Video-STaR

Self-Training Enables Video Instruction Tuning with Any Supervision

Orr Zohar

Xiaohan Wang

Yonatan Bitton

Idan Szpektor

Serena Yeung-Levy

Why do we care about Video-LMMs?

Augmented Reality (AR) Applications

Virtual Customer Support Agent

Robotics

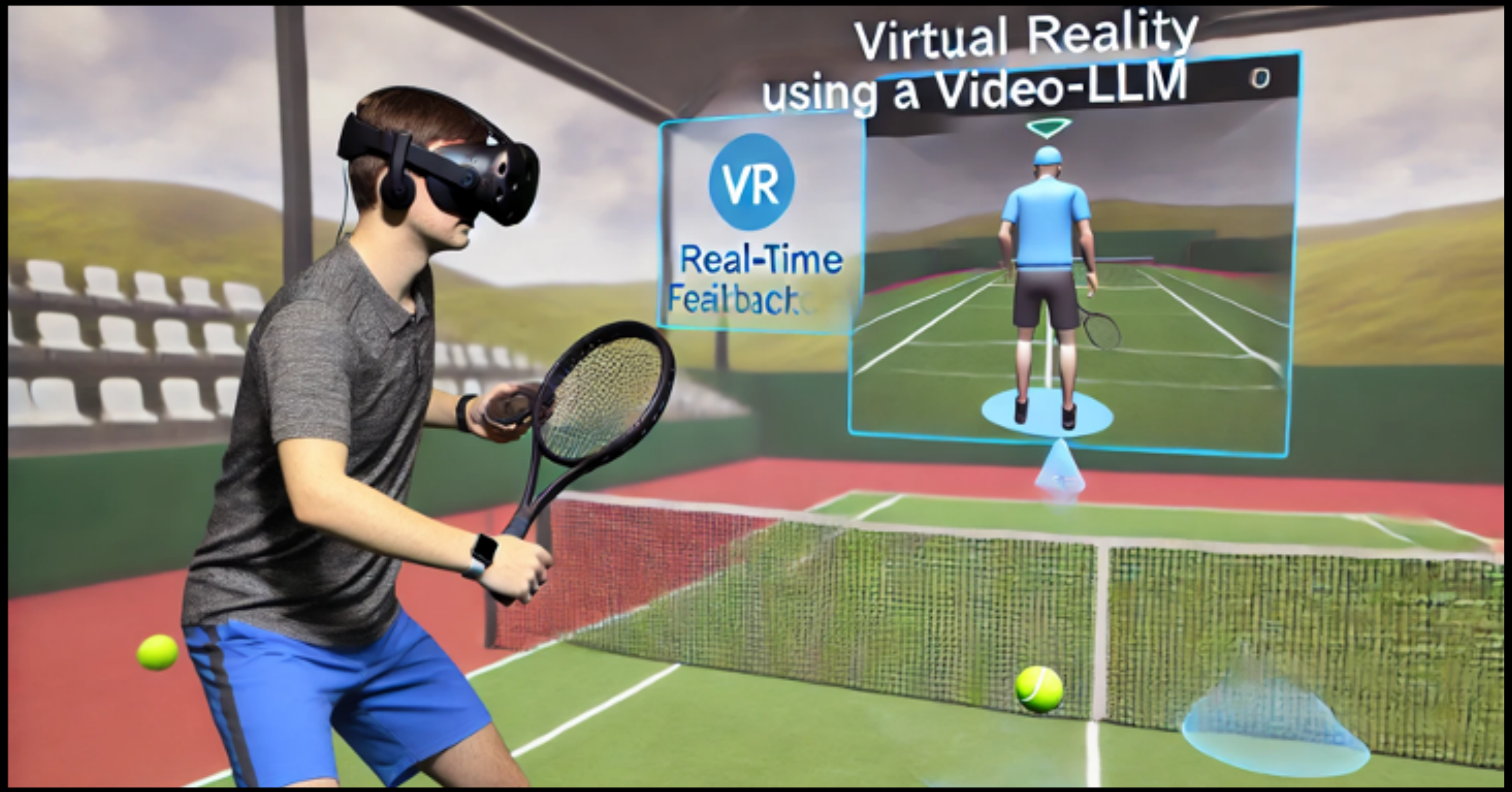
Judge

Coach

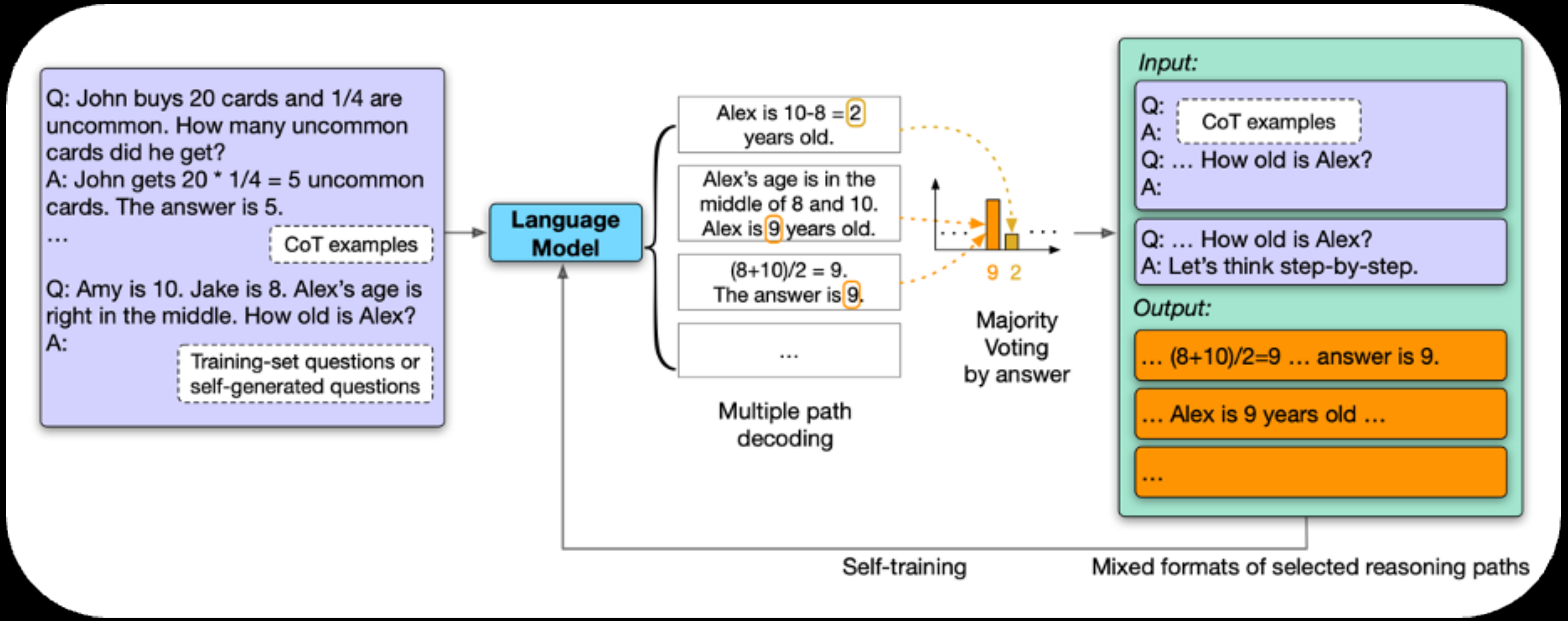
Surgical Assistants

Cooking Instructor

…



Background: Self-Training in Large Language Models



Q: John buys 20 cards and 1/4 are uncommon. How many uncommon cards did he get?
A: John gets 20 * 1/4 = 5 uncommon cards. The answer is 5.
...

Q: Amy is 10. Jake is 8. Alex's age is right in the middle. How old is Alex?
A:

Alex is 10-8 = 2 years old.

Alex's age is in the middle of 8 and 10. Alex is 9 years old.

(8+10)/2 = 9. The answer is 9.

...

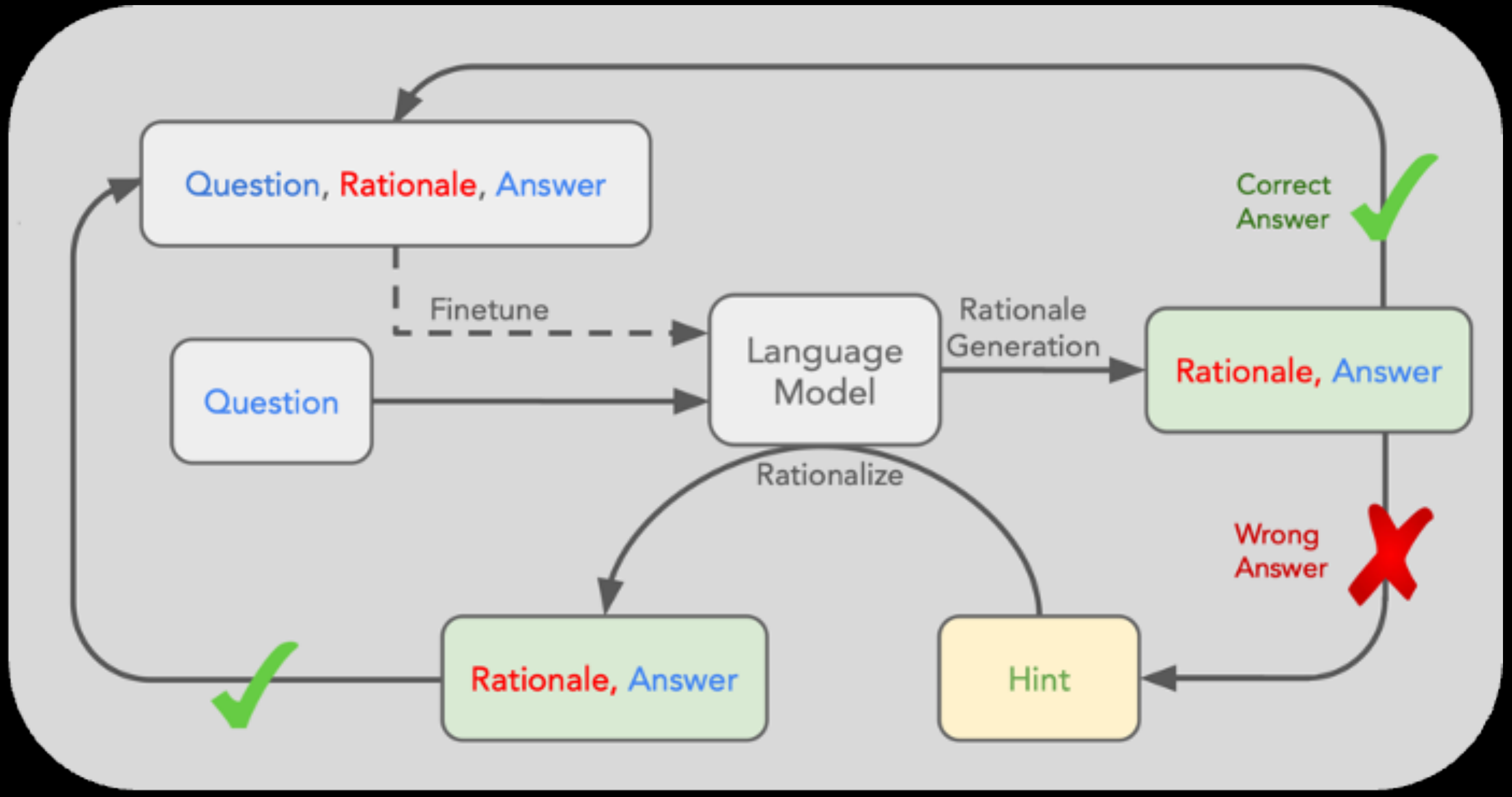
Input:
Q: CoT examples
A:
Q: ... How old is Alex?
A:

Q: ... How old is Alex?
A: Let's think step-by-step.

Output:
... (8+10)/2=9 ... answer is 9.
... Alex is 9 years old ...
...

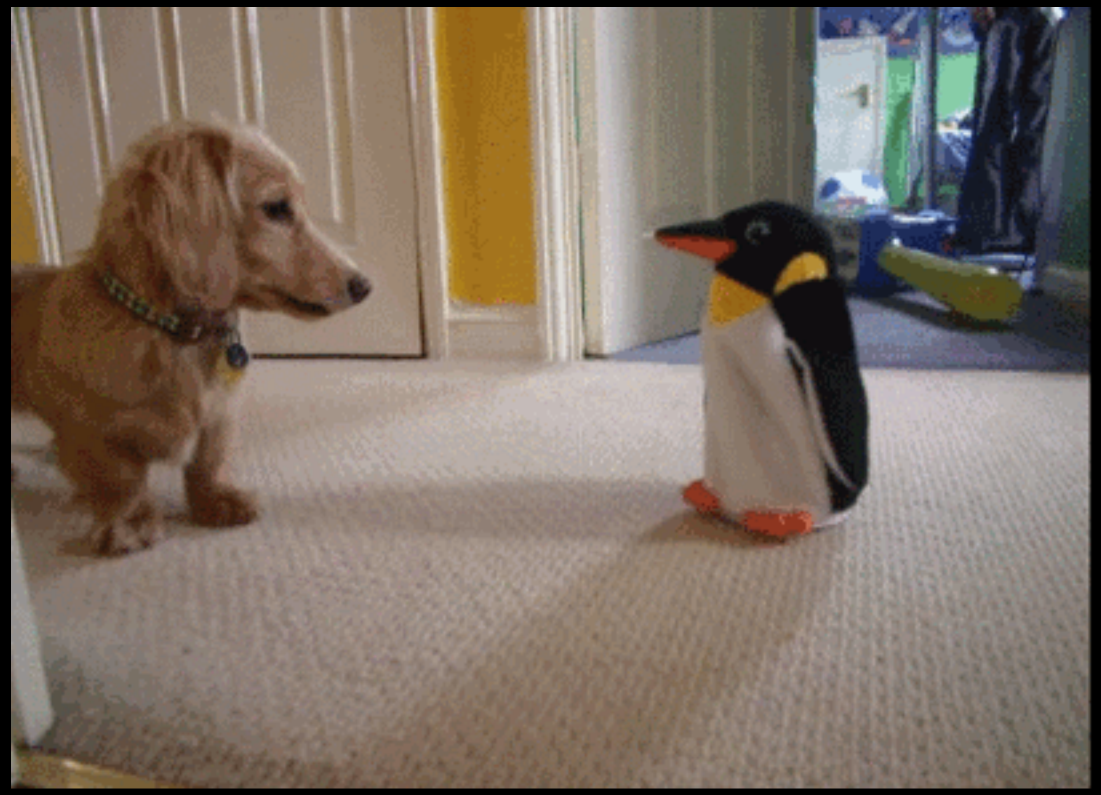
CoT examples
Training-set questions or self-generated questions
Language Model
Multiple path decoding
9 2
Majority Voting by answer
Self-training
Mixed formats of selected reasoning paths

Majority Voting
Large Language Models Can Self-Improve, Huang et. al, 2022

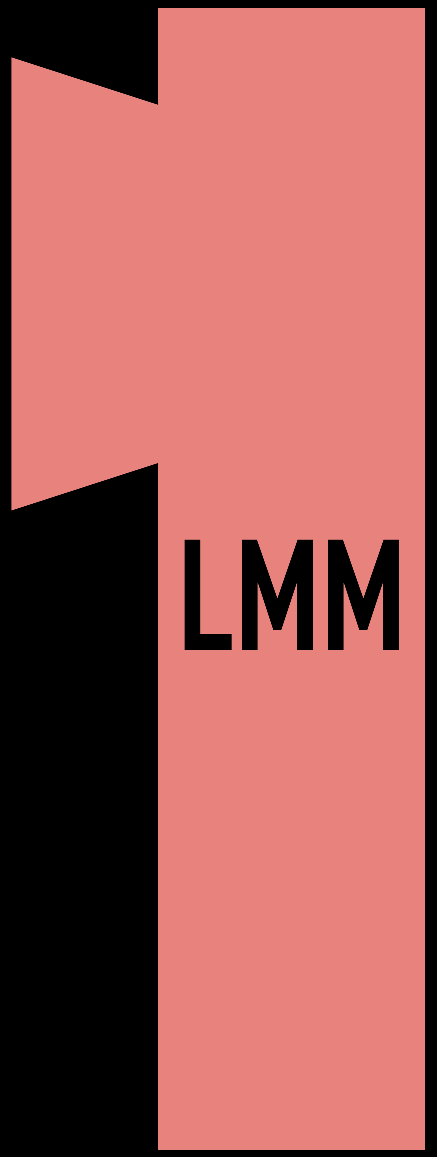


Rationalization
STaR: Bootstrapping Reasoning With Reasoning, Zelikman et. Al, 2020

Video Instruction Tuning



Question: What room does the dog first go in to?



