligned as possible using the MCATS plugin,<sup>13</sup> and using the Rokoko Studio Live plugin<sup>14</sup> to retarget the animations from the motion capture data to the character. In addition, a sun light source and large plane at the feet level of the character are added for shadow capture for more realistic videos. Lastly, a camera is added so that videos can be rendered from the camera’s viewpoint. We select a conservatively zoomed out viewpoint in order to make sure that the full action sequence is captured in the rendered output.

#### A.3 Render settings

Each frame is rendered with the following rendering configurations:

- Number of samples: 32
- Maximum number of light bounces: 1
- Resolution: 1280 × 720

<sup>12</sup><https://help.vicon.com/space/Shogun112/31229851/Place+markers+on+a+performer>

<sup>13</sup><https://github.com/absolute-quantum/cats-blender-plugin>

<sup>14</sup><https://github.com/Rokoko/rokoko-studio-live-blender>

- Adaptive threshold: 0.5
- Denoise using GPU: True
- Use persistent data: True
- Caustics reflective: False
- Caustics refractive: False
- Use light tree: Falses

We find these settings to strike a reasonable balance between video quality and render time.

We process rendering jobs in parallel on P100 and V100 GPUs, depending on availability. The final step of overlaying the frames with transparent backgrounds over various backgrounds are accelerated with parallel(Tange, 2024). Generative AI workloads were run locally on an RTX 4090.

## B Prompt Details

We provide the templates for our prompts here:

### B.1 Zero-shot Multiple Choice

What action is the person miming in this image/video?  
Choose the most accurate description from the options below.

- A. {options[0]}
- B. {options[1]}
- C. {options[2]}
- D. {options[3]}

Respond with just a single letter (A, B, C, or D).

### B.2 Zero-shot Free-form

What action is the person miming in this image/video?  
Describe the action in a single short phrase (under 5 words).

You can think out the action in a chain of thought, but please reply on the final line of your response, a single short phrase (under 5 words).

This action is being 'mimed' meaning backgrounds or objects that are relevant may not be present. Think about only the \*action\* taking place in the video, and give a response for what it looks like the character is "acting out" or doing "charades" of.

### B.3 CoT Multiple Choice

What action is the person miming in this image/video?  
Choose the most accurate description from the options below.

- A. {options[0]}
- B. {options[1]}
- C. {options[2]}
- D. {options[3]}

Carefully think through the answer, by detailing

the particular actions and movements that you see the person doing. Your output should contain your explanation, and then on a new line, a single letter corresponding to the answer you choose, with no punctuation. An example response is shown below:

'In the video, the person is moving a single arm back and forth, as if they are swinging a bat. This action is most accurately described by option B.

B'

### B.4 CoT Free-form

What action is the person miming in this image/video?  
Carefully think through the answer, by detailing the particular actions and movements that you see the person doing.

This action is being 'mimed' meaning backgrounds or objects that are relevant may not be present. Think about only the \*action\* taking place in the video, and give a response for what it looks like the character is "acting out" or doing "charades" of. Your output should contain your explanation, and then on a new line, a short phrase (under 5 words) corresponding to your answer, with no punctuation or answer prefix such as 'Answer:'

### B.5 Few-shot ICL Multiple Choice

What action is the person miming in this video?  
Choose from:

- A. {options[0]}
- B. {options[1]}
- C. {options[2]}
- D. {options[3]}

Answer with just a single letter (A, B, C, or D).  
Answer: <answer>

What action is the person miming in this video?  
Choose from:

- A. {options[0]}
- B. {options[1]}
- C. {options[2]}
- D. {options[3]}

Answer with just a single letter (A, B, C, or D).  
Answer: <answer>

...  
What action is the person miming in this video?  
Choose from:

- A. {options[0]}
- B. {options[1]}
- C. {options[2]}
- D. {options[3]}

Answer with just a single letter (A, B, C, or D).

### B.6 Few-shot ICL Free-form

What action is the person miming in this video?  
Describe the action in a single short phrase.  
Answer: <answer>

What action is the person miming in this video?  
Describe the action in a single short phrase.  
Answer: <answer>

...  
What action is the person miming in this video?  
Describe the action in a single short phrase.



## C Fine-tuning Details

The  $n = 5$  folds that we use for N-fold training for fine-tuning experiments are shown in Table 5. During FT, only the vision encoder is trained, while the text encoder remains frozen. We train for 7 epochs with an initial learning rate of  $2e-5$ , following a cosine learning rate schedule. The batch size is set to 8, and we use the AdamW optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . To optimize the balance between computational speed and precision, BF16 and TF32 are enabled. All models are trained using  $2 \times$  A100 GPUs.

