

EXACT: A Video-Language Benchmark for Expert Action Analysis

Supplementary Material

Our supplementary materials include the following sections: Section **S1**: Full Prompts for Constructing EXACT, Section **S2**: Annotation Interface Details, Section **S3**: Qualitative Results, Section **S4**: Evaluated Reasoning Dimensions in EXACT, Section **S5**: Sample statistics after each filtering stage.

S1 Full Prompts for Constructing EXACT

We provide the detailed prompts used for (1) pre-processing raw expert commentaries, (2) generating negative/distractor commentaries for EXACT, and (3) formatting inputs for VLMs.

Prompt for Pre-Processing Raw Expert Commentaries. We present the prompts used for pre-processing raw expert commentaries into Good Execution and Tips for Improvement in Figure 5 and Figure 7, respectively. Corresponding examples of Good Execution and Tips for Improvement are shown in Figure 6 and Figure 8.

You are given a raw, transcribed expert commentary from a dataset describing a participant performing an {activity}. These transcriptions may contain automatic speech recognition (ASR) errors, redundant phrases, off-topic remarks, or unstructured language. Your primary goal is to identify the good executions mentioned in the commentary and express them in a clean, concise, and coherent manner. Specifically:

1. Correct any transcription or grammatical errors.
2. Remove irrelevant, repetitive, or filler content.
3. Your goal is to determine the good executions mentioned in the commentary and write them in a coherent, concise manner.
4. If there are no good executions, respond with: "The expert mentions no good executions."

Figure 5: A prompt for pre-processing raw expert commentaries from the good execution (GE) category.

Sample 1

Input

A grand great rhythm and flow as he crosses over. This is a great picture of going off the correct foot, which is his right foot, raising that left knee. He’s got his eye and head up, total great body extension. He’s got in great spatial awareness of where he is in relation to the basket and the rim. He’s not too far under, he’s not too far out.

Output

The participant has a great rhythm and flow as he crosses over. He goes off of the correct foot, raises the correct knee, and has his head and eyes up. Overall, great body extension and spatial awareness under the rim.

Sample 2

Input

So here the dancer is jumping side to side, almost like a little short gallop. Really nice leg space. I definitely would recommend bending the knees a little bit more. And then just really like, yeah, I think if we bent the knees that would just kind of bring a better level and ease for him to jump side to side. Good job with the arms swaying back and forth on that part and then good eye connection to the audience facing front.

Output

The participant demonstrates nice leg space during the jumps. Additionally, the participant nicely sways their arms and makes good eye contact with the audience.

Figure 6: Samples of converting raw expert commentaries into concise good execution (GE) commentaries.

You are given a raw, transcribed expert commentary from a dataset describing a participant performing an {activity}. These transcriptions may contain automatic speech recognition (ASR) errors, redundant phrases, off-topic remarks, or unstructured language.

Your primary goal is to identify the tips for improvement mentioned in the commentary and express them in a clean, concise, and coherent manner. Specifically:

1. Correct any transcription or grammatical errors.
2. Remove irrelevant, repetitive, or filler content.
3. Your goal is to determine the tips for improvement mentioned in the commentary and write them in a coherent, concise manner.
4. If there are no tips for improvement, respond with: "The expert mentions no tips for improvement."

Figure 7: A prompt for pre-processing raw expert commentaries from the tips for improvement (TIPS) category.

Sample 1

Input

This is another example again of the person taking too heavy of a touch and then overextending his leg to be able to wrap around the ball. So actually you see now how he's contacting the ball. Instead of kind of catching it, he's actually having to almost like, his foot placement is much higher on the center of the ball here. He's actually coming like downward with the force in this direction when he's actually contacting it, which could actually potentially make the ball jump up when he hits it.

Output

The participant should try to take a lighter touch and focus on not overextending their leg to wrap around the ball. The participant should avoid coming downward when contacting the ball because it could make the ball jump up when they hit it.

Sample 2

Input

This is a really excellent shift that the player had. If they wanted to be as expressive, they could change the weight as they're going up to come into the fingerboard a little bit more so it comes across as a more expressive slide.

Output

The participant should change the weight as they go up to come into the fingerboard so it comes across as a more expressive slide.

Figure 8: Samples of converting raw expert commentaries into concise tips for improvement (TIPS) commentaries.

Prompt for Negative Commentary Generation. Here we present the prompts used for negative/distractor commentary generation. For Good Execution (GE) commentaries, we adopt two strategies, as shown in Figure 9. For Tips for Improvement commentaries, we use four strategies, as shown in Figure 10.

You are a bouldering expert tasked with creating **wrong but plausible commentary** to train a video understanding model. You will be given high-expertise bouldering commentary, and your task is to generate **four wrong comments** based on that expert commentary. These wrong comments should be grounded in visible actions from the video and appear reasonable but must either provide an incorrect justification or directly misinterpret the actions.

