Can Vision Language Models Understand Mimed Actions?

Hyundong Cho[1] **Spencer Lin**[2] **Tejas Srinivasan**[3] **Michael Saxon**[4]
Deuksin Kwon[2] **Natali T. Chavez**[5] **Jonathan May**[1]
Information Sciences Institute[1], Institute for Creative Technologies[2], Department of Computer Science[3]
University of Southern California
University of California, Santa Barbara[4] Aristotle University of Thessaloniki[5]
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 also 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.

Figure 1: Simplified illustration of a sample in MIME shown with a single frame from a video of a 3D male character miming a basketball shot. Humans achieve almost perfect accuracy on identifying mimed actions regardless of evaluation format, adversarial perturbations, and the absence of any salient context (e.g., basketball, court, basketball uniform). On the other hand, the best performing VLM achieves only 52.3% with a multiple choice format with four choices and 19.8% with free-form short answers even without any perturbations.

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), but also makes interaction engaging and natural (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; 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). Despite impressive achievements of VLMs on action recognition benchmarks (Qu et al., 2024; Kong and Fu, 2022;

Findings of the Association for Computational Linguistics: ACL 2025, pages 26744–26759

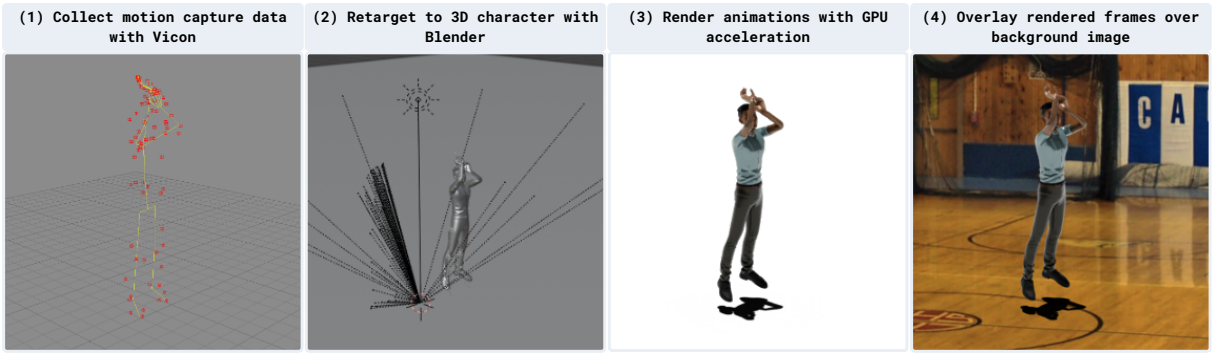

Figure 2: An overview of the pipeline for constructing MIME. (1) We first collect motion capture data of a mimed action on a Vicon stage. (2) Then, a 3D character is retargeted to our motion capture data in Blender, a computer graphics software. (3) Next, we render frames of the animation with a transparent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

Wang et al., 2023), we find that they cannot even reliably identify a subset of NVC that human adults without apraxia[1] 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; Yi et al., 2023). 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 **M**ime **I**dentification **M**ultimodal **E**valuation (MIME),[2] 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 easily 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 despite the contextual information provided by the answer choices and at most 19.8% with a free-from short answers format. These scores are even lower for videos with adversarial perturbations, for which all evaluated models achieving 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), suggesting that these models lack a robust understanding of the mimed actions in MIME.

To answer the second question, we conduct a preliminary exploration into whether existing methods can help 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. 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

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 evaluation 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 for conducting a systematic study

[1] A neurological disorder that disrupts the ability to plan and execute purposeful movements, despite having the physical ability to do so.

[2] Data and code for MIME is available https://justin-cho.com/mime.

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. In this section, we describe the data collection pipeline for MIME. An overview is shown in Figure 2.

2.1 Motion Capture

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. We exclude gestures such as hand-waving or thumbs-up as salient context is not missing in their enactment.

Next, we have two actors (one nonprofessional actor and one 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 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 shotputting, the actors reference YouTube videos of professional athletes.