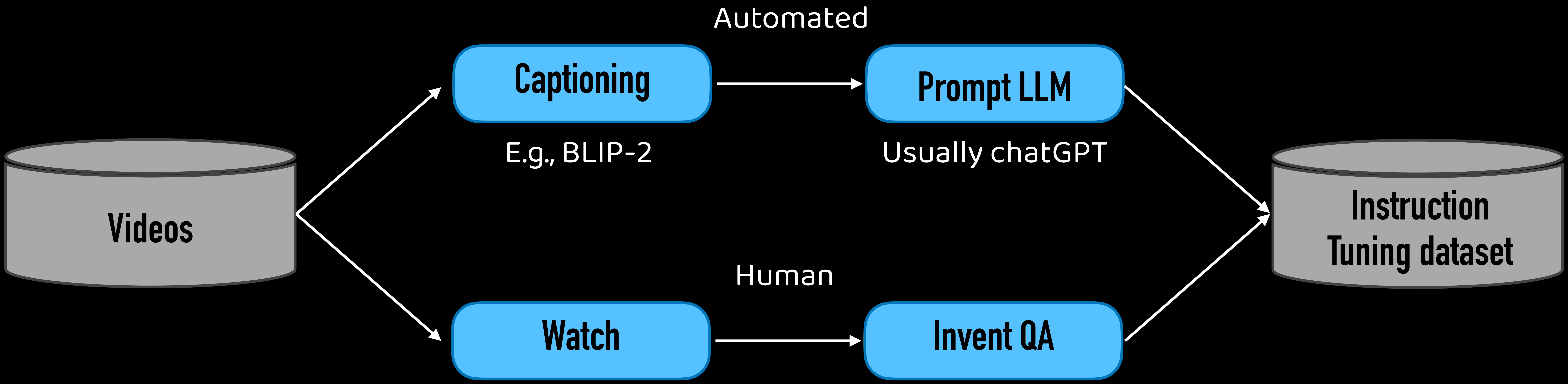
Answer: The room on the right

Video Instruction Tuning

However, collecting this data is hard because:

1. Requires annotators to watch long videos and come up with question-answer pairs
2. Is frequently ambiguous
3. Different tasks will require new datasets

Existing annotation approaches



Automated

Videos

Captioning

E.g., BLIP-2

Prompt LLM

Usually chatGPT

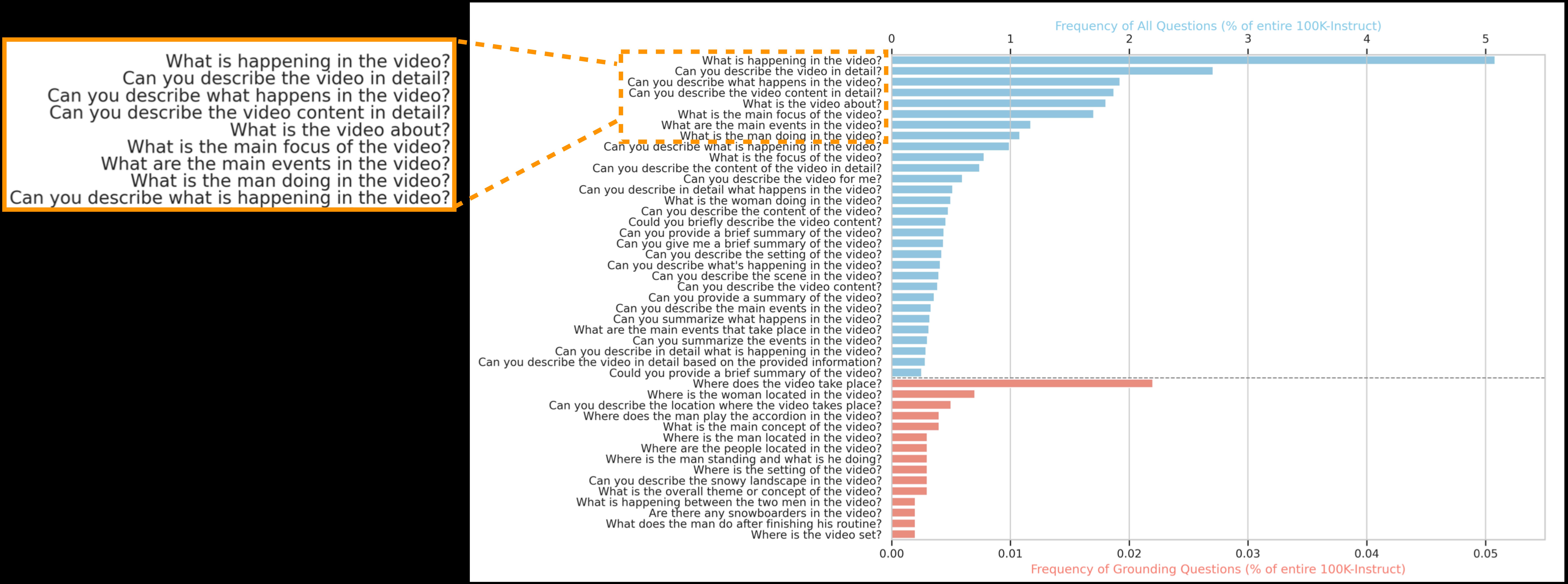
Instruction Tuning dataset

Human

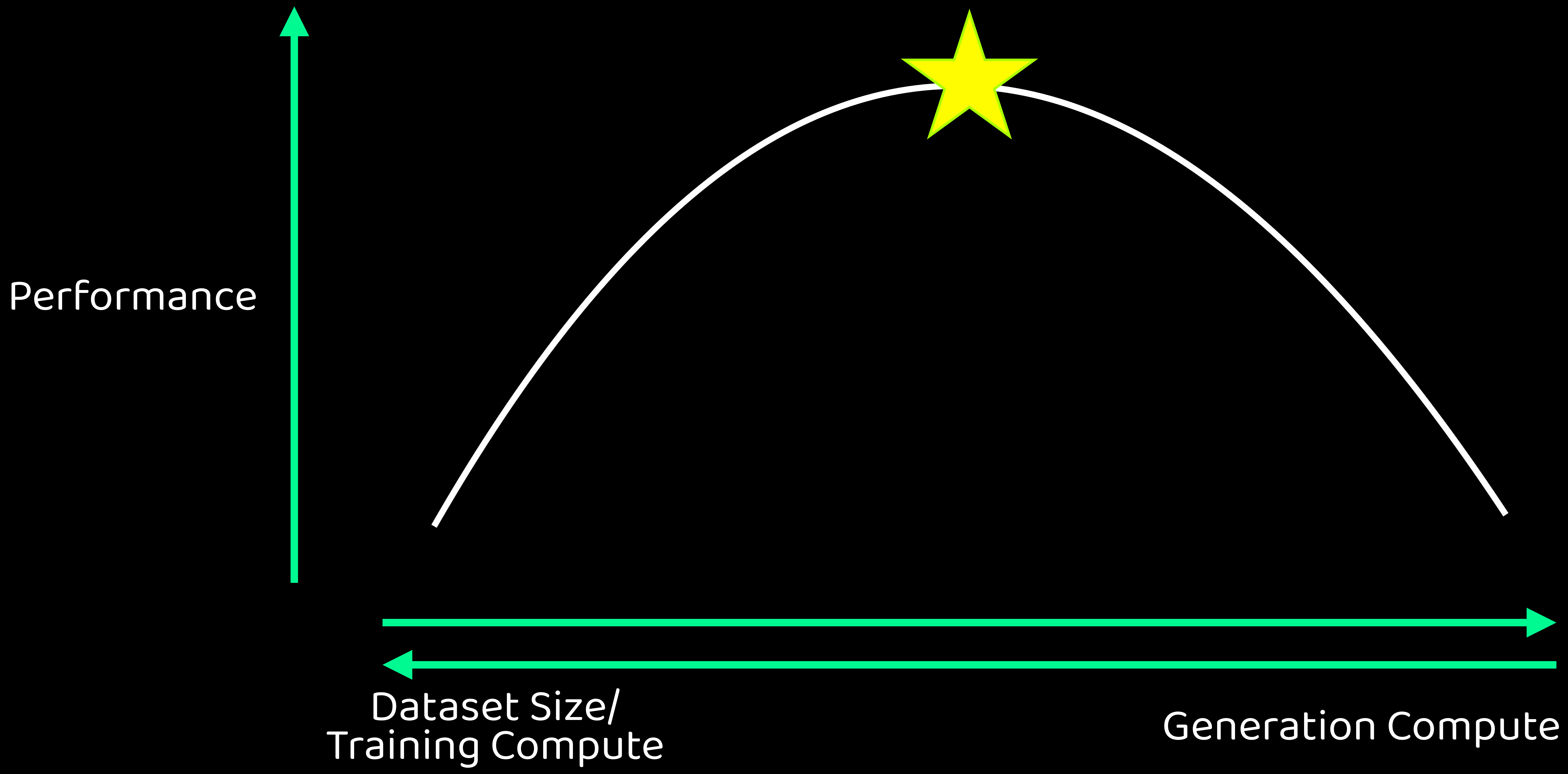
Watch

Invent QA

Resulting Video Instruction Tuning Datasets



Compute–Dataset Size Tradeoff



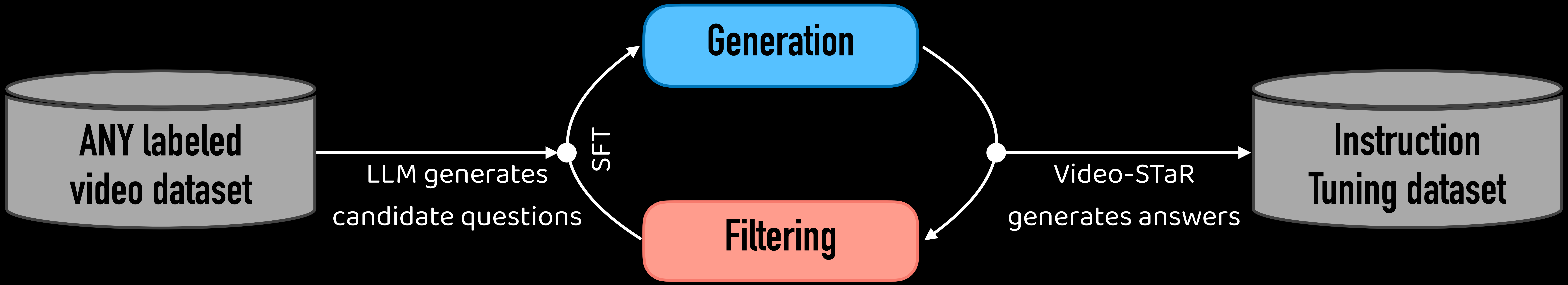
Video-STaR

Improve LMM Video-QA performance

Adapt LMMs to new applications

Video-STaR

Use any video label for video instruction tuning!


ANY labeled video dataset

LLM generates candidate questions

SFT

Generation

Filtering

Video-STaR generates answers

Instruction Tuning dataset

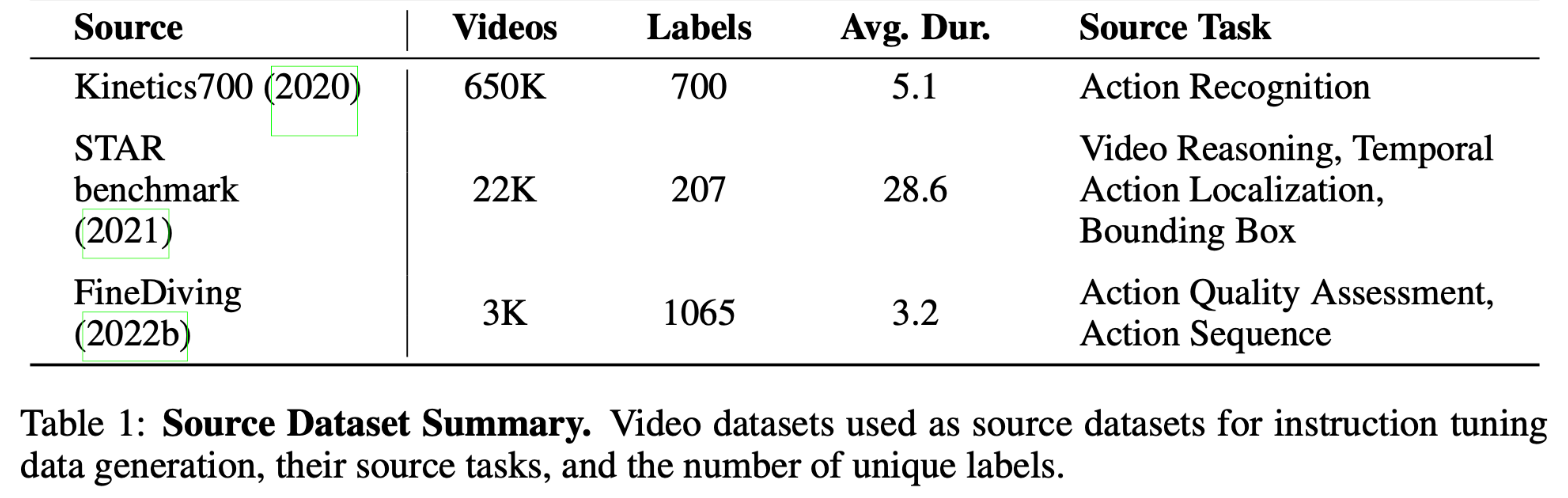
Video-STaR

Definitions

- Given: video's v_i with corresponding labels l_i: $\mathscr{D} = \{(v_i, l_i)\}_{i=1}^{d}$
- Goal: generate questions (q) and answer (a) pairs: $\hat{\mathscr{D}} = \{(v_i, q_i, a_i)\}_{i=1}^{d_f}$
- In cycle i, the instruction-tuned model $\hat{M}^{(i-1)\star}$ is used to generate training data.
- In cycle i, the instruction-tuned model $\hat{M}^{i\star}$ is produced, by training on the generated data.

Video-STaR

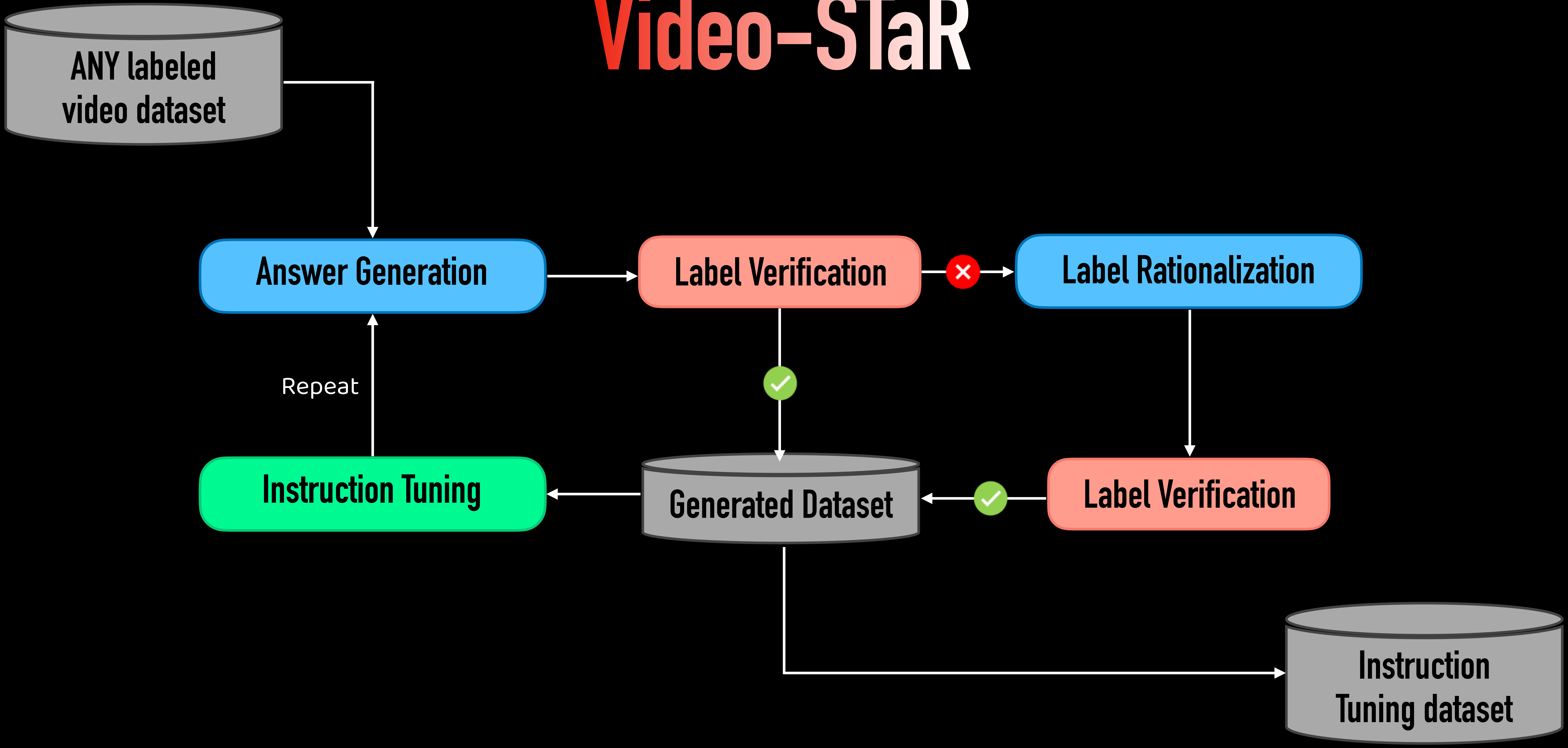
Our Source Datasets



Source	Videos	Labels	Avg. Dur.	Source Task
Kinetics700 (2020)	650K	700	5.1	Action Recognition
STAR benchmark (2021)	22K	207	28.6	Video Reasoning, Temporal Action Localization, Bounding Box
FineDiving (2022b)	3K	1065	3.2	Action Quality Assessment, Action Sequence

