

ExAct: A Video-Language Benchmark for Expert Action Analysis

Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin Zhang,
Oluwatumininu Oguntola, Gedas Bertasius

UNC Chapel Hill





Project Page

https://texaser.github.io/exact_project_page/

Motivation

- Modern VLMs lack expert-level understanding of skilled physical human activities.

Sports	Bike Repair	Music
		
<p>? Question: Which expert commentary best matches the provided video?</p>		
<ol style="list-style-type: none"> 1 The participant should keep his eyes on the ball ... to improve the angle at the backboard for the layup. ❌ 2 The participant should focus on spinning the ball harder off the backboard ... for the layup. ❌ 3 The participant should aim for a higher jump ... to achieve a better angle at the backboard. ✔️ 4 The participant should aim for a lower jump to maintain control ... for a more effective reverse layup. ❌ 5 The participant should take off from both feet to get more hang-time ... for the reverse layup. ❌ 	<ol style="list-style-type: none"> 1 The participant shows a useful technique by pedaling forwards to stop the wheel. ❌ 2 The participant shows a useful technique by holding the brake lever to stop the wheel. ❌ 3 The participant shows a useful technique by raising the bike on a stand to keep the wheel stationary. ❌ 4 The participant shows a useful technique by adjusting the derailleur to stop the wheel. ❌ 5 The participant shows a useful technique by pedaling backwards to stop the wheel from moving. ✔️ 	<ol style="list-style-type: none"> 1 The participant should concentrate on moving the left hand which will adjust the right hand's position. ❌ 2 The participant should adjust the right hand higher to avoid moving up and down for different keys. ✔️ 3 The participant should keep fingers close to the keyboard to access white and black keys with less movement. ❌ 4 The participant should place the right hand lower on the keyboard to reach keys without unnecessary motion. ❌ 5 The participant should position the right hand lower to reduce wrist strain and improve reach. ❌

Task: Given a short video, select the correct expert commentary from 5 options.