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.[3] 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.

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.

2.2 Blender File Creation

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.1.

2.3 Rendering

Characters We use free 3D characters from Mixamo.[4] 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[5] as aligned and misaligned backgrounds (e.g., (d,i) and (e,j) in Figure 3, respectively). We try 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.[6]

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.

[3] https://help.vicon.com/space/Shogun112/31229851/Place+markers+on+a+performer

[4] https://www.mixamo.com

[5] https://commons.wikimedia.org/

[6] 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.

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. 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 (≠background, e.g., living room).

3 Experimental Setup

3.1 Evaluation

MIME evaluates VLMs with two different question answering conditions, the choice condition and naming condition (Osiurak et al., 2012). 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. 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).[7] While this may make the multiple choice setup easier, it simplifies evaluation by preventing instances where multiple answer choices are valid.

[7] sentence-transformers/all-MiniLM-L6-v2

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 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 above prior to selecting distractors, 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 affect 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

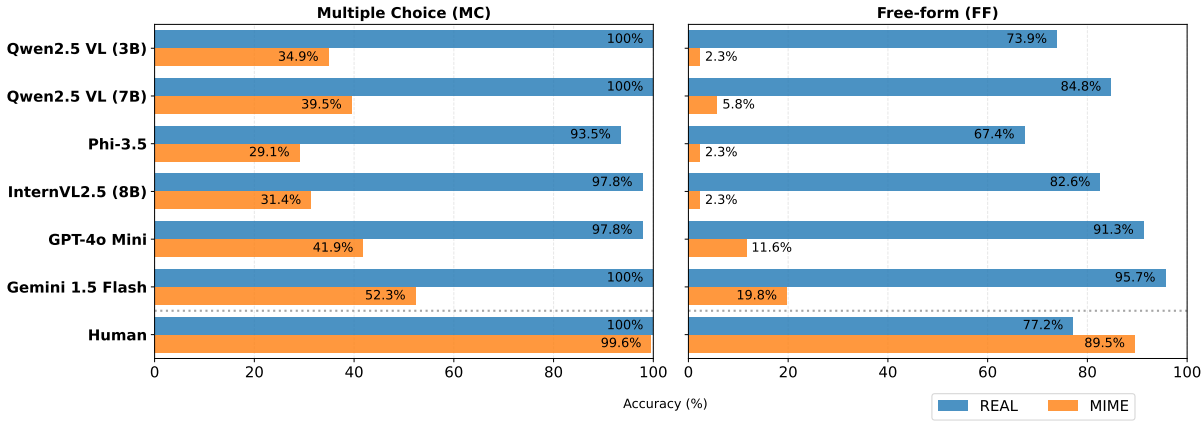

Figure 4: Performance comparison on the base setting of MIME and on the REAL dataset. Humans show equally strong performance on both MIME and REAL. VLMs struggle with MIME while achieving comparative performance on REAL, which suggests they lack a robust understanding of human actions.

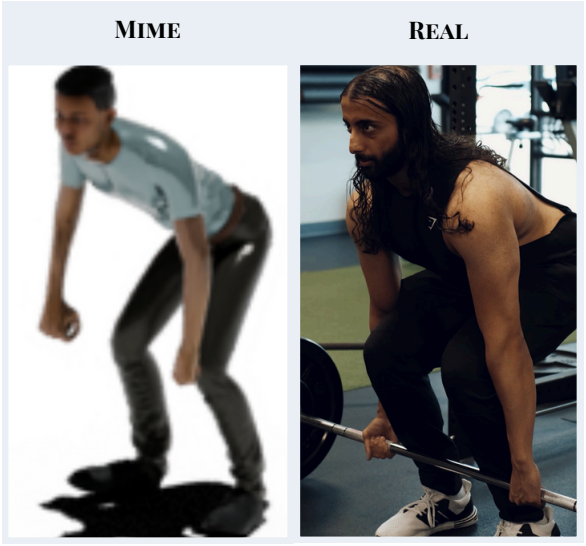

Figure 5: A frame from videos of deadlifting from MIME (left) and REAL (right). In MIME, salient objects are missing (e.g., barbell) and other hints (e.g., gym clothing).

mimed activities.