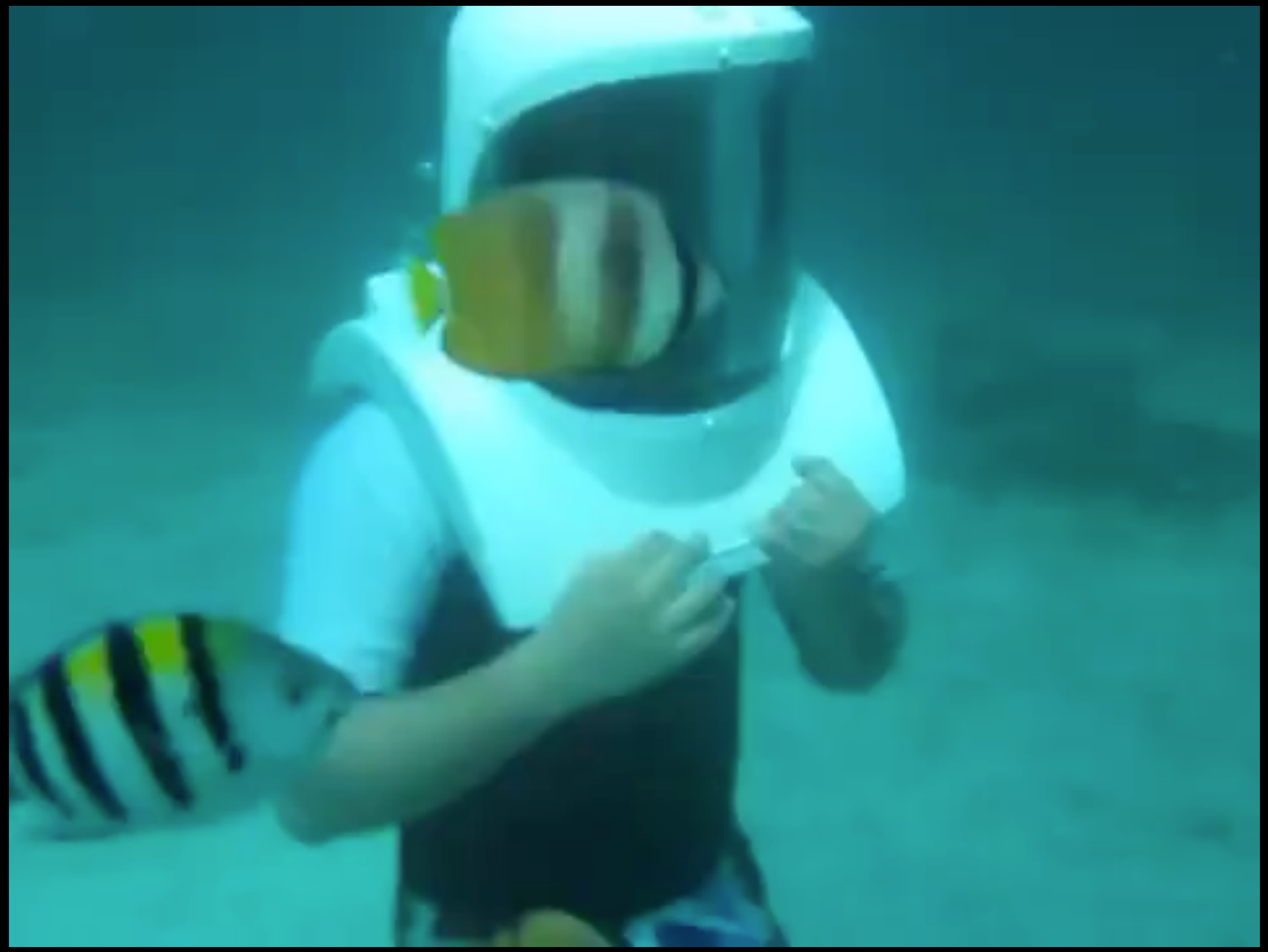
Table 1: **Source Dataset Summary.** Video datasets used as source datasets for instruction tuning data generation, their source tasks, and the number of unique labels.

Video-STaR

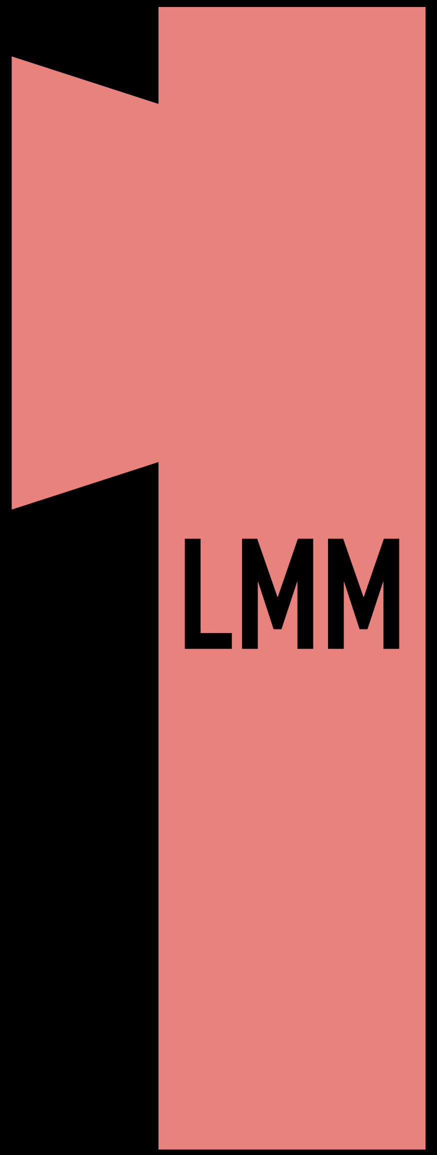

ANY labeled video dataset
Answer Generation
Label Verification
✗
Label Rationalization
✓
Repeat
Instruction Tuning
Generated Dataset
✓
Label Verification
Instruction Tuning dataset

Video-STaR

Answer Generation



What activity is happening in this video?

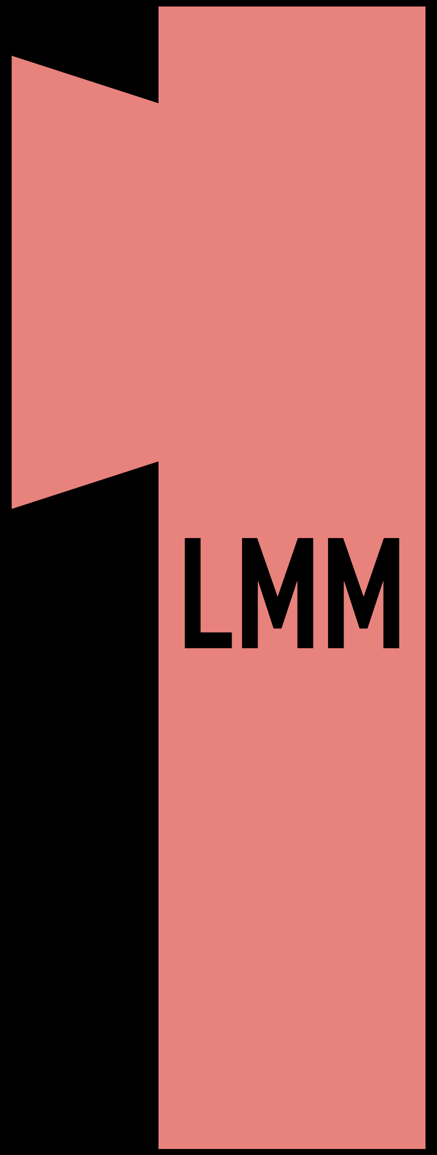


The video shows a man and a woman, both wearing helmets, swimming underwater and interacting with each other and a fish. Their helmets indicate they are engaged in helmet diving, an underwater activity that allows breathing and protection.

Video-STaR

Label Rationalization

Question: Which object was taken by the person?Answer: The shoe.
Can you rationalize the answer step-by-step? How can one arrive at this conclusion?



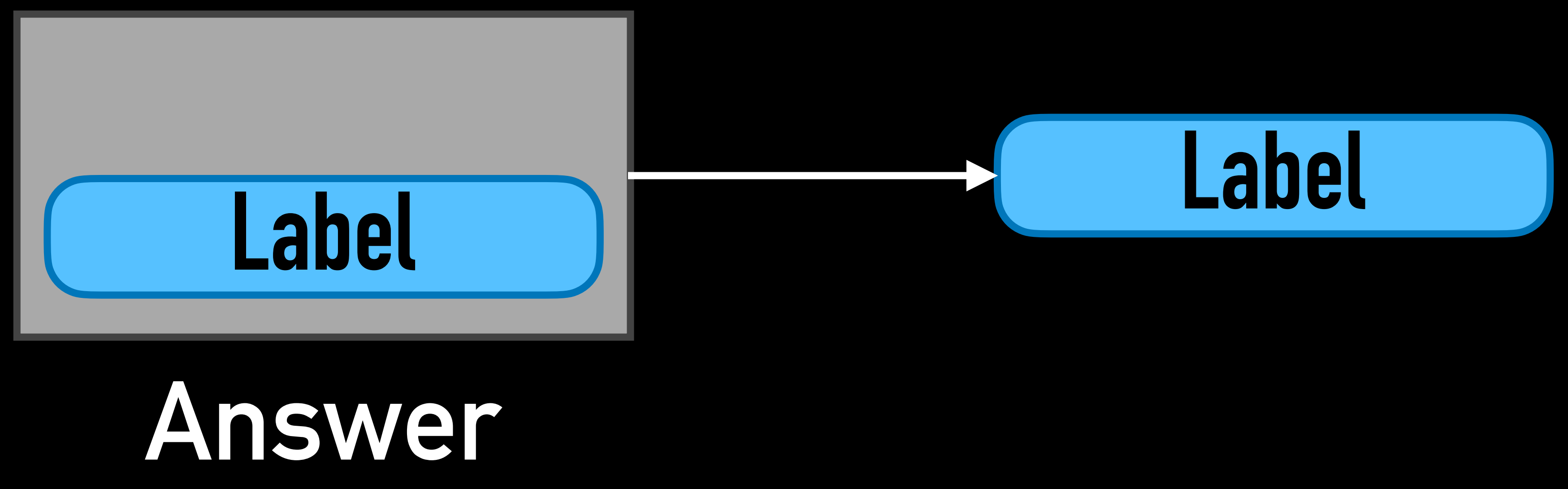
The video shows the person picking up an object from the floor, which is identified as a shoe. This action suggests the person intends to use the shoe, likely to wear it. Thus, the object taken by the person is the shoe.

Video-STaR

Label Verifier

Parser

Tasked with extracting the different video labels from the generated text



Label

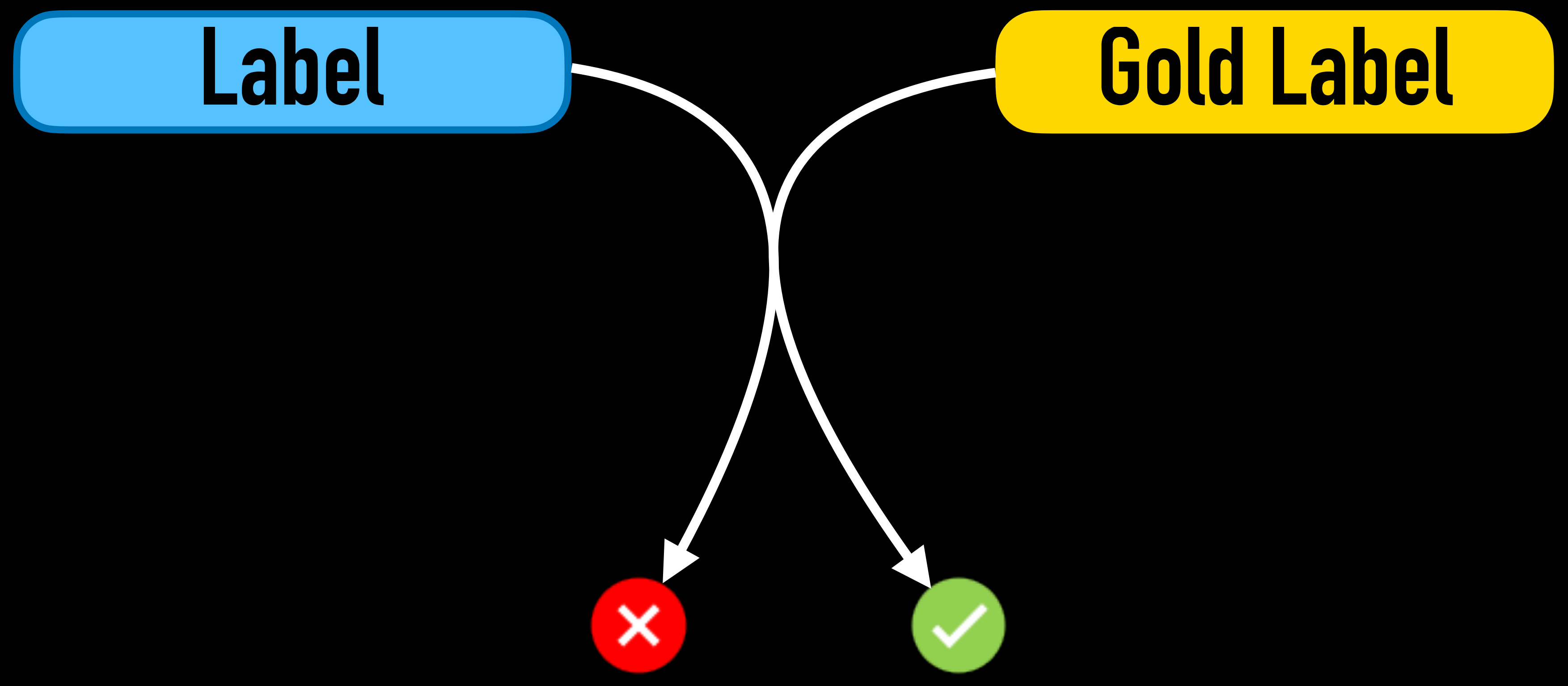
Label

Answer

Entity recognition,
regex, fuzzy logic, etc.

Verifier

Tasked with comparing the extracted label with the GT label.



Label

Gold Label

✗ ✓

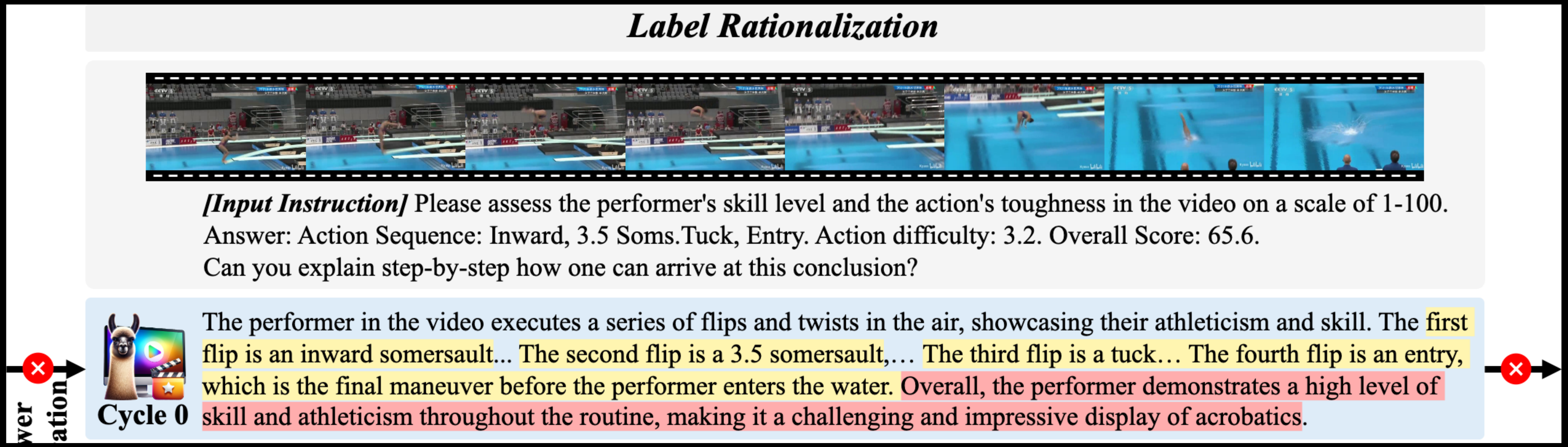
TAL/bbox -> IoU, text -> BERT embeddings
Floats -> abs/rel delta

Video-STaR

Data flow

Cycle 0

Answer Generation



Label Rationalization

[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100. Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6. Can you explain step-by-step how one can arrive at this conclusion?