## D Human Evaluation Details

Without any prior guidance, each participant is asked to answer the question in free-form format first after watching a sample in MIME and then answer the same question with the multiple choice format. The interface shown to our participants is illustrated in Figure 6. While this setup is efficient for collecting both free-form and multiple choice format results from a single participant, options shown for the multiple choice format in prior samples may provide contextual information for free-form format answers in the remaining samples. However, we find this effect to be negligible in comparison to the large performance gap between humans and models: human performance with the free-form format on the first samples that they annotate is  $\sim 88\%$  while that of the remaining samples is  $\sim 93\%$ . Each participant annotates half of the total samples (43) in one of the configurations shown in Figure 3 in order to keep annotator burden low.

## E Failure Mode Analysis Details

Distribution of failure modes and examples of each mode are shown in Table 6.

## F Alternative Methods for Creating MIME

### F.1 Live Action Footage

The ideal setup that does not introduce domain shift is to create an equivalent of MIME that contains live action footage. However, we intentionally avoid this option because of the difficulty to create systematic variations of the same mimed actions that enable robustness analysis and concerns of privacy of the actors that would be included in said dataset. Our attempts with removing objects and replacing backgrounds in each frame of the videos in REAL

produced inconsistent results, and even if they were consistent, we would need another method to perturb the actor in a way that the resulting footage remains photorealistic.

### F.2 Video Generation Models

We also explore video generation models for creating MIME and show sample outputs in Figure 7. For paid services, we test Sora<sup>15</sup> and Runway<sup>16</sup>, and for open-weight models, we use a variety of Hunyuan (Sun et al., 2024) fp16 and bf8 models using ComfyUI’s<sup>17</sup> recommended text-to-video Hunyuan workflow (Weijie Kong, 2024). Despite various prompts, all video models struggle to generate mimed actions and generate the action with the salient context still present in the video, even when explicitly asked not to include it (see Figure 7c) or ensuring it is not mentioned in the prompt (see Figure 7a). We also try prompts that are generated by language models, such as the output for the prompt: “Generate a prompt for a video generation model to generate a video of someone miming fencing such that the resulting video does not include any fencing equipment”. While this avoids producing salient context in some cases, it fails to produce a video that matches the intended action (e.g., dancing move shown for a prompt for fencing Figure 7b).

<sup>15</sup><https://openai.com/sora/>

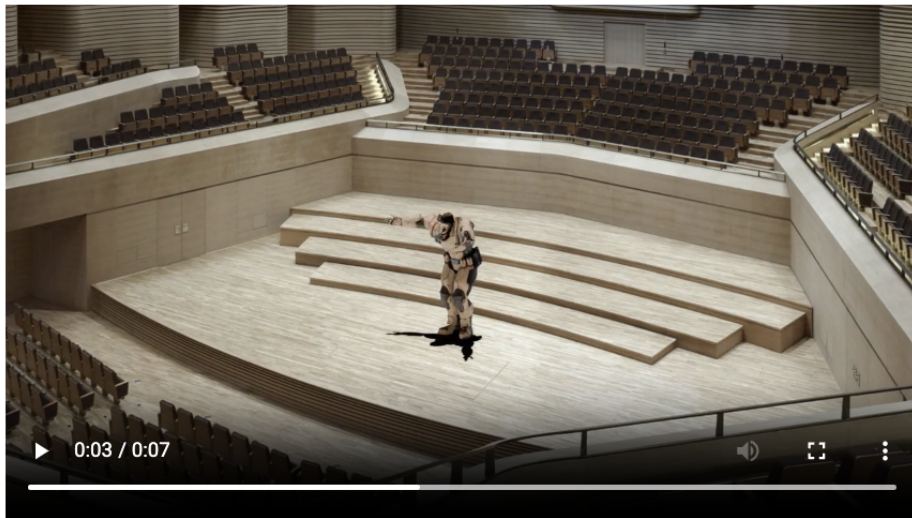
<sup>16</sup><https://runwayml.com/>

<sup>17</sup><https://github.com/comfyanonymous/ComfyUI>

## What am I Miming?

- In this experiment, you will be shown a video, and you must determine the action that is being mimed by the person in the video.
- Type the action that you think is being mimed in the text box.

Progress: 1 / 43



What action is being mimed by the person here? Answer in 1-2 words.

**Select the correct action from the following options**

**Option 1:** breast stroke swimming

**Option 2:** boxing uppercut

**Option 3:** dragging

**Option 4:** alternating single arm curls

**Option 1**

Option 2

Option 3

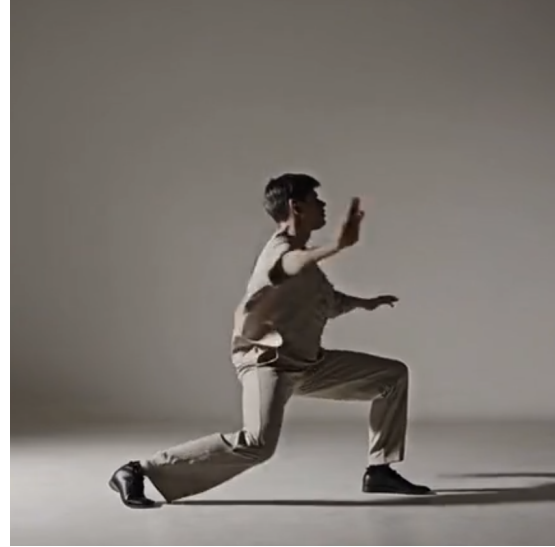
Option 4

Next question

Figure 6: Our interface for human evaluation. The evaluators can only attempt to answer the question after seeing the full video. After answering a free-form short answer question, they are asked to complete a multiple choice equivalent before moving on to the next sample.