For open-source models, we evaluate on (*i*) Qwen 2.5 VL Instruction (Team, 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-soure models, we evaluate on (*iv*) Gemini 1.5 Flash from Google (Team, 2024) and (*v*) GPT-4o mini from OpenAI.[8] 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 REAL Data

We ground the performance on MIME by measuring the performance on recognizing actions from real footage of the same set of actions in MIME. We collect a set of license-free videos of such footage and call it REAL. An example of a video from REAL and its corresponding sample in MIME is shown side by side in Figure 5. 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.

Videos for REAL are sourced through Pexels.[9] 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 using a zero-shot prompt are shown in Figure 4.

Results on REAL clearly indicate that all VLMs are able to identify actions when all of the salient

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

[9] https://www.pexels.com/

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	68.6	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	**73.3**	27.9	31.4	8.1	**44.2**	0.0	72.1	27.9	**36.1**	**5.8**
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	41.9	11.6	66.3	39.5	**37.2**	3.5	33.7	**8.1**	67.4	33.7	**36.1**	2.3
Gemini 1.5 Flash	**52.3**	**19.8**	68.6	**51.2**	**37.2**	**12.8**	**44.2**	**8.1**	**75.6**	**46.5**	**36.1**	3.5
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. =back. indicates aligned background (e.g., basketball shot on a basketball court), ≠back. indicates misaligned background (e.g., basketball shot in living room, see Figure 1). 😈 denotes using an adversarial character. Samples of each variant are shown in Figure 3. Humans are robust to all variations, but VLMs drop performance for adversarial perturbations while get a significant boost when exposed to signals from the background that are aligned with the action.

context is present (e.g., doing a deadlift in a gym with a barbell while wearing gym attire), achieving almost perfect scores for all models for the multiple choice format while showing only a minor drop for the free-form format. This is on par with human performance.

However, on MIME, the performance drops sharply, while human performance remains more or less the same, with only a 0.4% drop in multiple choice while there is actually a boost for free-form by 12.3%. Upon manual inspection, we find that this is not because human performance actually is worse with real footage, but rather because humans are more descriptive in their responses for the free-form format for the real footage and this produces more false negatives.

4.2 Character and Background Perturbations

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 multiple choice settings while scoring at least 89.5% in the free-form setting. These results indicate that while humans are able to ignore irrelevant information and only focus on the actions themselves, VLMs rely on other hints about the action present in the scene. These results are in line with the resuls on REAL.

4.3 Angle Variations

Next, we share results with various angle perturbations to observe whether VLMs are viewpoint-agnostic for identifying mime. We see that humans clearly are 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 the multiple choice format, the variance in accuracy is much larger for VLMs than humans, another indication of a lack of robustness in VLMs in comparison to humans.

4.4 Gender Variation

Lastly, although our dataset has been verified as 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 posi-

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.

Model	Method	MC		FF	
		Male	Female	Male	Female
Qwen 2.5 VL (3B)	Zero-shot	34.9	29.1	2.3	1.2
	CoT	29.1	37.2	0.0	2.3
Qwen 2.5 VL (7B)	Zero-shot	39.5	41.9	5.8	9.3
	CoT	41.9	46.5	8.1	10.5
Phi 3.5	Zero-shot	29.1	34.9	2.3	2.3
	CoT	41.9	33.7	4.7	2.3
InternVL2.5 (8B)	Zero-shot	31.4	33.7	2.3	5.8
	CoT	25.6	24.4	1.2	5.8
GPT-4o-mini	Zero-shot	41.9	44.2	11.6	12.8
	CoT	43.0	53.5	16.3	10.5
	Few-shot	74.4	65.1	9.3	10.5
Gemini 1.5 Flash	Zero-shot	52.3	47.7	19.8	20.9
	CoT	54.7	52.3	22.1	19.8
	Few-shot	57.0	59.3	14.0	22.1
Human	-	99.6	98.5	89.5	90.3

Table 3: Performance comparison on MIME for gender variations. Similar to angle variation results, results for VLMs vary largely depending on the gender, while human performance is consistent. However, there is no consistent performance advantage for a certain gender.

tive 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 in MIME, we are interested in whether there are simple methods that can surface VLMs' potential to understand mimed actions. Therefore, we attempt to improve their performance via various well-established methods that do not require excessive compute.