Cycle 0

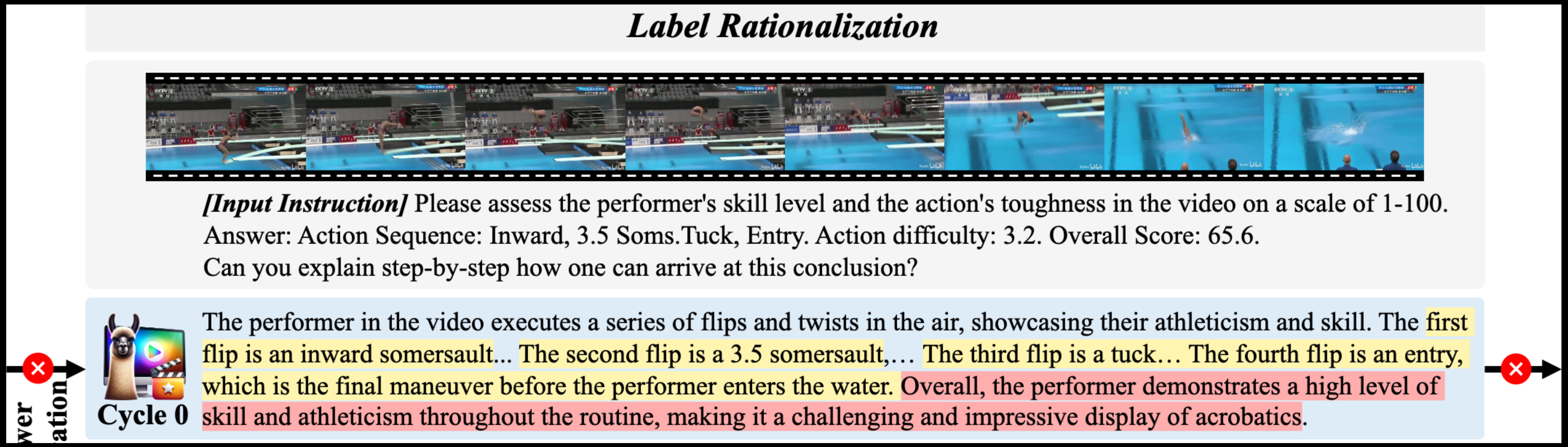
The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Video-STaR

Data flow

Cycle 0

Answer Generation → **Label Verification**



Label Rationalization

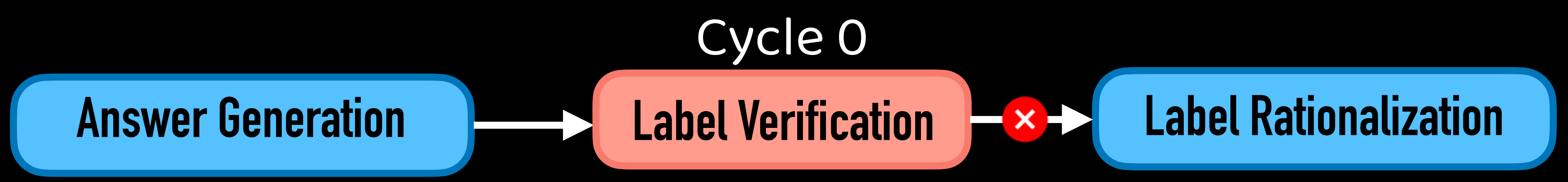
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100. Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6. Can you explain step-by-step how one can arrive at this conclusion?

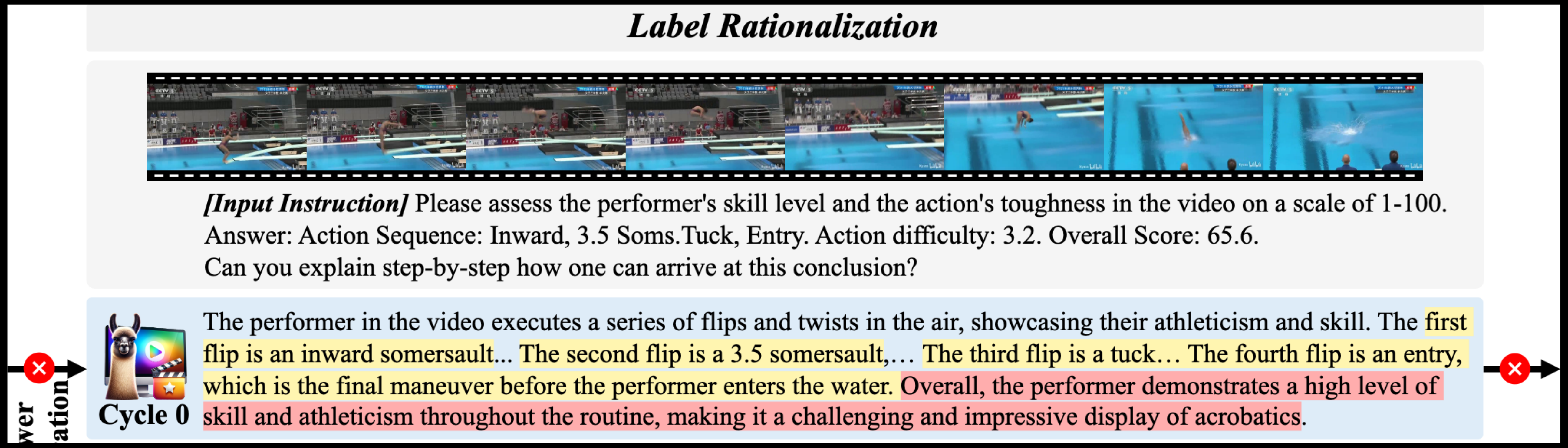
Cycle 0

The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Video-STaR

Data flow





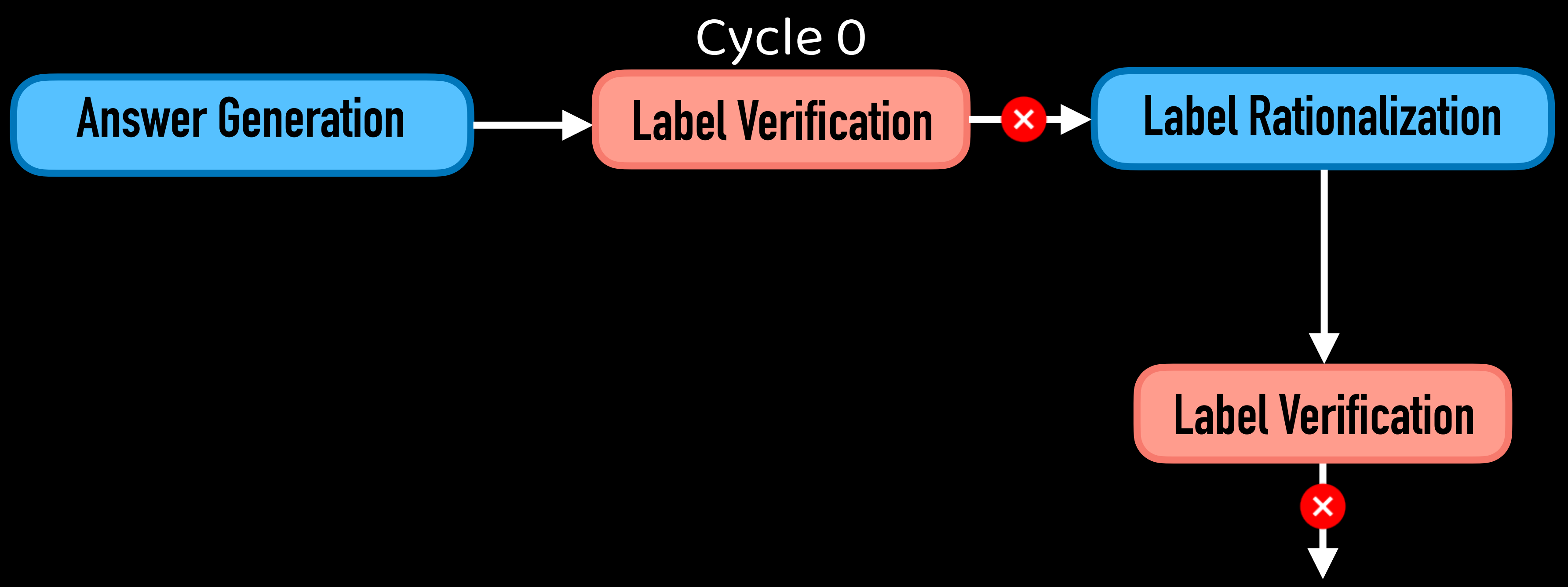
Label Rationalization

[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100. Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6. Can you explain step-by-step how one can arrive at this conclusion?

The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Video-STaR

Data flow



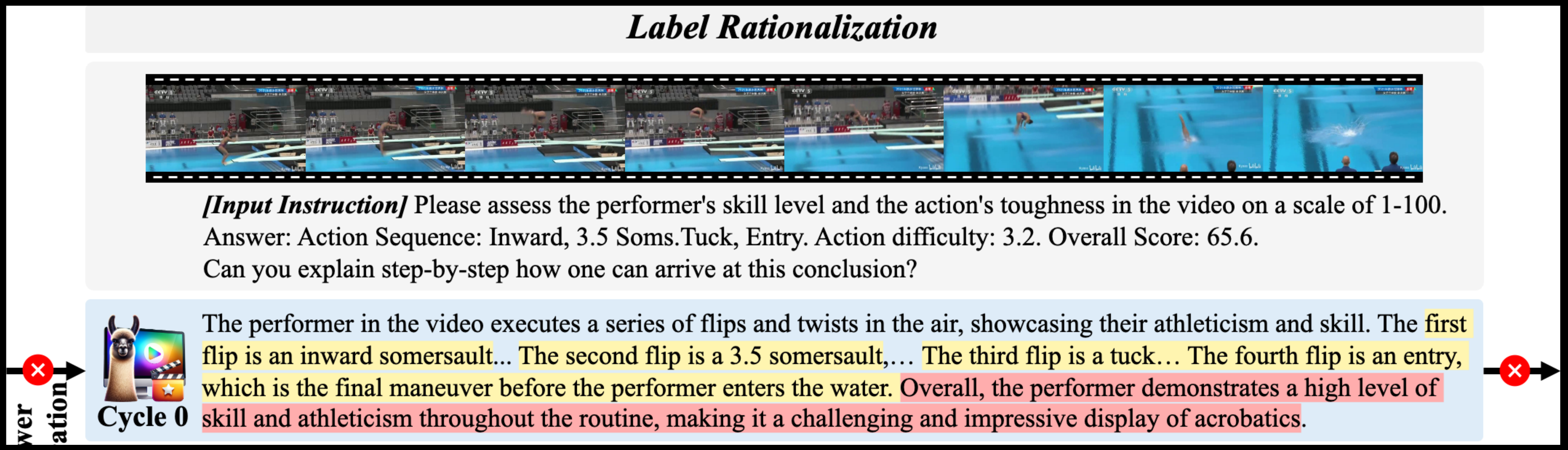
Cycle 0

Answer Generation

Label Verification

Label Rationalization

Label Verification



Label Rationalization

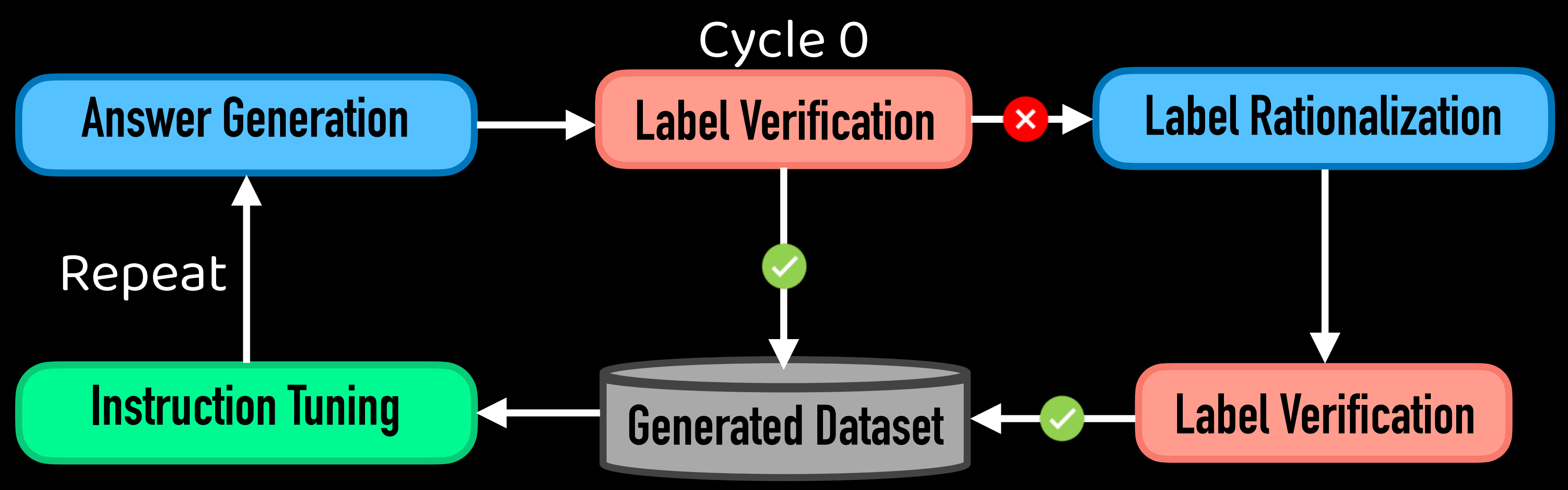
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100. Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6. Can you explain step-by-step how one can arrive at this conclusion?

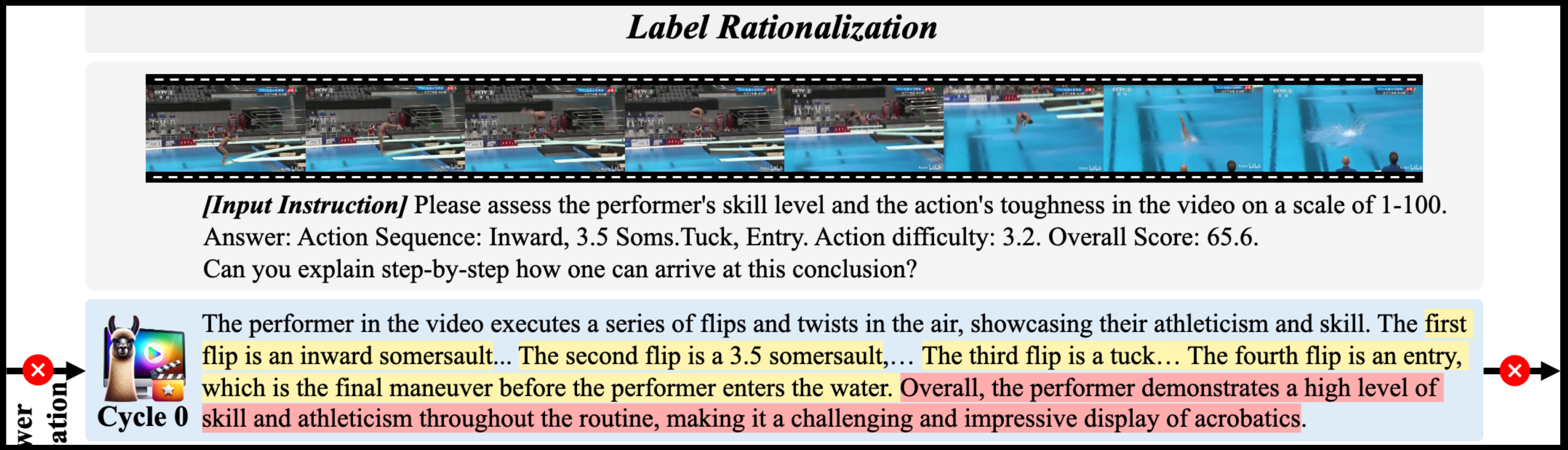
Cycle 0

The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Video-STaR

Data flow





Label Rationalization

[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.
Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6.
Can you explain step-by-step how one can arrive at this conclusion?