Requirements:

- All comments must be grounded in observable actions from the video and avoid references to non-visual elements.
- Match the length, detail, and complexity of the expert’s original comments without obvious stylistic differences.
- The difference between correct and incorrect comments should lie in the reasoning or specific actions mentioned.
- Do not generate comments that might sometimes be true; ensure the actions are *definitely* incorrect based on expert feedback.
- Avoid using negative adjective words such as “improper,” “bad,” “not good,” or “not perfect.”
- Ensure the incorrect comments appear plausible, limiting each to 1–2 subtle errors.
- Vary the type of error across the four generated comments. Keep all comments logical and coherent.

Some techniques for creating wrong comments:

- **Action replacement:** A key action in the original commentary is substituted with a plausible but incorrect alternative.
- **Absent-action insertion:** A new event or action is inserted that was never mentioned or shown.

Output Format:

Good execution: “The participant demonstrates a good initiation of upward movement, properly preparing their legs to generate momentum upwards.”

Wrong Comments:

- **Action replacement:** “The participant demonstrates a good initiation of downward movement, correctly preparing their arms to generate momentum downwards.”
- **Action replacement:** “The participant properly prepares their arms to generate momentum sideways, effectively aiding their lateral movement along the boulder.”
- **Absent-action insertion:** “The participant demonstrates a clever strategy by swinging their body to the side before leaping to the next hold, avoiding direct upward movement.”
- **Absent-action insertion:** “The participant initiates a powerful dyno by planting both feet on the foothold and lunging directly for the top of the wall, using agility over controlled movement.”

The high-expertise comment is as follows:

Figure 9: Prompt for negative/distractor commentary generation for good execution commentaries. We use a bouldering example for illustration.

You are a cooking expert tasked with creating **wrong but plausible commentary** to train a video understanding model. You will be given high-expertise cooking commentary, and your task is to generate **four wrong comments** based on that expert commentary. These wrong comments should be grounded in visible actions from the video and appear reasonable but must either provide an incorrect justification or directly misinterpret the actions.

Requirements:

- All comments must be grounded in observable actions from the video and avoid references to non-visual elements.
- Match the length, detail, and complexity of the expert's original comments without obvious stylistic differences.
- The difference between correct and incorrect comments should lie in the reasoning or specific actions mentioned.
- Do not generate comments that might sometimes be true; ensure the actions are *definitely* incorrect based on expert feedback.
- Avoid using negative adjective words such as "improper," "bad," "not good," or "not perfect."
- Ensure the incorrect comments appear plausible, limiting each to 1–2 subtle errors.
- Vary the type of error across the four generated comments. Keep all comments logical and coherent.

Some techniques for creating wrong comments:

- **Action misinterpretation:** Misinterpreting the mistake in an execution.
- **Incorrect technical reasoning:** Correctly identifying a flaw but providing an implausible or technically inaccurate explanation.
- **False cause–effect relationship:** Introducing a misleading causal link between the error and an unrelated factor.
- **Ineffective suggestion:** Proposing a correction to the execution that does not address the problem.

Output Format:

Tips for improvement: "The participant should add herbs, spices, and tea while waiting for the mixture to come to a simmer to improve efficiency and flavor infusion."

Wrong Comments:

- **Action misinterpretation:** "The participant should wait until the mixture has finished simmering before adding herbs, spices, and tea, as this prevents any flavors from being cooked out of the ingredients."
- **Incorrect technical reasoning:** "The participant should add herbs, spices, and tea while the mixture is boiling vigorously, as the intense heat heightens the flavor of these ingredients."
- **False cause–effect relationship:** "The participant should add herbs, spices, and tea just after the mixture stops simmering, as this allows the flavors to cool simultaneously with the dish for balanced taste."
- **Ineffective suggestion:** "The participant should blend the herbs, spices, and tea into a smooth paste before adding them to the mixture after it comes to a simmer, ensuring a more uniform flavor throughout."

The high-expertise comment is as follows:

Figure 10: Prompt for negative/distractor commentary generation from tips for improvement commentaries. We use a cooking example for illustration.

VLM Prompt for Processing EXACT. Here we provide the template for the input prompt to LLaVA-Video. The prompts used for other models are very similar, only with a different video separation token. The default image token is the video separation input for LLaVA-Video. The scenario prompt briefly introduces the activity being performed by the participant (*e.g.*, “The participant is practicing basketball.”). We also include time-related instructions, such as the total video duration and the timestamps of uniformly sampled frames in the prompt to guide the VLMs. The prompt presents five candidate answer options labeled **Option 1** to **Option 5**, including one correct answer and four distractors.

```
<DEFAULT_IMAGE_TOKEN>
The video is {video_time} seconds long, and {len(frames)} uniformly sampled frames occur at
{frame_time}.
{scenario_prompt}
Below are different feedback statements about the person’s performance in this video:
Option 1. Option 2. Option 3. Option 4. Option 5.
Based on what you observe in the video, which expert commentary best matches the provided video?
Just respond with the option number (1–5) and nothing else.
```

Figure 11: An input prompt to Vision-Language models (VLMs) for processing EXACT.