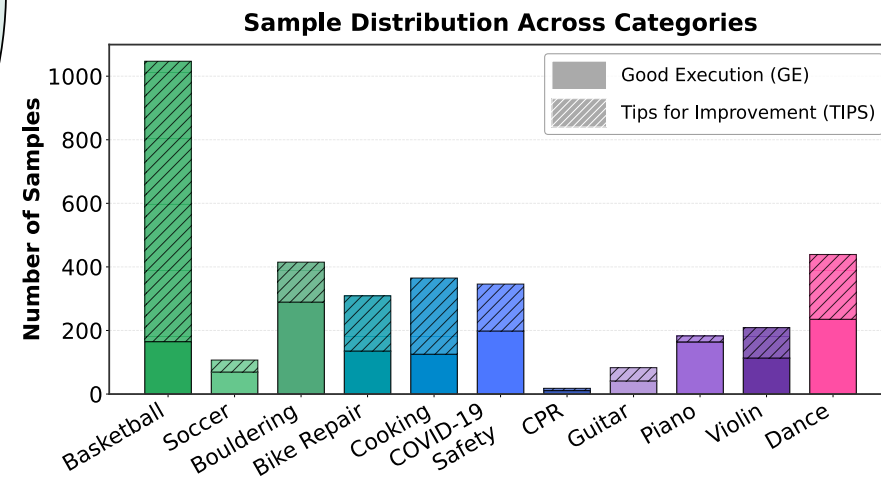
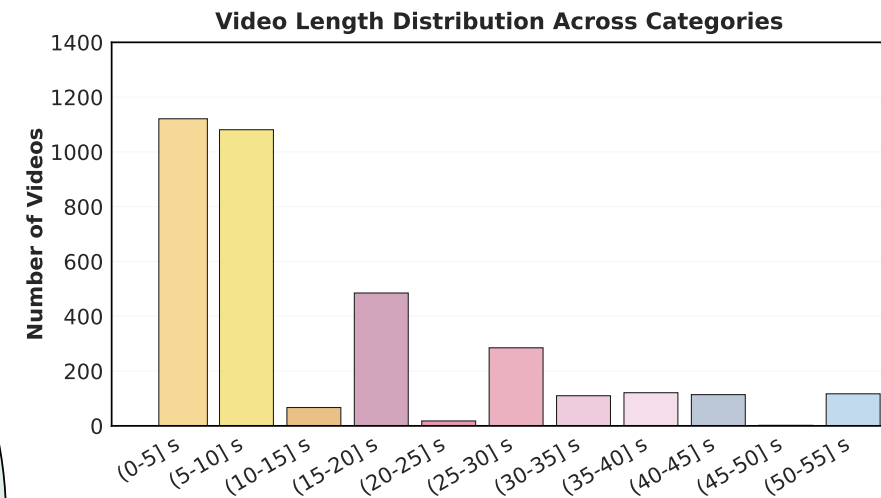
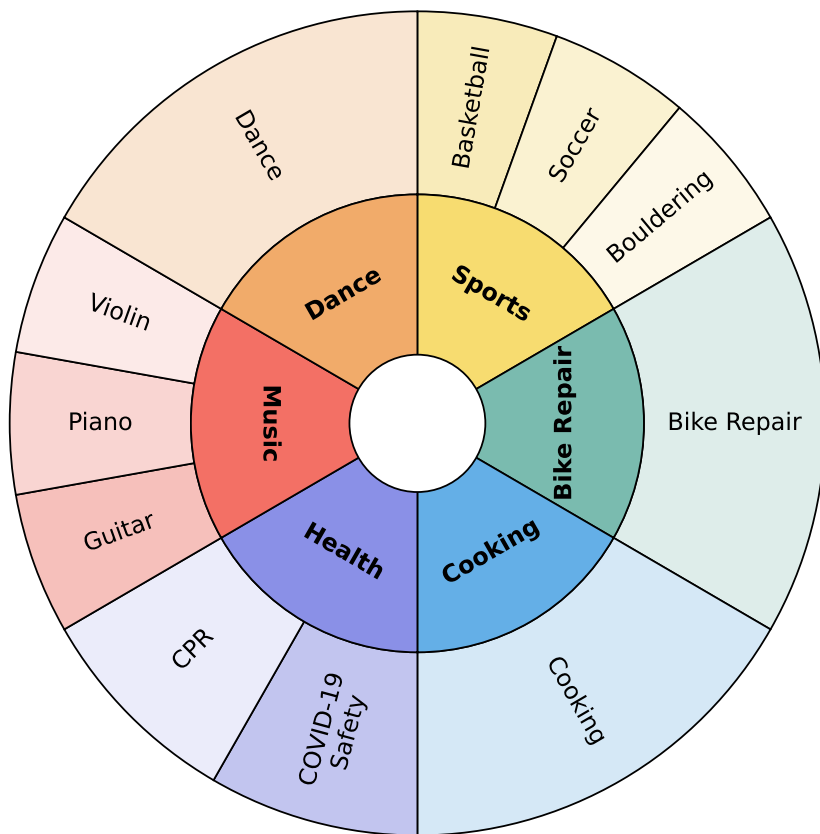
Comparison with Existing Datasets

- Most existing datasets lack expert-level annotations
- Provides free-form language annotations
- Supports MCQ-based evaluation

Dataset	Expert-level Knowledge	Free-form Language Annotations	MCQ Evaluation
<i>Coarse Action Recognition Datasets</i>			
Kinetics-700 [8]	✗	✗	✗
HowTo100M [33]	✗	✗	✗
UCF101 [45]	✗	✗	✗
HMDB [21]	✗	✗	✗
Moments in Time [34]	✗	✗	✗
Hollywood [43]	✗	✓	✗
ActivityNet-QA [51]	✗	✗	✓
<i>Fine-grained Action Recognition Datasets</i>			
Something-SomethingV2 [15]	✗	✗	✗
FineGym [42]	✗	✗	✗
Multisports [25]	✗	✗	✗
TemporalBench [6]	✗	✓	✓
<i>Video-Based Skill Assessment Datasets</i>			
JIGSAWS [1]	✓	✗	✗
Best [12]	✓	✗	✗
FineDiving [50]	✓	✗	✗
FP-Basket [4]	✓	✗	✗
BASKET [36]	✓	✗	✗
Aifit [13]	✓	✓	✗
<i>Skilled Activity Video-Language Datasets</i>			
VidDiffBench [5]	✓	✓	✗
EgoExo-Fitness [26]	✓	✓	✗
EgoExoLearn [19]	✓	✓	✗
Ego-Exo4D [16]	✓	✓	✗
EXACT (Ours)	✓	✓	✓