Cycle 0

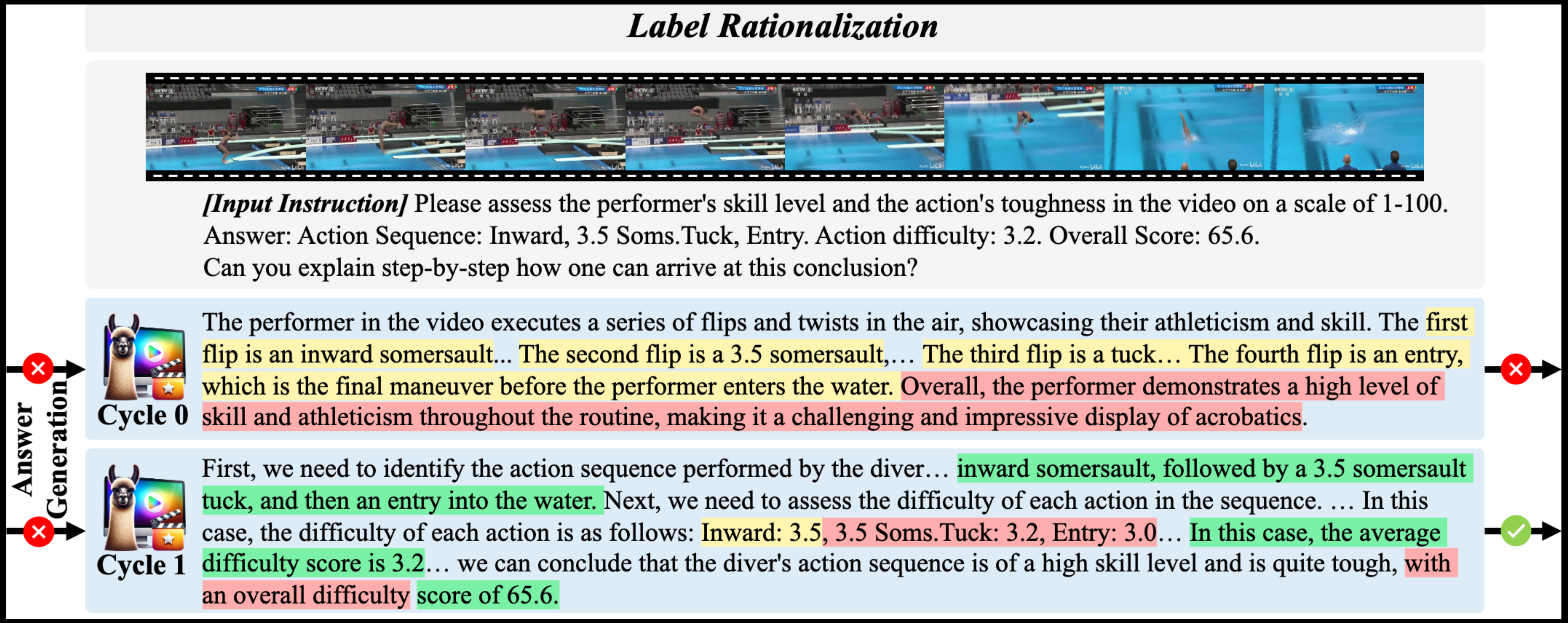
The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Video-STaR

Data flow

Cycle 1

Answer Generation



Label Rationalization

[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100. Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6. Can you explain step-by-step how one can arrive at this conclusion?

Cycle 0

The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Cycle 1

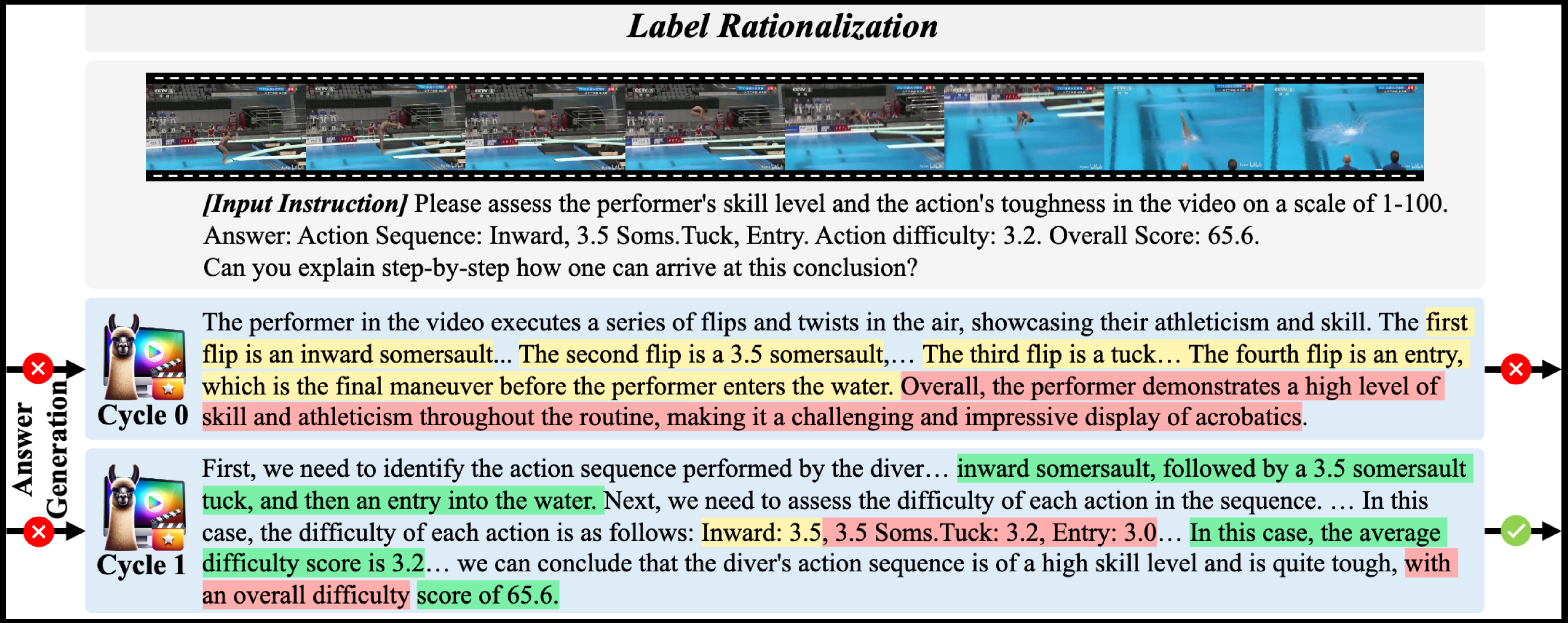
First, we need to identify the action sequence performed by the diver… inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water. Next, we need to assess the difficulty of each action in the sequence. … In this case, the difficulty of each action is as follows: Inward: 3.5, 3.5 Soms.Tuck: 3.2, Entry: 3.0… In this case, the average difficulty score is 3.2… we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall difficulty score of 65.6.

Video-STaR

Data flow

Cycle 1

Answer Generation → **Label Verification**



Label Rationalization

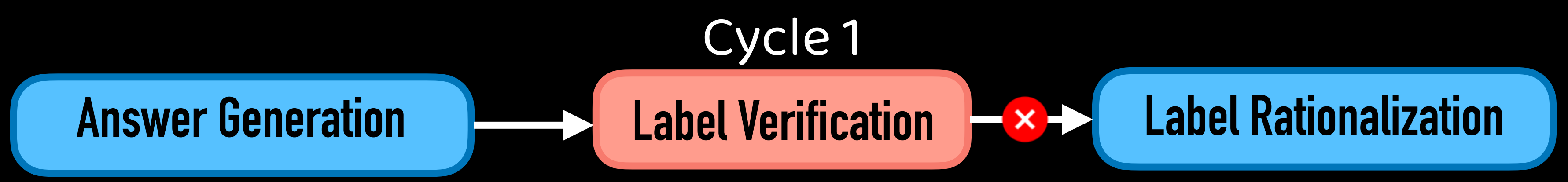
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.
Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6.
Can you explain step-by-step how one can arrive at this conclusion?

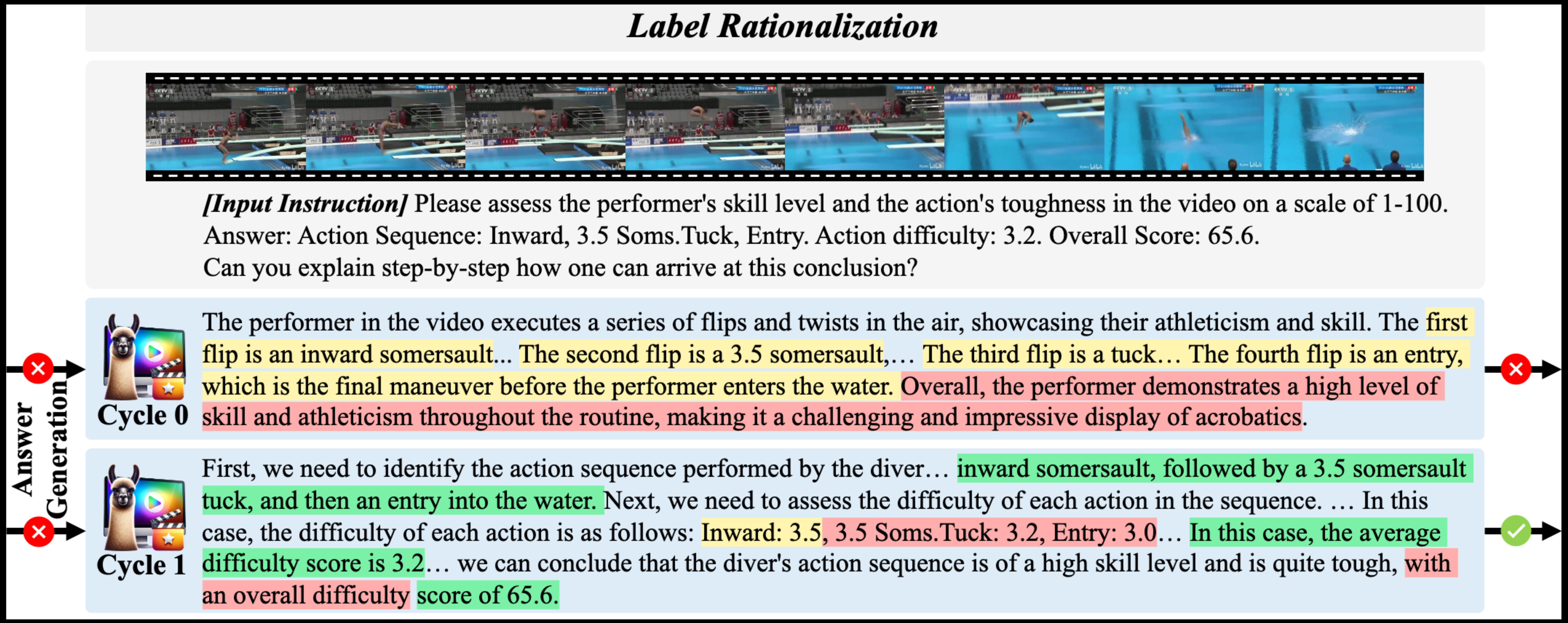
Cycle 0 The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Cycle 1 First, we need to identify the action sequence performed by the diver… inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water. Next, we need to assess the difficulty of each action in the sequence. … In this case, the difficulty of each action is as follows: Inward: 3.5, 3.5 Soms.Tuck: 3.2, Entry: 3.0… In this case, the average difficulty score is 3.2… we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall difficulty score of 65.6.

Video-STaR

Data flow





Label Rationalization

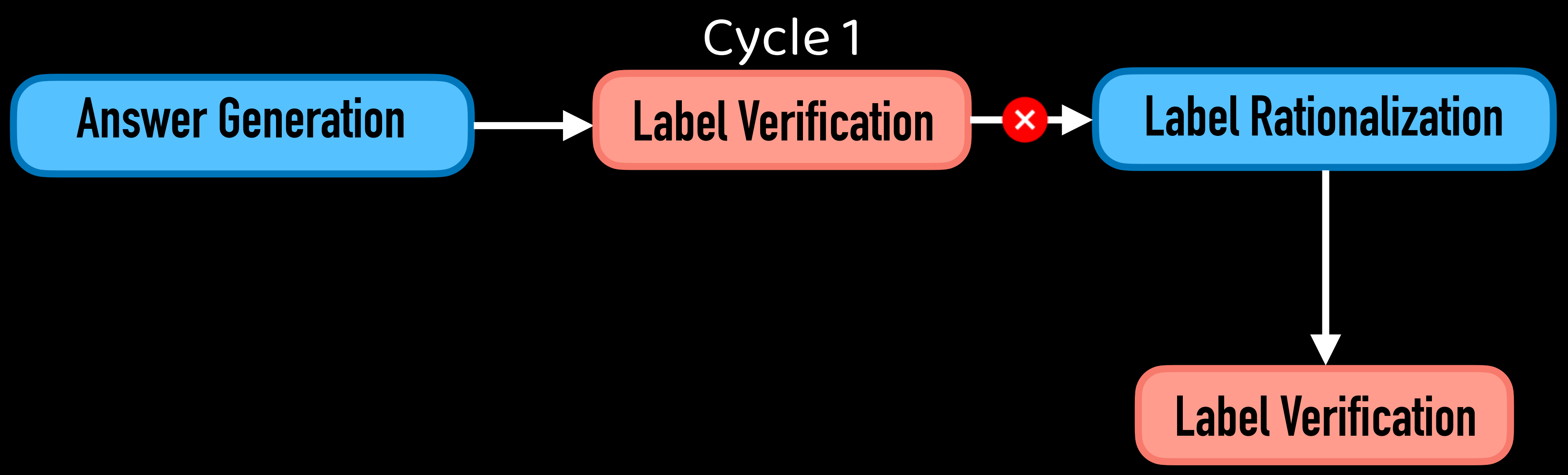
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.
Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6.
Can you explain step-by-step how one can arrive at this conclusion?

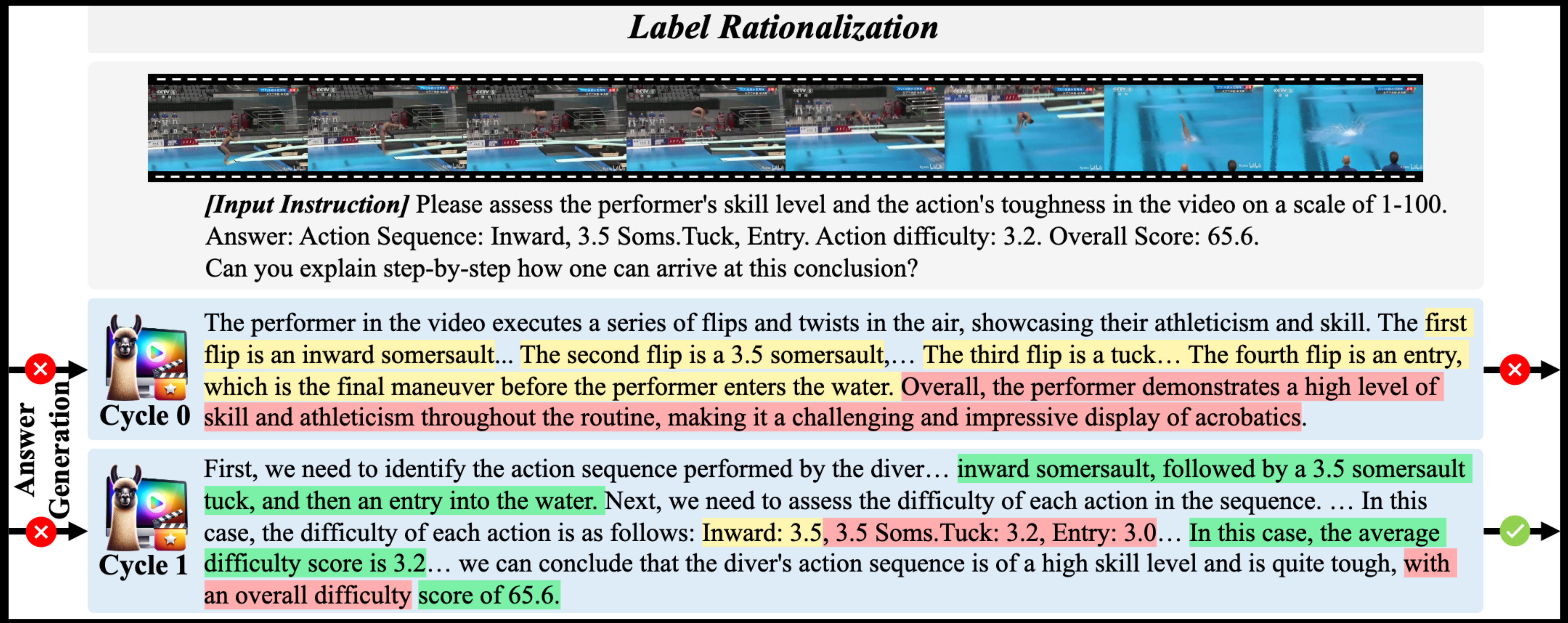
Cycle 0: The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

Cycle 1: First, we need to identify the action sequence performed by the diver… inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water. Next, we need to assess the difficulty of each action in the sequence. … In this case, the difficulty of each action is as follows: Inward: 3.5, 3.5 Soms.Tuck: 3.2, Entry: 3.0… In this case, the average difficulty score is 3.2… we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall difficulty score of 65.6.