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 randomly sample three other samples from MIME and provide them as in-context examples that the models can leverage to improve their predictions on the target sample. (*iv*) **Fine-tuning**: Lastly, we experiment with fine-tuning. Since we have a limited data size and compute budget, we fine-tune (FT) our model using a 5-fold validation approach with a 41/18/41 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). 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× A100 GPUs. Refer to further details for our fine-tuning setup in Appendix C.

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 the multiple choice format. 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 the free-form format, indicating that they continue to struggle without contextual information. 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.

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 fine-tuning results.

6 Related Work

6.1 Nonverbal Communication

Beginning with prior foundational work done in NVC recognition, Gao et al. (2017) proposed a model that jointly predicts action proposals and Cao et al. (2018) presented a real-time multi-person pose detection system. Carreira and Zisserman (2017) developed convolutional neural network architectures and advanced action recognition research. Such work paved the way for even more sophisticated methods such as a sports action recognition system that can analyze video input using a particle swarm optimization algorithm (Zhang and Hou, 2023). Although these studies have all made substantial contributions, they are done with non-VLM approaches.

6.2 Action Recognition

Currently, there exists several video datasets focusing on people engaging with objects and performing actions in everyday contexts (Tenorth et al., 2009; Goyal et al., 2017). Furthermore, there is also RareAct, a video dataset with rare and complex actions including some pantomimic gestures (Miech et al., 2020). However, these datasets have several shortcomings in that they are not focused on pantomimic actions and are not modifiable so that the same actions can be performed in a swappable adversarial background. Our motion capture dataset, on the other hand, allows for this flexibility.

Previous work testing VLM understanding of human actions has been limited to scenarios where a salient object, e.g. a sporting apparatus, is present in the input data (Kong and Fu, 2022; Sun et al., 2022). As such, it is dubious whether the VLM is truly understanding complex human body motions or if it is simply identifying the salient object. Overall, this presents ample motivation for further exploration into how AI understands the nuances of complex body motions through testing its interpretation of pantomimic actions.

7 Conclusion

In this work, we introduced MIME as a novel benchmark for assessing VLMs' understanding of NVC through the controlled study of mimed gestures. By constructing a dataset of 86 mimed actions with systematic perturbations in character, background, and viewpoint, we evaluated the robustness of both open-weight and API-based VLMs in recognizing nonverbal cues. Our findings reveal a significant gap between human and model performance, highlighting the limitations of current VLMs in understanding gesture-based communication. While humans remain highly robust to adversarial modifications, models struggle, particularly in free-form settings where recognition accuracy approaches zero under adversarial perturbations. These results

highlight the need for further research in enhancing VLMs' capacity to generalize across variations in nonverbal expressions and motivate future efforts to integrate context-aware, multimodal reasoning into vision-language models to bridge the gap in NVC fluency between human and AI.

Limitations

Despite the unique value of MIME that we presented in this work, there are several limitations worth mentioning.

First, the dataset used in this study is not photorealistic. This introduces potential domain gaps compared to real-world applications, where higher-quality visual data and more naturalistic human performances may contribute to better model understanding. Additionally, the dataset may exhibit distribution shifts that could artificially increase the difficulty of the task for machine learning models. However, given that humans can successfully interpret these mimed actions, we argue that models should also be capable of generalizing if they develop a robust understanding of nonverbal gestures and movements. This highlights the importance of evaluating model robustness under varying data distributions, particularly when dealing with abstract or indirect representations of actions.