Benchmark Overview

- 3,521 expert-curated video QA samples
- 11 activities across 6 domains: Sports, Music, Dance, Health, Cooking, Bike Repair



Benchmark Construction

Stage I: Pre-Processing Raw Expert Commentaries

Raw expert commentary

An additional thing to look for in terms of footwork in order to be able to open up his hips, so as the ball's been played in, again we're looking to move into this space to set ourselves up to release the ball. So by moving the non-kicking foot back slightly as the ball comes in, that should naturally help in terms of opening up his hips to face more in the direction that we want to go and then it opens up the space for the ball to travel through.



GPT-4o



The participant should adjust his footwork by moving the non-kicking foot back slightly as the ball comes in. This will help in opening up his hips to face more in the direction they intend to go, and it will also create space for the ball to travel through.

Self-contained feedback

Benchmark Construction

Stage I: Pre-Processing Raw Expert Commentaries

Raw expert commentary

An additional thing to look for in terms of footwork in order to be able to open up his hips, so as the ball's been played in, again we're looking to move into this space to set ourselves up to release the ball. So by moving the non-kicking foot back slightly as the ball comes in, that should naturally help in terms of opening up his hips to face more in the direction that we want to go and then it opens up the space for the ball to travel through.



GPT-4o

The participant should adjust his footwork by moving the non-kicking foot back slightly as the ball comes in. This will help in opening up his hips to face more in the direction they intend to go, and it will also create space for the ball to travel through.

Self-contained feedback

Stage II: Question and Answer Generation



? Which expert commentary best matches the provided video?

- 1 The participant should **move his non-kicking foot forward**... This helps square the hips and reduce interception. ✗
- 2 The participant should **cross his non-kicking foot over**... This helps generate more power... **for a longer pass**. ✗
- 3 The participant should adjust his footwork by **moving the non-kicking foot back**... This will help in opening up his hips... and **create space**... ✓
- 4 The participant should **rotate his torso away** from the target... This helps add **swerve to the ball and mislead the opponent**. ✗
- 5 The participant should **plant his non-kicking foot in line with the ball**... This helps **maintain hip stability and a straight pass path**. ✗

Benchmark Construction

Stage I: Pre-Processing Raw Expert Commentaries

Raw expert commentary

An additional thing to look for in terms of footwork in order to be able to open up his hips, so as the ball's been played in, again we're looking to move into this space to set ourselves up to release the ball. So by moving the non-kicking foot back slightly as the ball comes in, that should naturally help in terms of opening up his hips to face more in the direction that we want to go and then it opens up the space for the ball to travel through.



GPT-4o

The participant should adjust his footwork by moving the non-kicking foot back slightly as the ball comes in. This will help in opening up his hips to face more in the direction they intend to go, and it will also create space for the ball to travel through.

Self-contained feedback

Stage II: Question and Answer Generation



? Which expert commentary best matches the provided video?

- 1 The participant should **move his non-kicking foot forward**... This helps square the hips and reduce interception. ✗
- 2 The participant should **cross his non-kicking foot over**... This helps generate more power... **for a longer pass**. ✗
- 3 The participant should adjust his footwork by **moving the non-kicking foot back**... This will help in opening up his hips... and **create space**... ✓
- 4 The participant should **rotate his torso away** from the target... This helps add **swerve to the ball and mislead the opponent**. ✗
- 5 The participant should **plant his non-kicking foot in line with the ball**... This helps **maintain hip stability and a straight pass path**. ✗

Stage III: Generated Question Answer Filtering

1. Length Similarity Filtering



2. Blind-LLM filtering



Benchmark Construction

Stage I: Pre-Processing Raw Expert Commentaries

Raw expert commentary

An additional thing to look for in terms of footwork in order to be able to open up his hips, so as the ball's been played in, again we're looking to move into this space to set ourselves up to release the ball. So by moving the non-kicking foot back slightly as the ball comes in, that should naturally help in terms of opening up his hips to face more in the direction that we want to go and then it opens up the space for the ball to travel through.



GPT-4o

The participant should adjust his footwork by moving the non-kicking foot back slightly as the ball comes in. This will help in opening up his hips to face more in the direction they intend to go, and it will also create space for the ball to travel through.