Video-STaR

Data flow




Answer Generation
Cycle 0
Cycle 1

Label Rationalization

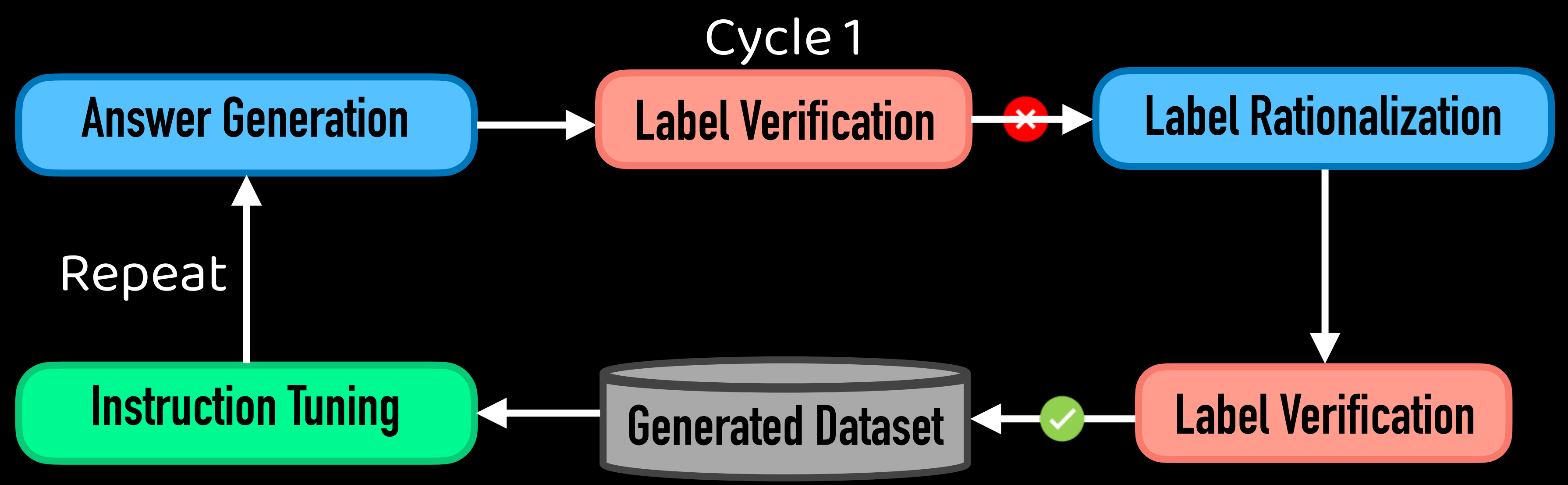
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.
Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6.
Can you explain step-by-step how one can arrive at this conclusion?

Cycle 0: The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics. ✗

Cycle 1: First, we need to identify the action sequence performed by the diver… inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water. Next, we need to assess the difficulty of each action in the sequence. … In this case, the difficulty of each action is as follows: Inward: 3.5, 3.5 Soms.Tuck: 3.2, Entry: 3.0… In this case, the average difficulty score is 3.2… we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall difficulty score of 65.6. ✓

Video-STaR

Data flow


Cycle 1

Answer Generation

Label Verification

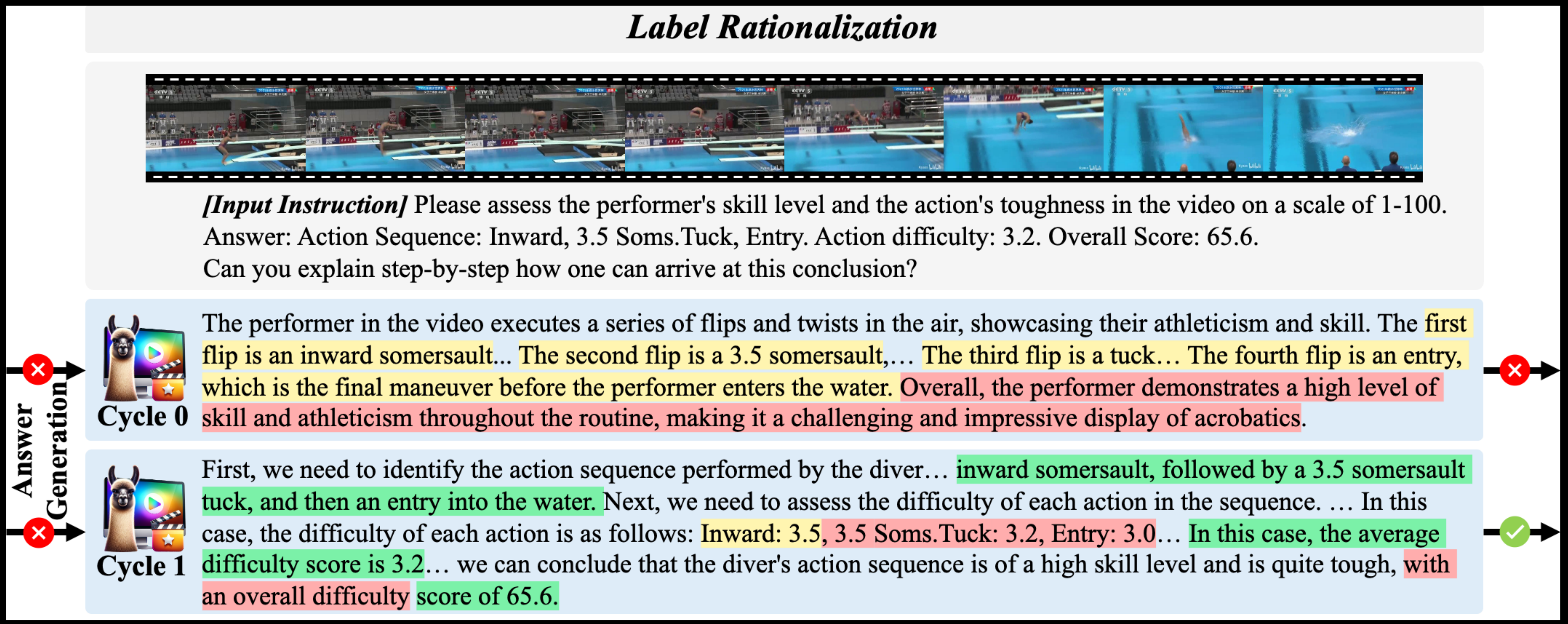
Label Rationalization

Label Verification

Generated Dataset

Instruction Tuning

Repeat



Label Rationalization

[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100. Answer: Action Sequence: Inward, 3.5 Soms.Tuck, Entry. Action difficulty: 3.2. Overall Score: 65.6. Can you explain step-by-step how one can arrive at this conclusion?

The performer in the video executes a series of flips and twists in the air, showcasing their athleticism and skill. The first flip is an inward somersault... The second flip is a 3.5 somersault,… The third flip is a tuck… The fourth flip is an entry, which is the final maneuver before the performer enters the water. Overall, the performer demonstrates a high level of skill and athleticism throughout the routine, making it a challenging and impressive display of acrobatics.

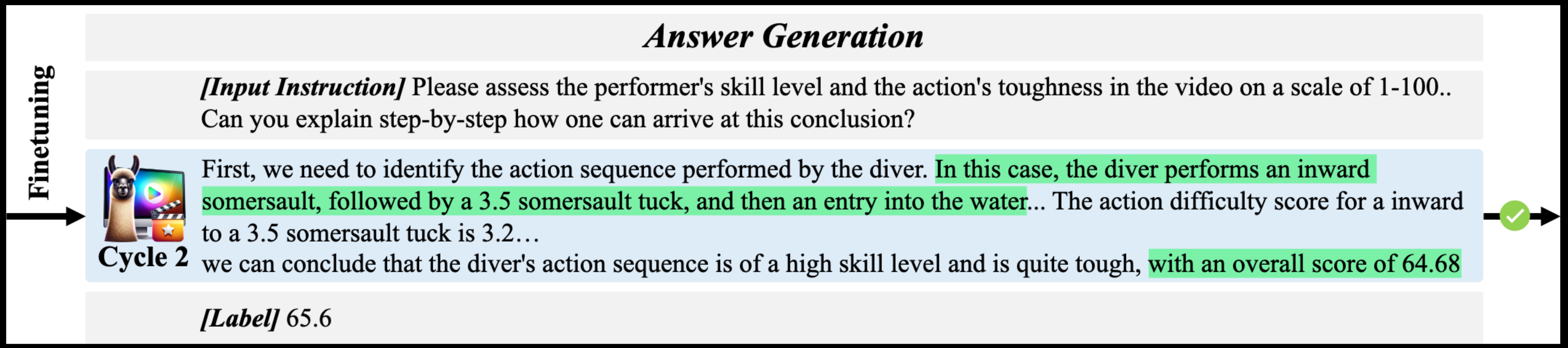
First, we need to identify the action sequence performed by the diver… inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water. Next, we need to assess the difficulty of each action in the sequence. … In this case, the difficulty of each action is as follows: Inward: 3.5, 3.5 Soms.Tuck: 3.2, Entry: 3.0… In this case, the average difficulty score is 3.2… we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall difficulty score of 65.6.

Video-STaR

Data flow

Cycle 2

Answer Generation →



Answer Generation

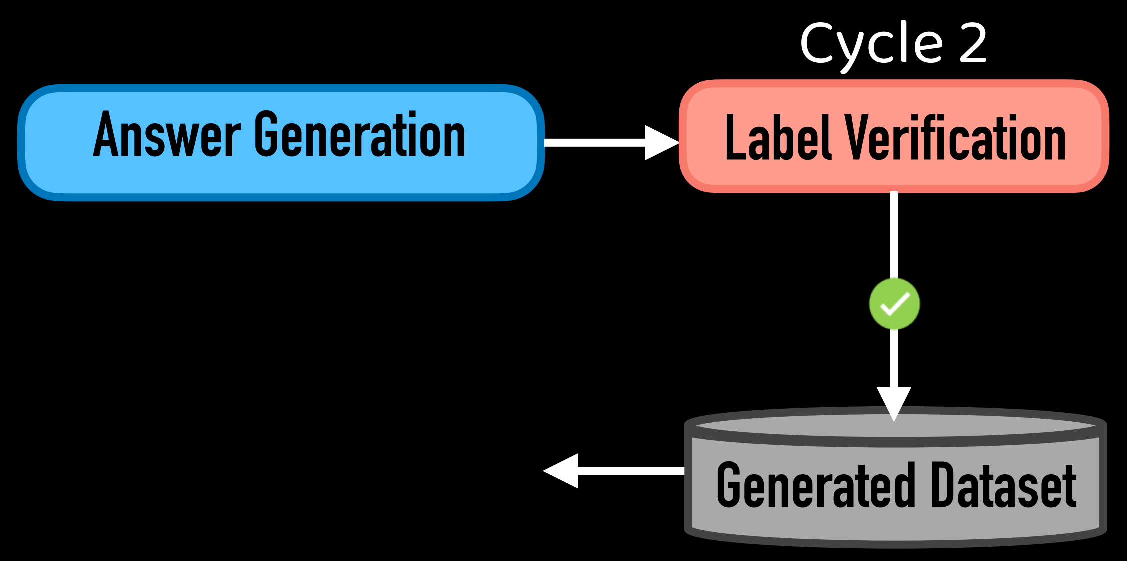
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.. Can you explain step-by-step how one can arrive at this conclusion?

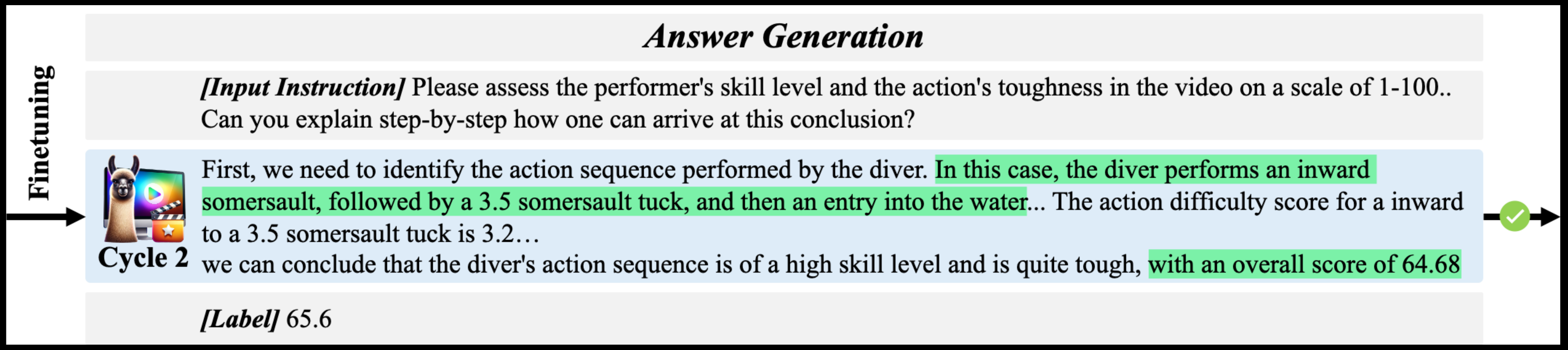
First, we need to identify the action sequence performed by the diver. In this case, the diver performs an inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water... The action difficulty score for a inward to a 3.5 somersault tuck is 3.2…
we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall score of 64.68

[Label] 65.6

Video-STaR

Data flow




Finetuning
Cycle 2

Answer Generation

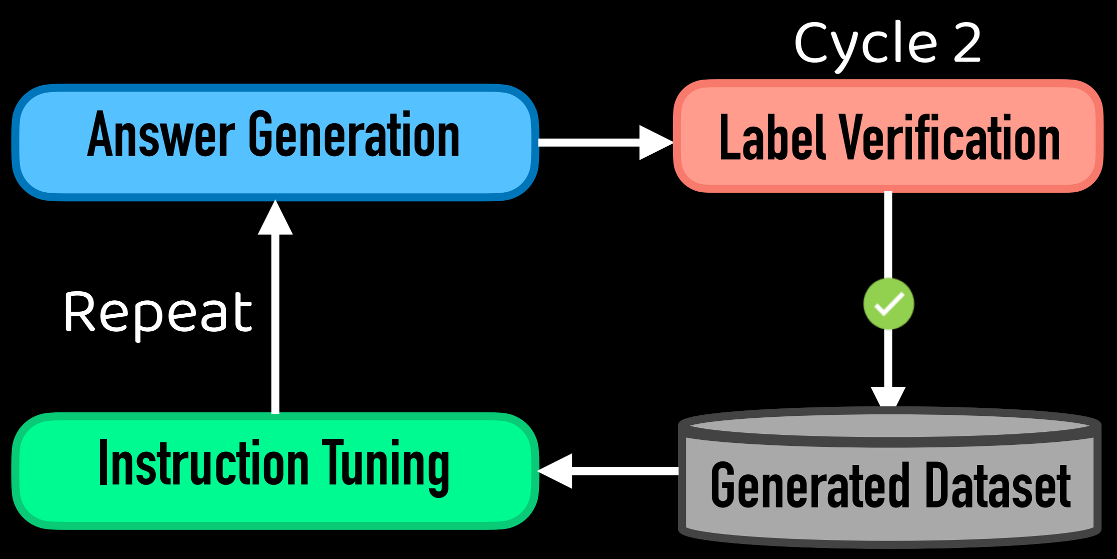
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.. Can you explain step-by-step how one can arrive at this conclusion?

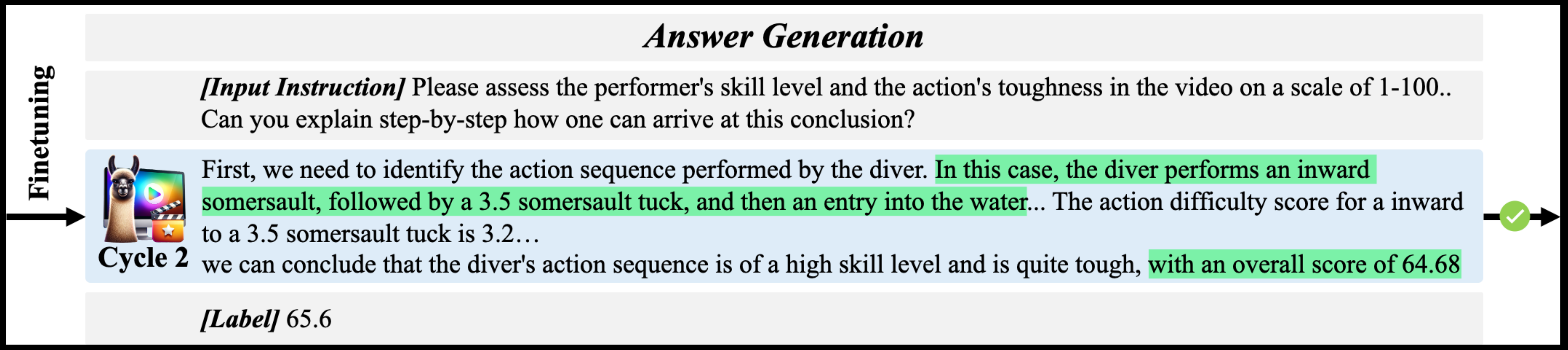
First, we need to identify the action sequence performed by the diver. In this case, the diver performs an inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water... The action difficulty score for a inward to a 3.5 somersault tuck is 3.2…
we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall score of 64.68

[Label] 65.6

Video-STaR

Data flow




Finetuning
Cycle 2

Answer Generation

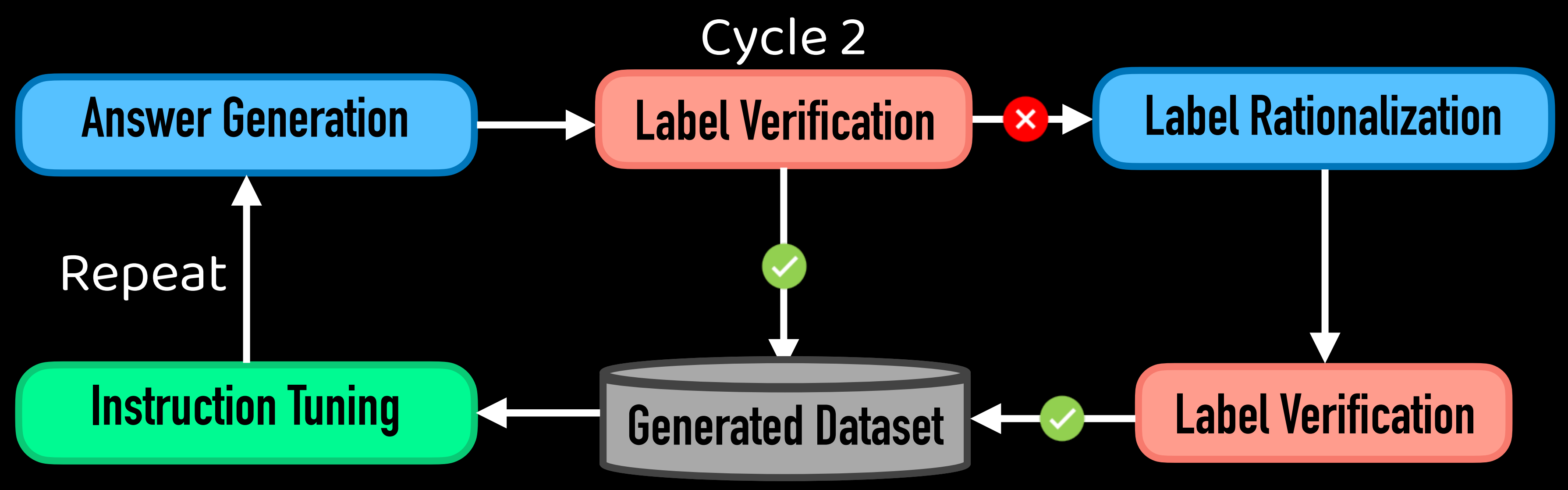
[Input Instruction] Please assess the performer's skill level and the action's toughness in the video on a scale of 1-100.. Can you explain step-by-step how one can arrive at this conclusion?