(a) OpenAI Sora's output with prompt: still shot without background of someone miming typing sitting by a desk without any objects on it.



(b) OpenAI Sora's output with LM-generated prompt: Generate a high-quality video of a person performing mime movements that resemble fencing. The individual should use expressive body language, dynamic footwork, and precise hand gestures to create the illusion of fencing without any actual fencing equipment, such as swords or protective gear. The performance should be fluid and theatrical, emphasizing exaggerated parries, lunges, and ripostes to convey the essence of fencing through mime alone. The person should be dressed in neutral or casual clothing suitable for a performance, with a simple background that keeps the focus on their movement.



(c) Runway's output with prompt: Generate a video of a person miming a fencing match without any fencing equipment. The person should perform precise exaggerated fencing movements such as lunges, parries, and ripostes. Their footwork should be light and agile, moving back and forth as if engaged in a real bout.



(d) Hunyuan-Large's (Sun et al., 2024) output with prompt: Man acting like shooting an arrow without anything in his hands. This should be a mimed action without any props.

Figure 7: Snapshots of outputs from various video generation models to generate mimed actions. All models that we tested failed to produce videos that either did **not** include the action's key object (e.g., keyboard while typing, bow and arrow while shooting an arrow) or correctly act out the intended action.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Volleyball001	Climbing001	DrinkingCoffee001	ConsoleGaming01	ArmCurls001
VolleyballServe	Climbing01	ShootingAHandgun001	Darts001	ArmCurls01
WeightedSquat002	DeadLift001	ShootingARifle001	Bowling003	ArmCurls03
CheckingWatch001	Deadlift01	ShootingHandgun01	Bowling01	Baseball004
CheckingWatch01	Archery001	Basketball001	Weightlifting001	BaseballPitch002
Swimming001	Archery01	BasketballLayup001	Violin002	BaseballPitch02
Swimming002	Driving002	BasketballLayup02	ShotPut001	CheckingPhone002
Swimming03	Driving003	BasketballShot02	ShotPut01	WatchingTV01
Swimming04	Soccer003	Boxing001	DrivingSitting001	SittingAndWriting001
Swimming06	SoccerShot01	Boxing03	DrivingSittingDown03	TakingPhotoWithCamera001

Table 5: The action IDs in MIME that are divided into five folds we use for our fine-tuning setup.

Configuration	Mode	%	Mimed Action	Prediction	Generated CoT Description
Base +blank	○	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.
	✓	3.5	Backstroke swimming	Falling down	The person in the video is lying on their back. Their arms are initially by their sides, then they slowly raise their arms above their head, then bring them back down to their sides. Their legs are spread slightly apart. The movements suggest the person is acting out falling and lying on the ground.
Base =back.	○	18.6	Baseball swing	Pitching a baseball	They go through the windup motion, bringing their arm back and then forward in a throwing motion. Their legs and body also shift their weight from one foot to the other, mimicking the motion of a pitcher.
	🕒	23.3	Playing piano	Pulling something heavy	The person in the video is crouched down, and moving their arms and hands in a way that mimics pulling something heavy. Their body language suggests exertion and strain.
	●	7.0	Frontstroke swimming	Belly flopping	The person in the video is lying on their stomach on the sand, their legs extended behind them, and their arms outstretched in front. Their body is relatively flat against the ground. One arm is slightly raised, possibly suggesting a wave-like motion. The overall posture and movement suggest a person mimicking the action of someone swimming, or more specifically, diving or belly-flopping into the water.
	💡	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.
Base ≠back.	○	14.0	Drinking out of a cup	Looking through binoculars	They hold their hands up to their eyes in a characteristic manner, mimicking the way one would hold and adjust binoculars to look through them. Their hand movements are consistent with focusing and looking through an imaginary device.
	🕒	38.4	Throwing dart	Answering phone call	They bring their hand to their ear as if holding a phone, and then extend their arm and point as if speaking to someone on the other end of the call.
	●	10.5	Deadlifting	Picking up a spill	The person in the video is bending over, appearing to pick up something from the floor. Their hands move as if they are gathering or collecting a substance. Their body is positioned as if they are carefully handling whatever is on the ground.
	😬	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.

Table 6: Gemini 1.5 Flash with CoT’s prediction category distributions (\*Note % indicates proportion of predictions, including correct ones) and corresponding examples, as discussed in §5.3. ○: completely incorrect description. 🕒: partially correct description of action. ●: correct description with only minor discrepancies but incorrect interpretation. ✓: correct description and valid alternative interpretation. Correct descriptions are in blue and incorrect descriptions are in orange.