Self-contained feedback

Stage II: Question and Answer Generation



? Which expert commentary best matches the provided video?

- 1 The participant should **move his non-kicking foot forward**... This helps square the hips and reduce interception. ✗
- 2 The participant should **cross his non-kicking foot over**... This helps generate more power... **for a longer pass**. ✗
- 3 The participant should adjust his footwork by **moving the non-kicking foot back**... This will help in opening up his hips... and **create space**... ✓
- 4 The participant should **rotate his torso away** from the target... This helps add **swerve to the ball and mislead the opponent**. ✗
- 5 The participant should **plant his non-kicking foot in line with the ball**... This helps **maintain hip stability and a straight pass path**. ✗

Stage III: Generated Question Answer Filtering

1. Length Similarity Filtering



2. Blind-LLM filtering



deepseek



Qwen



书生



GPT-4o



Llama

Stage IV: Final Expert Review and Validation



Data Examples



?

Question: Which expert commentary best matches the provided video?

A

The participant needs to improve on providing enough arc accuracy and rotation on their jump shot to ensure the ball reaches the midpoint between the rims.

B

The participant should work on keeping a stiffer wrist during the release to maintain stability, which will ensure the ball travels precisely to the center of the hoop.

C

The participant should aim to add more spin to the ball to create a backspin effect, which will assist the ball in reaching the center point of the hoop.

D

The participant should concentrate on jumping higher to increase the shot's velocity, which will make the ball accurately land in the midpoint between the rims.

E

The participant needs to focus on reducing the arc of their jump shot to increase momentum, which will help the ball reach the midpoint between the rims.

Data Examples



?

Question: Which expert commentary best matches the provided video?

A

The participant needs **to improve on providing enough arc accuracy and rotation on their jump shot** to ensure the ball reaches the midpoint between the rims.

B

The participant should work on **keeping a stiffer wrist during the release to maintain stability**, which will ensure the ball travels precisely to the center of the hoop.

C

The participant should aim to **add more spin to the ball to create a backspin effect**, which will assist the ball in reaching the center point of the hoop.

D

The participant should concentrate on **jumping higher to increase the shot's velocity**, which will make the ball accurately land in the midpoint between the rims.

E

The participant needs to focus on **reducing the arc of their jump shot to increase momentum**, which will help the ball reach the midpoint between the rims.

Data Examples



?

Question: Which expert commentary best matches the provided video?

A

The participant needs **to improve on providing enough arc accuracy and rotation on their jump shot** to ensure the ball reaches the midpoint between the rims.

B

The participant should work on **keeping a stiffer wrist during the release to maintain stability**, which will ensure the ball travels precisely to the center of the hoop.

C

The participant should aim to **add more spin to the ball to create a backspin effect**, which will assist the ball in reaching the center point of the hoop.

D

The participant should concentrate on **jumping higher to increase the shot's velocity**, which will make the ball accurately land in the midpoint between the rims.

E

The participant needs to focus on **reducing the arc of their jump shot to increase momentum**, which will help the ball reach the midpoint between the rims.

Model response:

GPT-4o: Option A



Gemini 1.5 Pro: Option A



LLaVA-Video: Option A



Human:

Human Expert: Option A



Human Non-Expert: Option A



Data Examples



?

Question: Which expert commentary best matches the provided video?

A

The participant executes a nice jump to the side with well-bent knees while clapping her hands above her head to maintain rhythm and movement.

B

The participant executes a nice jump to the side with well-bent knees and performs a nice roll with her upper body to maintain rhythm and movement.

C

The participant executes a nice jump to the side with well-bent knees and performs a smooth cartwheel to maintain rhythm and movement.