First, we need to identify the action sequence performed by the diver. In this case, the diver performs an inward somersault, followed by a 3.5 somersault tuck, and then an entry into the water... The action difficulty score for a inward to a 3.5 somersault tuck is 3.2…
we can conclude that the diver's action sequence is of a high skill level and is quite tough, with an overall score of 64.68

[Label] 65.6

Video-STaR

Data flow



Cycle 2

Answer Generation

Label Verification

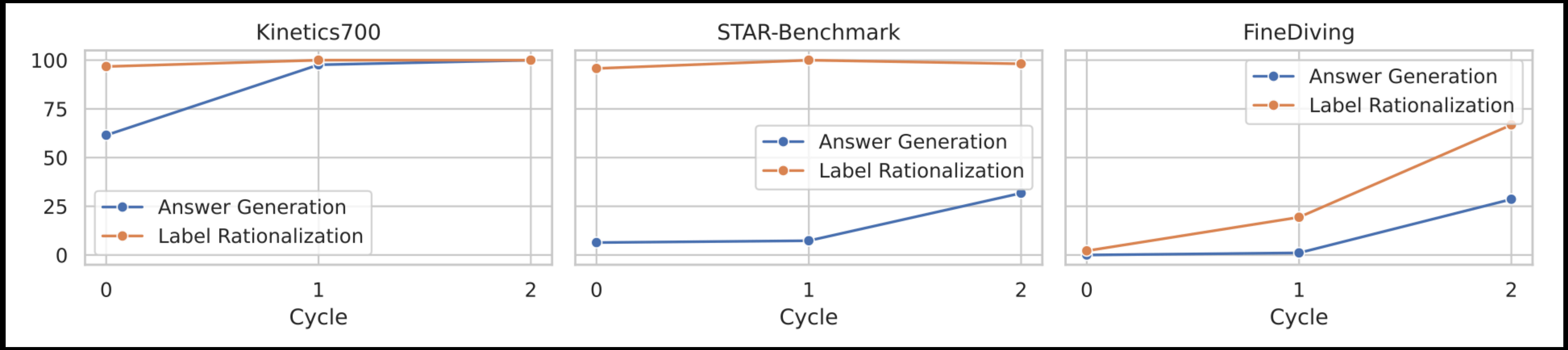
Label Rationalization

Repeat

Instruction Tuning

Generated Dataset

Label Verification

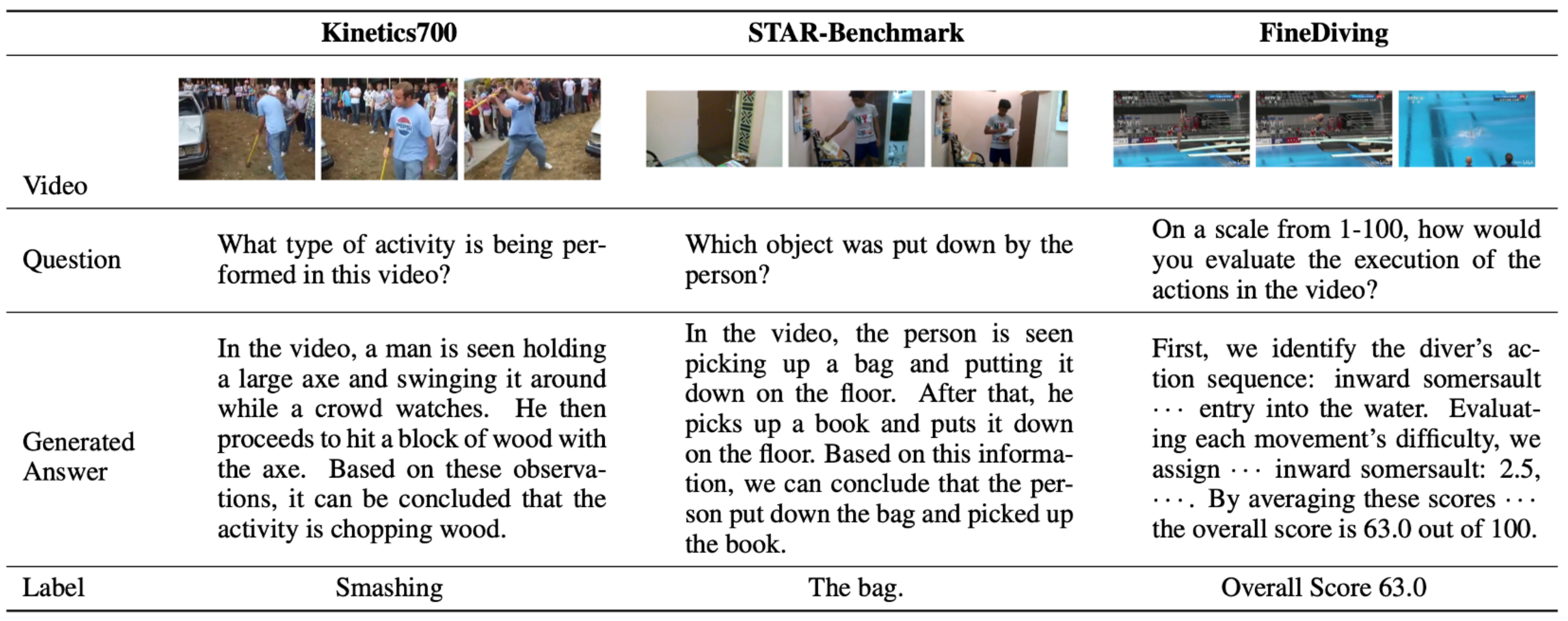


Kinetics700

STAR-Benchmark

FineDiving

Source/Generated Datasets



	Kinetics700	STAR-Benchmark	FineDiving
Video			
Question	What type of activity is being performed in this video?	Which object was put down by the person?	On a scale from 1-100, how would you evaluate the execution of the actions in the video?
Generated Answer	In the video, a man is seen holding a large axe and swinging it around while a crowd watches. He then proceeds to hit a block of wood with the axe. Based on these observations, it can be concluded that the activity is chopping wood.	In the video, the person is seen picking up a bag and putting it down on the floor. After that, he picks up a book and puts it down on the floor. Based on this information, we can conclude that the person put down the bag and picked up the book.	First, we identify the diver's action sequence: inward somersault ··· entry into the water. Evaluating each movement's difficulty, we assign ··· inward somersault: 2.5, ···. By averaging these scores ··· the overall score is 63.0 out of 100.
Label	Smashing	The bag.	Overall Score 63.0

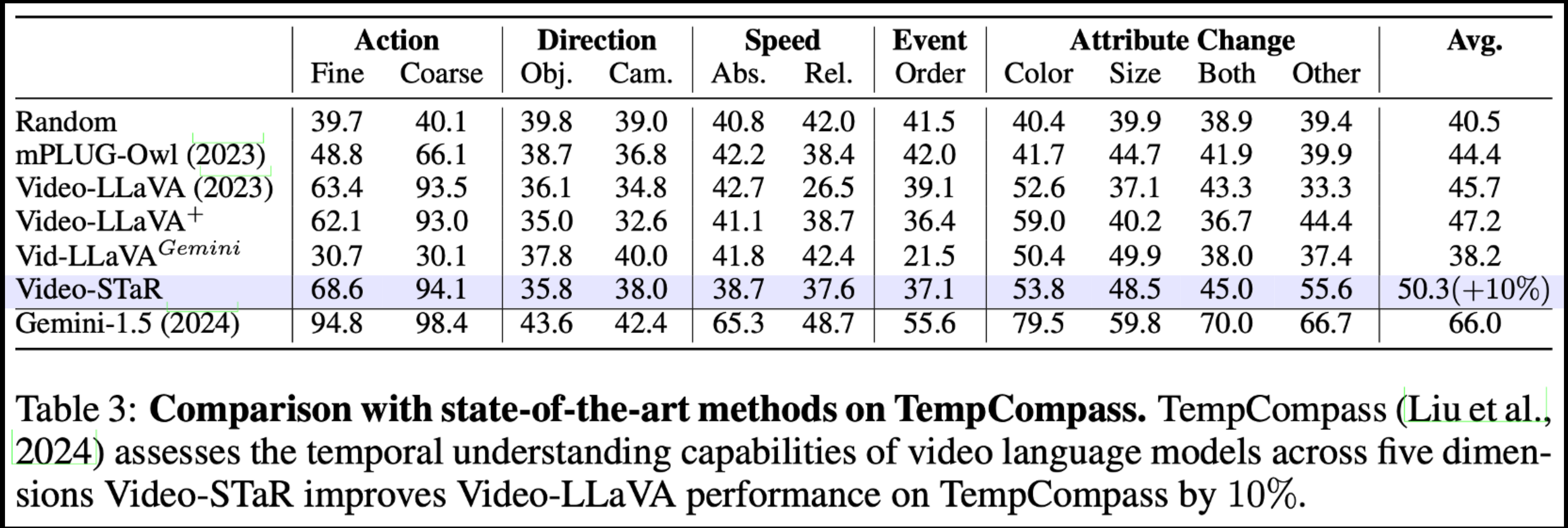
Video-STaR

Results

Baselines:

- Video-LLaVA: Source model
- Video-LLaVA+: Utilize the same datasets, and convert to QA using templates
- Video-LLaVA-Gemini: Use Gemini to label 1000 examples per dataset (distillation comparison)

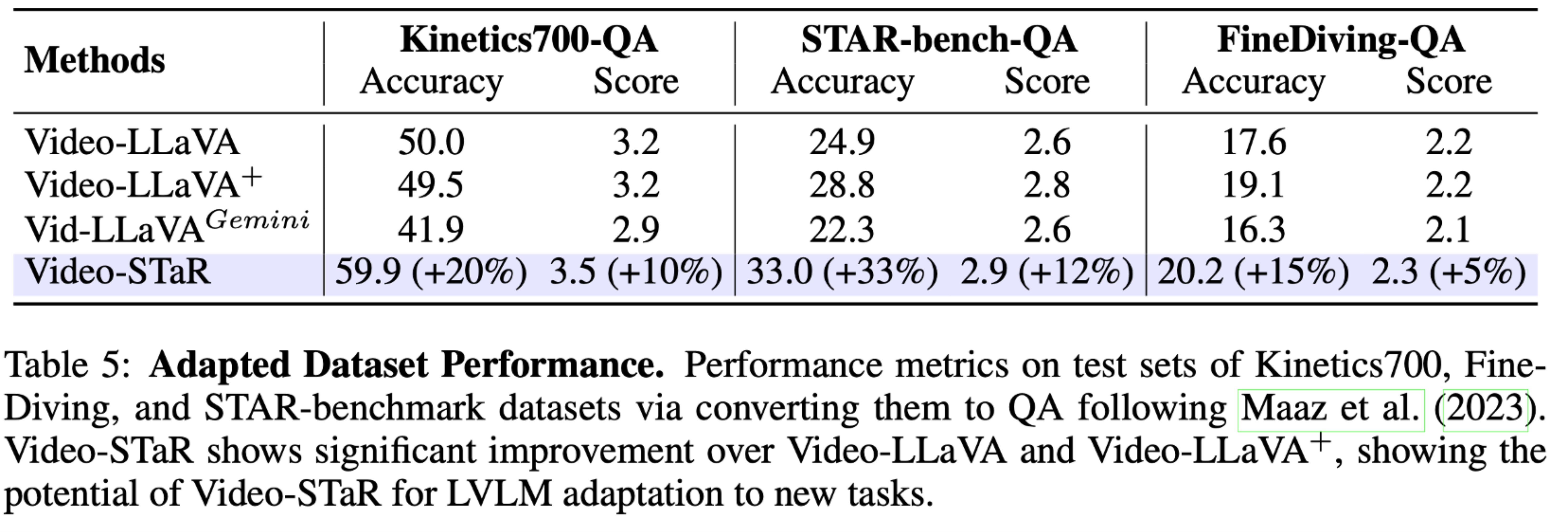
TempCompass



	Action		Direction		Speed		Event	Attribute Change				Avg.
	Fine	Coarse	Obj.	Cam.	Abs.	Rel.	Order	Color	Size	Both	Other	
Random	39.7	40.1	39.8	39.0	40.8	42.0	41.5	40.4	39.9	38.9	39.4	40.5
mPLUG-Owl (2023)	48.8	66.1	38.7	36.8	42.2	38.4	42.0	41.7	44.7	41.9	39.9	44.4
Video-LLaVA (2023)	63.4	93.5	36.1	34.8	42.7	26.5	39.1	52.6	37.1	43.3	33.3	45.7
Video-LLaVA^{+}	62.1	93.0	35.0	32.6	41.1	38.7	36.4	59.0	40.2	36.7	44.4	47.2
Vid-LLaVA^{Gemini}	30.7	30.1	37.8	40.0	41.8	42.4	21.5	50.4	49.9	38.0	37.4	38.2
Video-STaR	68.6	94.1	35.8	38.0	38.7	37.6	37.1	53.8	48.5	45.0	55.6	50.3(+10%)
Gemini-1.5 (2024)	94.8	98.4	43.6	42.4	65.3	48.7	55.6	79.5	59.8	70.0	66.7	66.0

Table 3: **Comparison with state-of-the-art methods on TempCompass.** TempCompass (Liu et al., 2024) assesses the temporal understanding capabilities of video language models across five dimensions Video-STaR improves Video-LLaVA performance on TempCompass by 10%.