Another limitation 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 attempts were conducted with a limited sample size, which likely led 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 MIME-related tasks.

Lastly, it is important to consider that human performance on MIME-based tasks is inherently subjective and may vary across individuals due to differences in prior exposure, cultural context, and interpretative abilities. While human-level performance provides a useful benchmark, it does not fully account for potential ambiguities in mimed actions, which could influence both model and human understanding. Future work should consider incorporating diverse human evaluations, as well as multi-modal learning approaches that leverage audio, text, and additional context cues to enhance model comprehension of mimed actions.

Ethical Considerations

Though VLMs present an exciting trove of possibilities, prudent ethical considerations should be made in their use. No AI is infallible and caution must always be taken with regards to accuracy and reliability. In the same vein, VLMs are not fully explainable, making it difficult to trace back to a point of failure if there is a lapse in the model's judgment. As with the handling of any sensitive data, especially video, special care should be taken to ensure the privacy, security and protection of data. Finally, it should be acknowledged that there exists inherit bias in VLMs based on their training data, so as to not disproportionately harm certain groups or reinforce harmful stereotypes.

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 Santacroce, 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).

Michael Argyle and Peter Trower. 1979. Person to person: ways of communicating. *(No Title)*.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

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.

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.

J. Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ramakant Nevatia. 2017. Turn tap: Temporal unit regression network for temporal action proposals. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3648–3656.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. *IEEE International Conference on Computer Vision*.

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.

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.

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.

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.

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.

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.

Chanjun Park, Yoonna Jang, Seolhwa Lee, Jaehyung Seo, Kisu Yang, and Heuiseok 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. Llms are good action recognizers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18395–18406.

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.

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.

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, Qiuhong 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.

Ole Tange. 2024. Gnu parallel 20240522 ('tbilisi').

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Qwen Team. 2025. Qwen2.5-vl.

Moritz Tenorth, Jan Bandouch, and Michael Beetz. 2009. The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1089–1096.

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.

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. Hunyuanvideo: A systematic framework for large video generative models.

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.

Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. 2023. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12965–12976.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Youming Zhang and Xingchen Hou. 2023. Application of video image processing in sports action recognition based on particle swarm optimization algorithm. *Preventive medicine*, page 107592.

Appendix

A MIME Details

A.1 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,[10] and using the Rokoko Studio Live plugin[11] 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.2 Render settings

Each frame is rendered with the following rendering configurations:

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

[10] https://github.com/absolute-quantum/cats-blender-plugin

[11] 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 doing 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 doing 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 doing 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 doing 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 doing 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).

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

D Human Evaluation Details

The interface for our human evaluation is shown in Appendix D. In total, we had 56 unique internal participants participate in our human evaluation study. Our participants cover a wide demographic: there are eight unique nationalities and their ages range from early 20s to mid 40s. They are all educated at the college level or beyond. While they are all located in the same city, we believe their diverse international backgrounds provide a reasonable approximation of human performance for our human evaluation setup.

E MIME via Video Generation Models

We also explore alternative video generation models for creating MIME and show sample outputs in Figure 7. For paid services, we test Sora[12] and Runway[13], and for open-weight models, we use a variety of Hunyuan (Sun et al., 2024) fp16 and bf8 models using ComfyUI's[14] recommended text-to-video Hunyuan workflow (Weijie Kong, 2024). All video models struggle to generate mimed actions and generate the action with the key object still present in the video, even when explicitly asked not to include it (see Figure 7c)or not mentioning it in the prompt (see Figure 7a. We also try with 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 objects in some cases, it fails to produce a video that matches to the intended action (e.g., dancing move shown for a prompt for fencing Figure 7b).