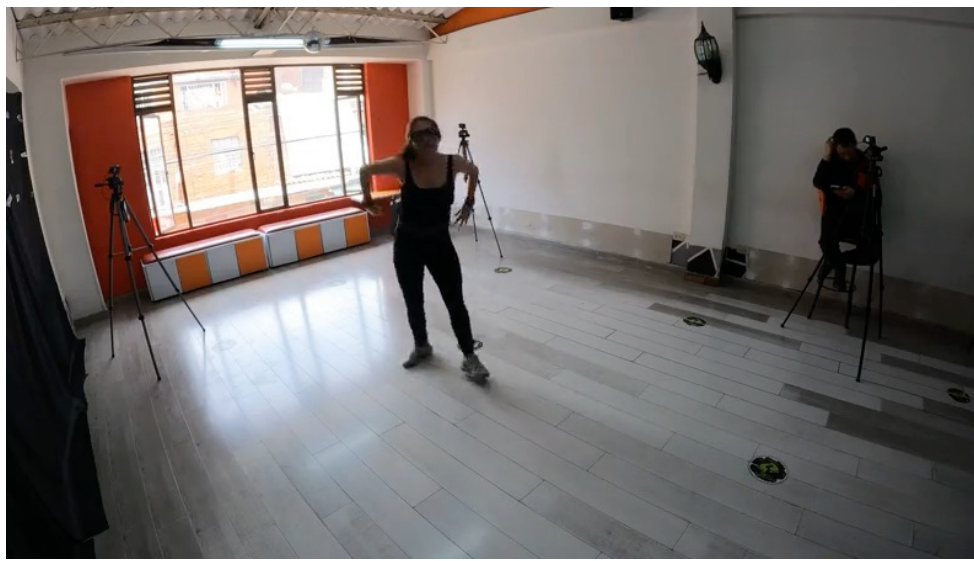
D

The participant executes a nice jump upwards with locked knees and performs a rigid turn with her upper body to maintain rhythm and movement.

E

The participant executes a nice leap to the front with well-straightened legs and performs a graceful arm sweep to maintain rhythm and movement.

Data Examples



?

Question: Which expert commentary best matches the provided video?

A

The participant executes a nice jump **to the side with well-bent knees while clapping her hands above her head** to maintain rhythm and movement.

B

The participant executes a nice jump **to the side with well-bent knees and performs a nice roll with her upper body** to maintain rhythm and movement.

C

The participant executes a nice jump **to the side with well-bent knees and performs a smooth cartwheel** to maintain rhythm and movement.

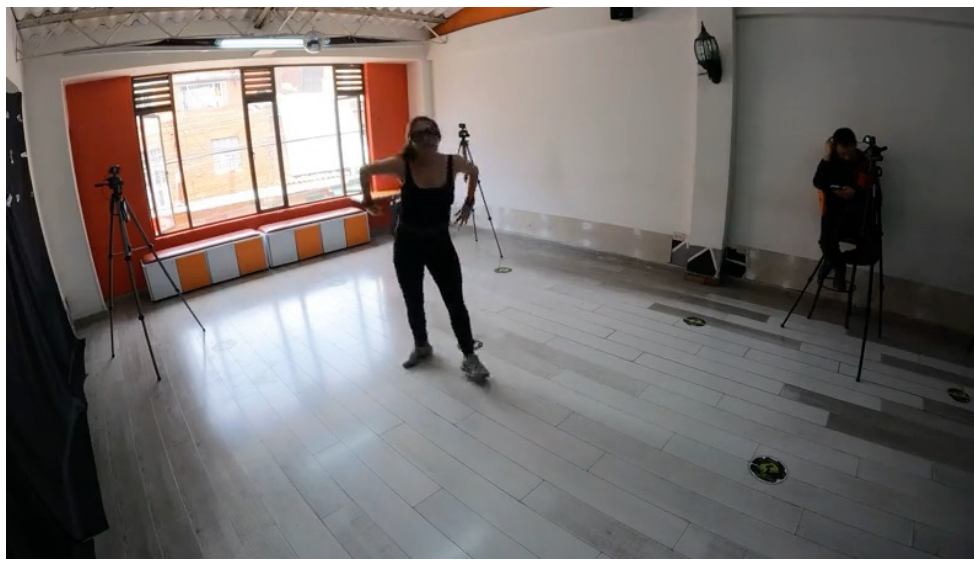
D

The participant executes a nice jump **upwards with locked knees and performs a rigid turn with her upper body** to maintain rhythm and movement.

E

The participant executes a nice leap **to the front with well-straightened legs and performs a graceful arm sweep** to maintain rhythm and movement.