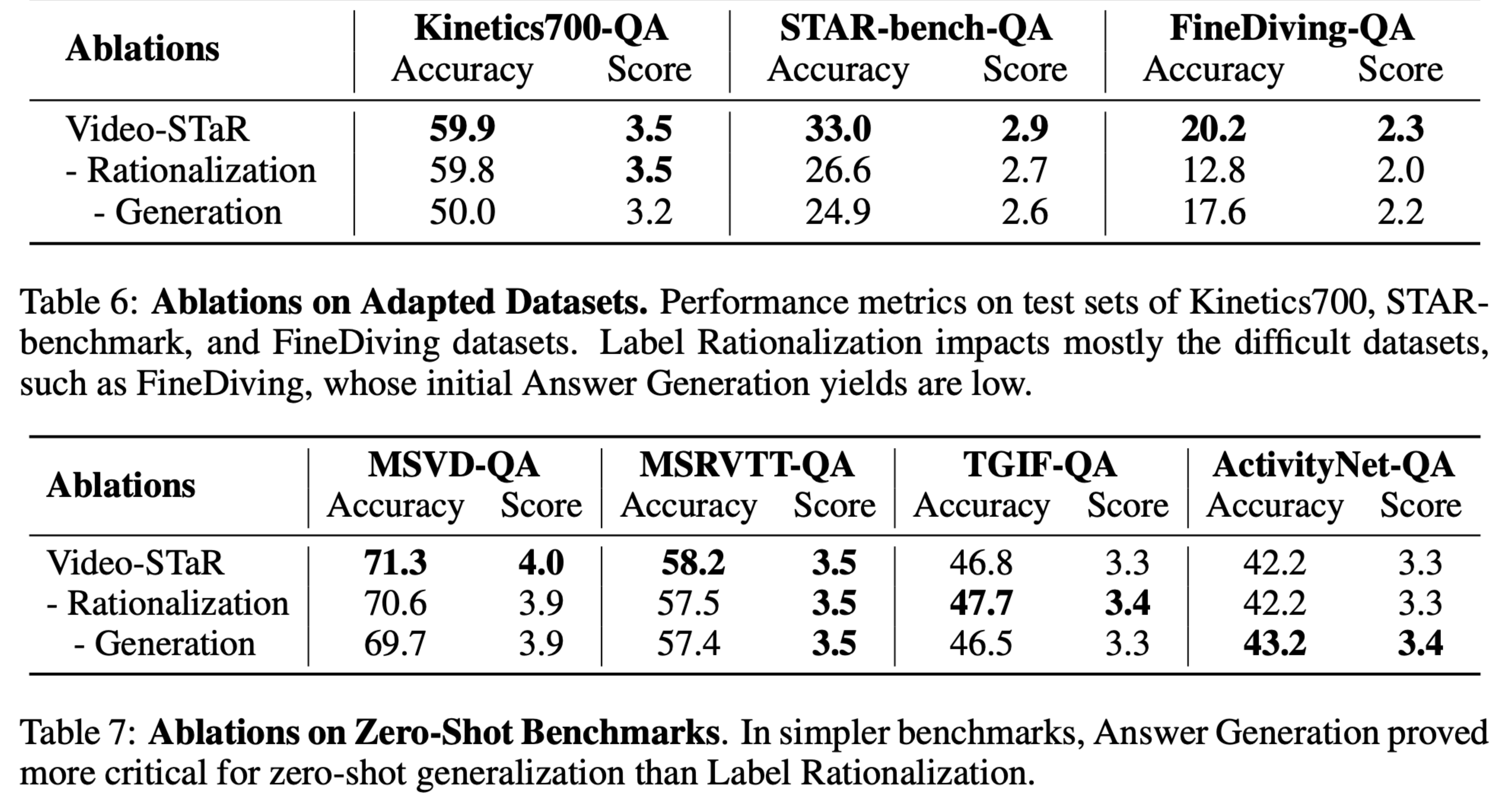
Adapted Datasets



Methods	**Kinetics700-QA**		**STAR-bench-QA**		**FineDiving-QA**	
	Accuracy	Score	Accuracy	Score	Accuracy	Score
Video-LLaVA	50.0	3.2	24.9	2.6	17.6	2.2
Video-LLaVA^{+}	49.5	3.2	28.8	2.8	19.1	2.2
Vid-LLaVA^{Gemini}	41.9	2.9	22.3	2.6	16.3	2.1
Video-STaR	59.9 (+20%)	3.5 (+10%)	33.0 (+33%)	2.9 (+12%)	20.2 (+15%)	2.3 (+5%)

Table 5: **Adapted Dataset Performance.** Performance metrics on test sets of Kinetics700, FineDiving, and STAR-benchmark datasets via converting them to QA following Maaz et al. (2023). Video-STaR shows significant improvement over Video-LLaVA and Video-LLaVA^{+}, showing the potential of Video-STaR for LVLM adaptation to new tasks.

Ablations



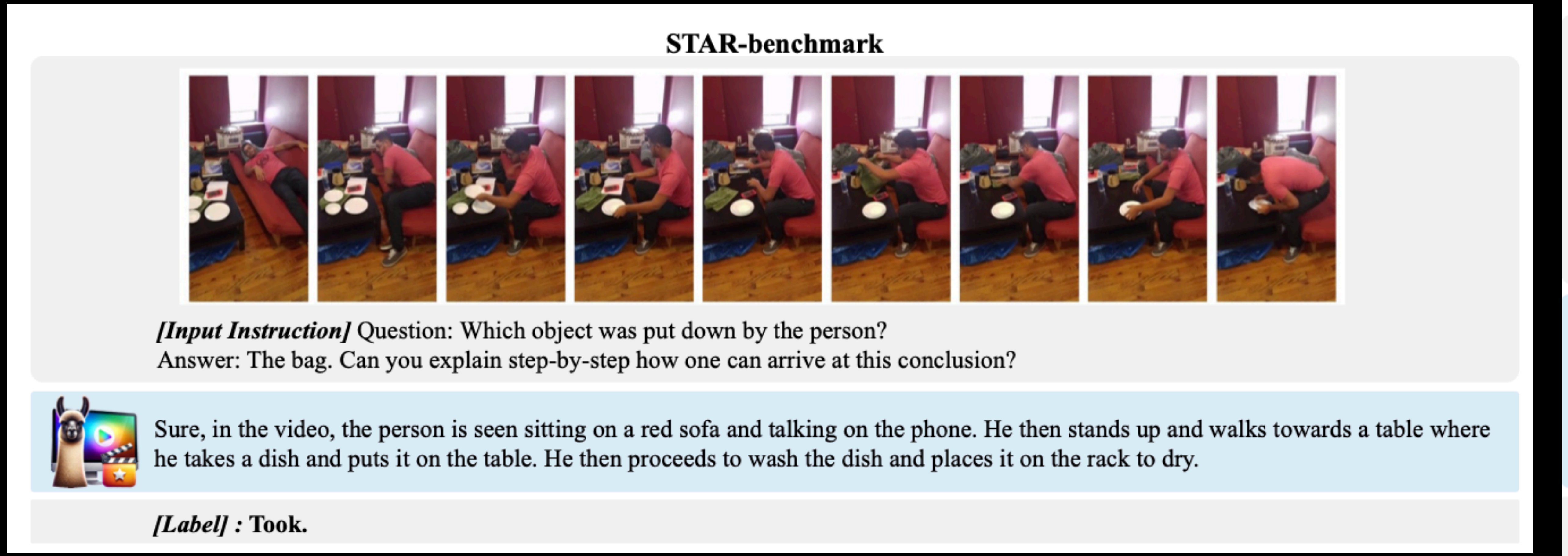
Ablations	Kinetics700-QA Accuracy	Kinetics700-QA Score	STAR-bench-QA Accuracy	STAR-bench-QA Score	FineDiving-QA Accuracy	FineDiving-QA Score
Video-STaR	**59.9**	**3.5**	**33.0**	**2.9**	**20.2**	**2.3**
- Rationalization	59.8	**3.5**	26.6	2.7	12.8	2.0
- Generation	50.0	3.2	24.9	2.6	17.6	2.2

Table 6: **Ablations on Adapted Datasets.** Performance metrics on test sets of Kinetics700, STAR-benchmark, and FineDiving datasets. Label Rationalization impacts mostly the difficult datasets, such as FineDiving, whose initial Answer Generation yields are low.

Ablations	MSVD-QA Accuracy	MSVD-QA Score	MSRVTT-QA Accuracy	MSRVTT-QA Score	TGIF-QA Accuracy	TGIF-QA Score	ActivityNet-QA Accuracy	ActivityNet-QA Score
Video-STaR	**71.3**	**4.0**	**58.2**	**3.5**	46.8	3.3	42.2	3.3
- Rationalization	70.6	3.9	57.5	**3.5**	**47.7**	**3.4**	42.2	3.3
- Generation	69.7	3.9	57.4	**3.5**	46.5	3.3	**43.2**	**3.4**

Table 7: **Ablations on Zero-Shot Benchmarks**. In simpler benchmarks, Answer Generation proved more critical for zero-shot generalization than Label Rationalization.

Failure Cases: Hallucination

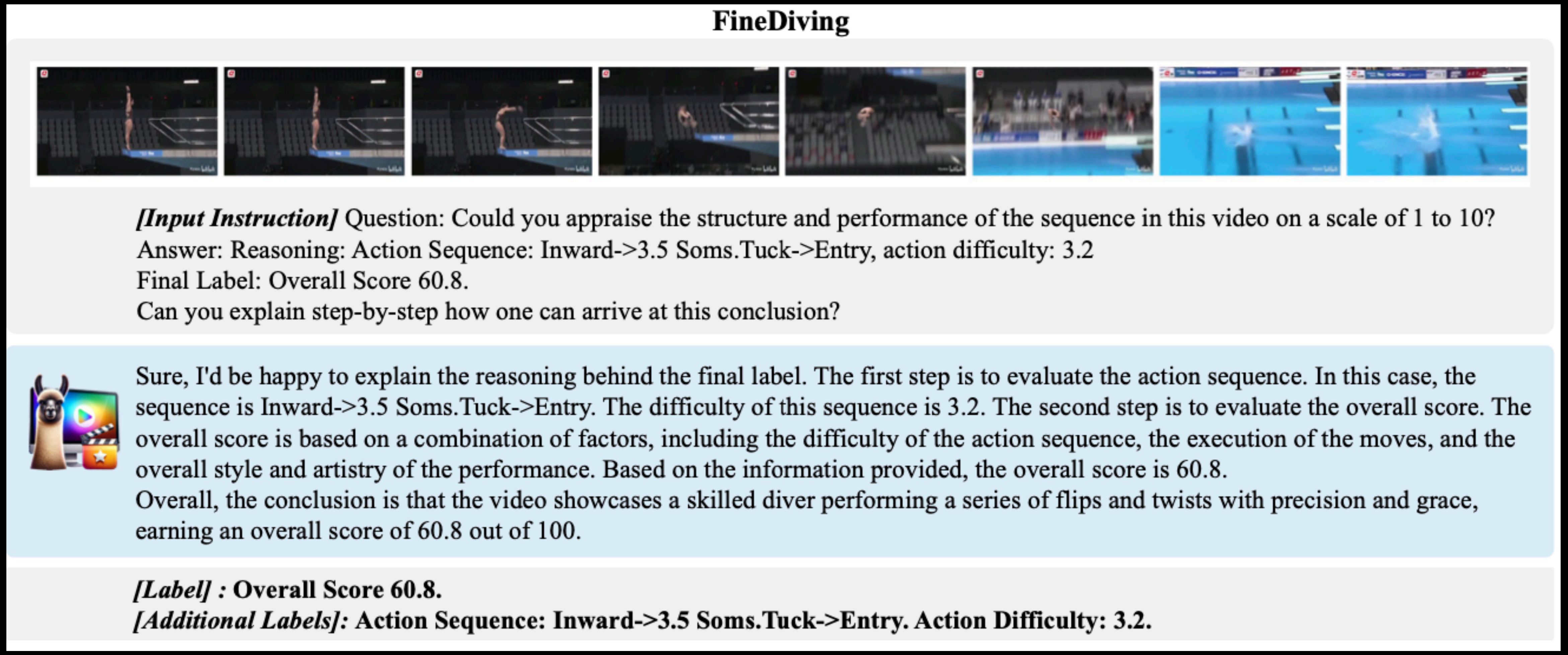


STAR-benchmark

[Input Instruction] Question: Which object was put down by the person?
Answer: The bag. Can you explain step-by-step how one can arrive at this conclusion?

Sure, in the video, the person is seen sitting on a red sofa and talking on the phone. He then stands up and walks towards a table where he takes a dish and puts it on the table. He then proceeds to wash the dish and places it on the rack to dry.

[Label] : **Took.**



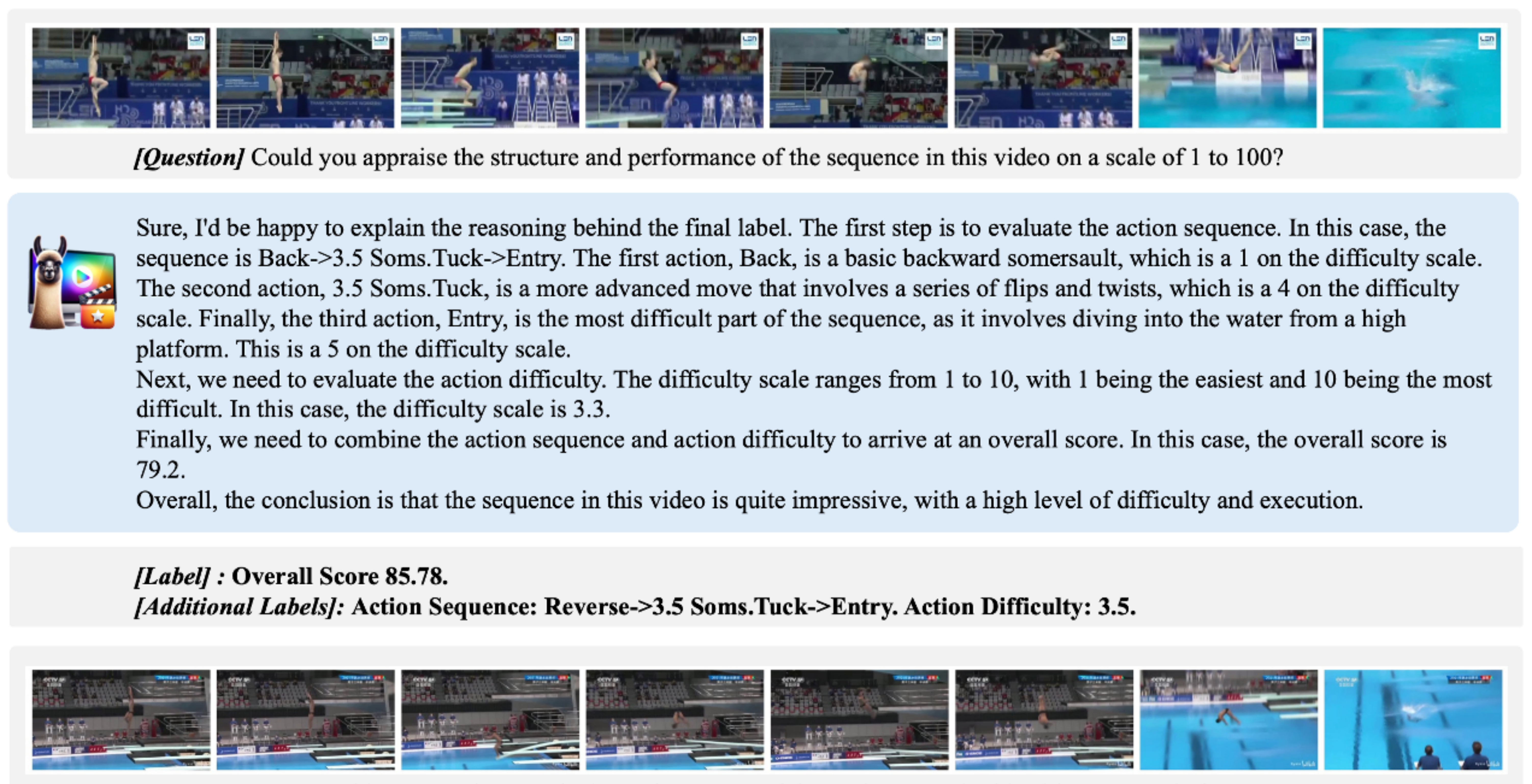
FineDiving

[Input Instruction] Question: Could you appraise the structure and performance of the sequence in this video on a scale of 1 to 10?
Answer: Reasoning: Action Sequence: Inward->3.5 Soms.Tuck->Entry, action difficulty: 3.2
Final Label: Overall Score 60.8.
Can you explain step-by-step how one can arrive at this conclusion?

Sure, I'd be happy to explain the reasoning behind the final label. The first step is to evaluate the action sequence. In this case, the sequence is Inward->3.5 Soms.Tuck->Entry. The difficulty of this sequence is 3.2. The second step is to evaluate the overall score. The overall score is based on a combination of factors, including the difficulty of the action sequence, the execution of the moves, and the overall style and artistry of the performance. Based on the information provided, the overall score is 60.8.
Overall, the conclusion is that the video showcases a skilled diver performing a series of flips and twists with precision and grace, earning an overall score of 60.8 out of 100.

[Label] : **Overall Score 60.8.**
[Additional Labels]: **Action Sequence: Inward->3.5 Soms.Tuck->Entry. Action Difficulty: 3.2.**

LVLMs as a judge in Olympic events?



[Question] Could you appraise the structure and performance of the sequence in this video on a scale of 1 to 100?

Sure, I'd be happy to explain the reasoning behind the final label. The first step is to evaluate the action sequence. In this case, the sequence is Back->3.5 Soms.Tuck->Entry. The first action, Back, is a basic backward somersault, which is a 1 on the difficulty scale. The second action, 3.5 Soms.Tuck, is a more advanced move that involves a series of flips and twists, which is a 4 on the difficulty scale. Finally, the third action, Entry, is the most difficult part of the sequence, as it involves diving into the water from a high platform. This is a 5 on the difficulty scale.

Next, we need to evaluate the action difficulty. The difficulty scale ranges from 1 to 10, with 1 being the easiest and 10 being the most difficult. In this case, the difficulty scale is 3.3.

Finally, we need to combine the action sequence and action difficulty to arrive at an overall score. In this case, the overall score is 79.2.

Overall, the conclusion is that the sequence in this video is quite impressive, with a high level of difficulty and execution.

***[Label] :* Overall Score 85.78.**

***[Additional Labels]:* Action Sequence: Reverse->3.5 Soms.Tuck->Entry. Action Difficulty: 3.5.**

[Question] How would you rate the complexity and accuracy of the performance in this video on a scale from 1 to 100?

Thanks for tuning in!

https://orrzohar.github.io/projects/video-star/

SCAN ME

Orr Zohar

Xiaohan Wang

Yonatan Bitton

Idan Szpektor

Serena Yeung-Levy