S2 Annotation Interface Details

In this section, we provide more details related to the annotation interface used for annotating EXACT. We develop a user-friendly web-based platform tailored for human annotators. A total of 16 experts participated in the annotation process, each with at least ten years of experience in their respective domains. The website is built using `github.io`, with `Formspree` used to collect submission data from annotators. All annotators are compensated at a rate of \$50 per hour. The interface begins with an introduction to the project, followed by detailed annotation guidelines. It supports saving progress, allowing annotators to complete their assigned tasks over multiple sessions rather than in a single sitting. Each expert is only assigned samples within their domain of expertise. Upon completing their assigned tasks, annotators submit their responses via the form. Figure 12 and Figure 13 shows our data annotation interface.

Guidelines for Annotators. Please follow the instructions below when annotating:

1. Carefully watch the video clip.
2. Read all five options and select the one you believe is correct.
3. Click the “Confirm Selection” button to submit your answer.
4. After submitting your selection, the ground-truth answer will be revealed. Please then evaluate the sample based on the following criteria:
 - (a) Does the video clearly support the action described in the ground-truth commentary?
 - (b) Do any of the other options also appear valid based on the video?
 - (c) Are there any language issues, such as grammatical errors or illogical phrasing, in any of the options?
5. Click “Continue” to move on to the next sample.

Cooking Technique Evaluation

Please review each cooking video and select the option you believe is most accurate

ⓘ Your progress is automatically saved as you go

Instructions:

1. Watch the video clip carefully
2. Read all options and select the one you believe is correct
3. Click the "Confirm Selection" button to submit your answer
4. After your selection, the ground truth will be revealed. Please assess the sample quality by verifying:
 - Whether the video clearly shows the action described in the ground truth
 - Whether any other options (besides the ground truth) are also supported by the video
 - If there are any language issues (grammatically incorrect or illogical options)
5. Click "Continue" to proceed to the next question

Question 1 of 239

Progress saved

Figure 12: User instructions for the annotation platform interface.

Question 1 of 239 Progress saved

Phase 2: Quality Assessment

0:00 / 0:54

This is a participant practicing cooking, here are some expert feedbacks (good execution). Please select the option you believe is correct:

☐ The participant correctly uses the cut side down technique, ensuring each slice is followed by a gentle brush to remove debris.

☐ The participant correctly uses the tip-to-tip slicing method when cutting.

☐ The participant correctly uses the claw grip technique while cutting.

☒ The participant correctly uses the cut side down technique when cutting.

☐ The participant correctly uses the cut side down technique, applying salt to the edges for enhanced grip.

Ground Truth:
The participant correctly uses the cut side down technique when cutting.

Correct! You selected the most appropriate feedback option.

⌚ Step 2: Please complete the quality assessment below

How would you rate the overall quality of this sample?

☐ **Good quality** - The video and the ground truth match well

☒ **Poor quality** - There are issues with this sample

Please identify any issues with this sample (if any):

☐ The video does NOT clearly contain the action described in the ground truth

☒ One or more of the other options (not ground truth) ARE also supported by the video

☐ There are language issues (grammatically incorrect, illogical, or other language problems)

☐ Other issues

Additional Comments (Optional)

Any comments about your selection

Submit Quality Assessment


Figure 13: Two-phase manual review interface on the annotation website.

S3 Qualitative Results

In this section, we present four QA examples (Figures 14–17) that range from easy to difficult.

Sample 1 (Figure 14) represents a relatively easy example. All models, as well as both human experts and non-experts, correctly identify the correct answer.

Sample 2 (Figure 15) presents a moderately challenging sample. Only human experts and some of the best models identify the correct answer.



? Question: Which expert commentary best matches the provided video?

- 1 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. ✓
- 2 The player 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.
- 3 The player 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.
- 4 The player should concentrate on jumping higher to increase the shot's velocity, which will make the ball accurately land in the midpoint between the rims.
- 5 The player 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	Human
GPT-4o: Option 1 ✓	Human Expert: Option 1 ✓
Gemini 1.5 Pro: Option 1 ✓	Human Non-Expert: Option 1 ✓
LLaVA-Video: Option 1 ✓	

Figure 14: Sample 1 (Basketball): All models, as well as both human experts and non-experts, select the correct answer.

Sample 3 (Figure 16) presents a more difficult case. Only human experts and GPT-4o select the correct answer.