Data Examples



- ?** **Question:** Which expert commentary best matches the provided video?
- A** The participant executes a nice jump **to the side with well-bent knees while clapping her hands above her head** to maintain rhythm and movement.
 - B** The participant executes a nice jump **to the side with well-bent knees and performs a nice roll with her upper body** to maintain rhythm and movement.
 - C** The participant executes a nice jump **to the side with well-bent knees and performs a smooth cartwheel** to maintain rhythm and movement.
 - D** The participant executes a nice jump **upwards with locked knees and performs a rigid turn with her upper body** to maintain rhythm and movement.
 - E** The participant executes a nice leap **to the front with well-straightened legs and performs a graceful arm sweep** to maintain rhythm and movement.

Model response:

GPT-4o: Option A

Gemini 1.5 Pro: Option A

LLaVA-Video: Option A

✗

✗

✗

Human:

Human Expert: Option B

Human Non-Expert: Option C

✓

✗

Benchmark Results

Model	Overall (%)	Results by Domain (%)					
		Sports	Bike Repair	Cooking	Health	Music	Dance
Random Choice	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Human Non-Expert	61.86	62.97	55.02	66.58	71.43	54.11	59.22
Human Expert	82.02	82.09	81.23	80.27	87.09	80.21	81.55
Open-source VLMs							
PerceptionLM-8B [10]	24.65	24.22	28.16	25.75	22.53	22.95	26.42
VideoLLaMA3-7B [53]	26.38	26.64	23.30	29.32	26.65	23.79	27.79
InternVL2.5-78B [9]	33.48	31.93	36.57	33.70	37.91	32.00	34.62
LLaVA-OneVision-72B [22]	35.44	33.65	43.04	33.42	35.44	30.53	43.51
Qwen2.5-VL-72B-Instruct [3]	35.67	35.62	37.86	33.97	36.26	32.63	38.50
LLaVA-Video-72B [55]	41.58	41.81	42.72	44.11	32.42	38.74	48.52
Proprietary VLMs							
Gemini 1.5 Pro [46]	43.91	42.83	52.10	51.78	41.21	41.89	39.86
GPT-4o [20]	44.70	43.47	52.75	46.30	53.30	33.89	46.70
GPT-4.1 [20]	50.89	51.37	58.90	54.25	51.10	40.84	51.48
Gemini 2.5 Pro [11]	55.35	52.58	65.05	58.36	60.71	53.05	53.98

Benchmark Results

Model	Overall (%)	Results by Domain (%)					
		Sports	Bike Repair	Cooking	Health	Music	Dance
Random Choice	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Human Non-Expert	61.86	62.97	55.02	66.58	71.43	54.11	59.22
Human Expert	82.02	82.09	81.23	80.27	87.09	80.21	81.55
Open-source VLMs							
PerceptionLM-8B [10]	24.65	24.22	28.16	25.75	22.53	22.95	26.42
VideoLLaMA3-7B [53]	26.38	26.64	23.30	29.32	26.65	23.79	27.79
InternVL2.5-78B [9]	33.48	31.93	36.57	33.70	37.91	32.00	34.62
LLaVA-OneVision-72B [22]	35.44	33.65	43.04	33.42	35.44	30.53	43.51
Qwen2.5-VL-72B-Instruct [3]	35.67	35.62	37.86	33.97	36.26	32.63	38.50
LLaVA-Video-72B [55]	41.58	41.81	42.72	44.11	32.42	38.74	48.52
Proprietary VLMs							
Gemini 1.5 Pro [46]	43.91	42.83	52.10	51.78	41.21	41.89	39.86
GPT-4o [20]	44.70	43.47	52.75	46.30	53.30	33.89	46.70
GPT-4.1 [20]	50.89	51.37	58.90	54.25	51.10	40.84	51.48
Gemini 2.5 Pro [11]	55.35	52.58	65.05	58.36	60.71	53.05	53.98