[12] https://openai.com/sora/
[13] https://runwayml.com/
[14] https://github.com/comfyanonymous/ComfyUI

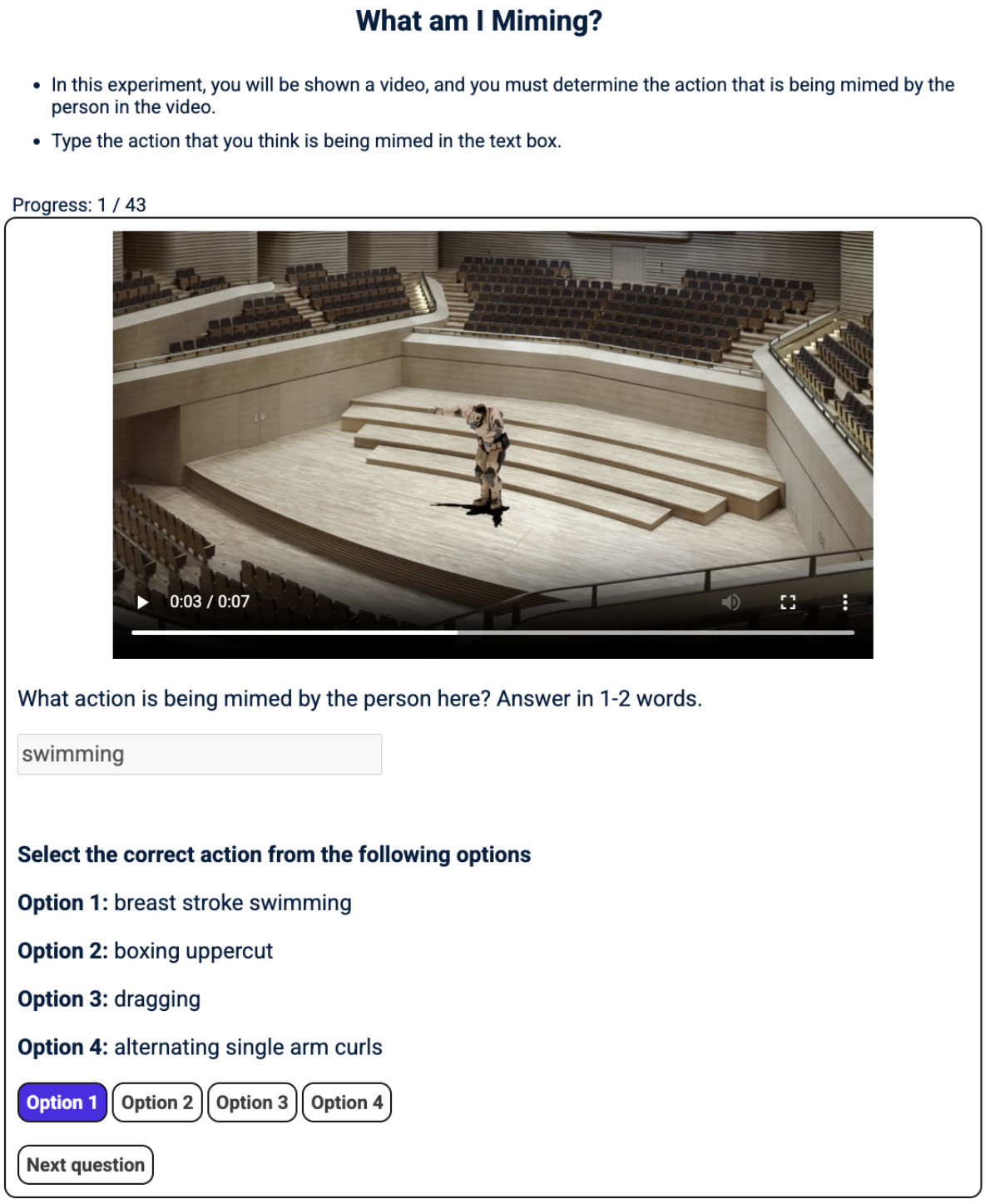

Figure 6: Our interface for human evaluation. The evaluators can only attempt to answer the question after seeing the full video. After answering a free-form short answer question, they are asked to complete a multiple choice equivalent before moving on to the next sample.

(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.