Sample 4 (Figure 17) illustrates a particularly difficult case. None of the models are able to select the correct answer. Only human experts succeed. This highlights a significant gap between current VLM capabilities and expert-level understanding, especially for tasks that require nuanced, domain-specific reasoning.

These examples collectively highlight the limitations of current VLMs in expert-level understanding of physical human skills.




? Question: Which expert commentary best matches the provided video?

- 1 The participant lowers their shoulder to focus on playing at the tip, improving their ability to execute fast passages.
- 2 The musician prominently uses their wrist to achieve playing close to the frog, allowing better transition between dynamics.
- 3 The violinist skillfully extends their upper arm away from the body to maintain control while playing near the frog, helping in balancing the instrument.
- 4 The participant effectively uses their upper arm, bringing it closer to their body to play near the frog. ✓
- 5 The violinist utilizes finger strength to move towards the tip of the bow, ensuring a brighter sound while keeping the bow under control.

Model response		Human	
GPT-4o: Option 4	✓	Human Expert: Option 4	✓
Gemini 1.5 Pro: Option 4	✓	Human Non-Expert: Option 3	✗
LLaVA-Video: Option 3	✗		

Figure 15: Sample 2 (Violin): Some models and human experts select the correct answer.




? Question: Which expert commentary best matches the provided video?

- 1 The participant does an excellent job at drying the swab with a cloth to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
- 2 The participant does an excellent job at shaking the tube to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
- 3 The participant does an excellent job at pressing the swab to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
- 4 The participant does an excellent job at swirling the swab in the air to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
- 5 The participant does an excellent job at squeezing the tube to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results. ✓

Model response	Human
GPT-4o: Option 5 ✓	Human Expert: Option 5 ✓
Gemini 1.5 Pro: Option 1 ✗	Human Non-Expert: Option 1 ✗
LLaVA-Video: Option 1 ✗	

Figure 16: Sample 3 (COVID-19 Safety): Only GPT-4o and human experts select the correct answer.



? Question: Which expert commentary best matches the provided video?

- 1 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.
- 2 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. ✓
- 3 The participant executes a nice jump to the side with well-bent knees and performs a smooth cartwheel to maintain rhythm and movement.
- 4 The participant executes a nice jump upwards with locked knees and performs a rigid turn with her upper body to maintain rhythm and movement.
- 5 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		Human	
GPT-4o: Option 1	✗	Human Expert: Option 2	✓
Gemini 1.5 Pro: Option 1	✗	Human Non-Expert: Option 3	✗
LLaVA-Video: Option 1	✗		

Figure 17: Sample 4 (Dance): None of the models select the correct answer. Only human experts identify the correct response.

S4 Evaluated Capabilities in EXACT

This section details the reasoning capabilities implicitly evaluated by our QA framework. Although each question follows a simple multiple-choice format, correctly answering them requires diverse capabilities that extend beyond visual recognition. The evaluated reasoning types include:

- **Fine-grained recognition:** Detecting subtle differences in body positioning, movement patterns, and technique execution (e.g., “plays with a very tight wrist” vs. “uses a hybrid picking technique”).
- **Temporal reasoning:** Understanding timing precision, action phases, and temporal dependencies (e.g., “releases the ball too early” vs. “too late after the peak jump”).
- **Causal understanding:** Linking technique flaws to performance outcomes (e.g., “doesn’t bend knees...reduces jump height” vs. “doesn’t bend knees...prevents ball spin”).
- **Procedural reasoning:** Evaluating adherence to task sequences (e.g., “checks pulse before calling for help” vs. “delays calling for help and requesting an AED”).
- **Spatial & geometric reasoning:** Understanding angles, trajectories, and spatial relationships (e.g., “flattens the arc on a left-handed layup” vs. “lowers the entry angle to avoid rotation”).
- **Domain-specific expertise:** Applying professional knowledge across sports, music, healthcare, cooking, bike repair, and dance (e.g., “maintains CPR rate at 100–120 bpm” vs. “120–140 bpm”).
- **Physical understanding:** Reasoning about biomechanics, force generation, and energy transfer (e.g., “angles knees inward for power” vs. “opens stance toward the rim”).

S5 Sample statistics after each filtering stage

Starting from over 400k expert commentaries in Ego-Exo4D, Stage I reduced the number of samples to 108k ($\approx -73\%$). Stage III’s length similarity filtering further reduced it from 108k to 60k ($\approx -44\%$). The blind-LLM filtering further shrank the sample size from 60k to 4.5k ($\approx -92.5\%$). Finally, Stage IV expert verification led to 3.5k samples ($\approx -22\%$ of the 4.5k), leaving about 0.9% of the original samples. These statistics demonstrate that Stage III is critical as it filters low-quality or easily “cheatable” samples before sending them to the experts for final verification.