Benchmark Results

Model	Overall (%)	Results by Domain (%)					
		Sports	Bike Repair	Cooking	Health	Music	Dance
Random Choice	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Human Non-Expert	61.86	62.97	55.02	66.58	71.43	54.11	59.22
Human Expert	82.02	82.09	81.23	80.27	87.09	80.21	81.55
Open-source VLMs							
PerceptionLM-8B [10]	24.65	24.22	28.16	25.75	22.53	22.95	26.42
VideoLLaMA3-7B [53]	26.38	26.64	23.30	29.32	26.65	23.79	27.79
InternVL2.5-78B [9]	33.48	31.93	36.57	33.70	37.91	32.00	34.62
LLaVA-OneVision-72B [22]	35.44	33.65	43.04	33.42	35.44	30.53	43.51
Qwen2.5-VL-72B-Instruct [3]	35.67	35.62	37.86	33.97	36.26	32.63	38.50
LLaVA-Video-72B [55]	41.58	41.81	42.72	44.11	32.42	38.74	48.52
Proprietary VLMs							
Gemini 1.5 Pro [46]	43.91	42.83	52.10	51.78	41.21	41.89	39.86
GPT-4o [20]	44.70	43.47	52.75	46.30	53.30	33.89	46.70
GPT-4.1 [20]	50.89	51.37	58.90	54.25	51.10	40.84	51.48
Gemini 2.5 Pro [11]	55.35	52.58	65.05	58.36	60.71	53.05	53.98

Benchmark Results

Model	Overall (%)	Results by Domain (%)					
		Sports	Bike Repair	Cooking	Health	Music	Dance
Random Choice	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Human Non-Expert	61.86	62.97	55.02	66.58	71.43	54.11	59.22
Human Expert	82.02	82.09	81.23	80.27	87.09	80.21	81.55
Open-source VLMs							
PerceptionLM-8B [10]	24.65	24.22	28.16	25.75	22.53	22.95	26.42
VideoLLaMA3-7B [53]	26.38	26.64	23.30	29.32	26.65	23.79	27.79
InternVL2.5-78B [9]	33.48	31.93	36.57	33.70	37.91	32.00	34.62
LLaVA-OneVision-72B [22]	35.44	33.65	43.04	33.42	35.44	30.53	43.51
Qwen2.5-VL-72B-Instruct [3]	35.67	35.62	37.86	33.97	36.26	32.63	38.50
LLaVA-Video-72B [55]	41.58	41.81	42.72	44.11	32.42	38.74	48.52
Proprietary VLMs							
Gemini 1.5 Pro [46]	43.91	42.83	52.10	51.78	41.21	41.89	39.86
GPT-4o [20]	44.70	43.47	52.75	46.30	53.30	33.89	46.70
GPT-4.1 [20]	50.89	51.37	58.90	54.25	51.10	40.84	51.48
Gemini 2.5 Pro [11]	55.35	52.58	65.05	58.36	60.71	53.05	53.98

Benchmark Results

Model	Overall (%)	Results by Domain (%)					
		Sports	Bike Repair	Cooking	Health	Music	Dance
Random Choice	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Human Non-Expert	61.86	62.97	55.02	66.58	71.43	54.11	59.22
Human Expert	82.02	82.09	81.23	80.27	87.09	80.21	81.55
Open-source VLMs							
PerceptionLM-8B [10]	24.65	24.22	28.16	25.75	22.53	22.95	26.42
VideoLLaMA3-7B [53]	26.38	26.64	23.30	29.32	26.65	23.79	27.79
InternVL2.5-78B [9]	33.48	31.93	36.57	33.70	37.91	32.00	34.62
LLaVA-OneVision-72B [22]	35.44	33.65	43.04	33.42	35.44	30.53	43.51
Qwen2.5-VL-72B-Instruct [3]	35.67	35.62	37.86	33.97	36.26	32.63	38.50
LLaVA-Video-72B [55]	41.58	41.81	42.72	44.11	32.42	38.74	48.52
Proprietary VLMs							
Gemini 1.5 Pro [46]	43.91	42.83	52.10	51.78	41.21	41.89	39.86
GPT-4o [20]	44.70	43.47	52.75	46.30	53.30	33.89	46.70
GPT-4.1 [20]	50.89	51.37	58.90	54.25	51.10	40.84	51.48
Gemini 2.5 Pro [11]	55.35	52.58	65.05	58.36	60.71	53.05	53.98

Conclusion

- We introduce ExAct, a new benchmark for evaluating expert-level understanding of skilled human activities via multiple-choice QA.
- Our results show a large performance gap between current VLMs and human experts.
- ExAct serves as a rigorous and necessary benchmark for advancing expert-level video-language models.

Thank you!



Project Page



Hugging Face