

VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment

Supplementary Material

In this supplementary material, we discuss

1. Additional results and analysis, both quantitative and qualitative (Appendix A);
2. Benchmark creation, quality control process, and some statistics (Appendix B);
3. Model prompts used in both setups: entailment and multiple-choice (Appendix C); and
4. Limitations (Appendix D).

A. Additional Results

In Appendix A.1, we present scatter plots of entailment scores for all models across all tests, expanding Fig. 3 from the main paper. Next, we present the positive and negative entailment scores that are used in StrictVLE (expanding the analysis in Sec. 5.2) in Appendix A.2. We experiment with Chain-of-Thought prompting in Appendix A.3 and present ablations for number of sampled frames in Appendix A.4. The multiple-choice (MC) evaluation results are discussed in Appendix A.5 and the human evaluation setup in Appendix A.6. Finally, we share some qualitative results of LLaVA-OneVision-72B on our benchmark in Appendix A.7.

A.1. Scatter plot of entailment scores

To analyze entailment scores, we present scatter plots for all models on the benchmark subset (150 samples) in Fig. 4. The ideal scenario is when all samples lie in the bottom-right quadrant (points in dark green, quadrant in light yellow), which indicates that the model confidently entails the correct caption while rejecting the negative caption, leading to a 100% StrictVLE accuracy. However, in practice, we observe two undesirable cases: (i) the points are concentrated in the top-right quadrant, indicating a strong bias towards responding ‘Yes’ regardless of whether the caption is aligned or misaligned; and (ii) the points are clustered around the diagonal, indicating that the model exhibits similar confidence levels when saying ‘Yes’ to both the positive and negative captions. Major takeaways are highlighted below:

- P-LLaVA has most of its points concentrated in the top-right quadrant, indicating a strong bias towards responding ‘Yes’ regardless of whether the caption is positive or negative, which also explains its near 0% StrictVLE accuracy.
- Owl-Con and Video-LLaVA are strongly clumped near the diagonal in the top-right quadrant (except for the Control Test): indicating that they tend to respond ‘Yes’ and have similar entailment scores for both the positive and negative captions. Owl-Con appears to be worse than Video-LLaVA with more points in the top-right quadrant.
- Between LLaVA-OneVision-7B (OV-7B) and LLaVA-OneVision-7B-Si (OV-7B-SI), we see that the points in OV-7B-SI are more clustered near the diagonal while LLaVA-OneVision-7B is more diffused except for AgCref. This is expected as it is hard for a model trained on single images to distinguish between the positive and the negative caption and nearly impossible for the EvChr and AgCref. In contrast, both models perform well on the Control Test since the replacements come from a totally different or random video, making it easier for the models to classify with sufficient confidence.
- For Qwen2-VL-7B (QVL-7B) except for Control Test, the points for all the other tests are concentrated in the top-right corner while additionally being clustered near the diagonal for the EvChr. QVL-7B performs worse than OV-7B even though both models are trained using the same base Qwen2 7B parameter LLM.
- Finally, on LLaVA-OneVision-72B, we see that many points are below the diagonal and would score correct on ClassicVLE. However, roughly half of them (on average) are in the bottom right quadrant indicating difficulty of the best model to predict ‘Yes’ for the positive caption and ‘No’ for the negative caption respectively.

A.2. Analyzing Entailment Scores for StrictVLE

Continuing from findings of Sec. 5.2 in the main paper, we analyze whether a model finds it easier to classify C^+ or C^- in Tab. 7 for all tests. Each cell in the table reports two numbers: the first is the accuracy of positive captions, and the second is the accuracy of negative captions when the positive caption is correct.

An interesting observation (as also noted in the main paper) is that as model size increases, the positive caption accuracy decreases while the negative caption accuracy improves. This holds for both variants: OV-7B to OV-72B and QVL-7B to QVL-72B, and indicates that small models are eager to say ‘Yes’ for both captions, while larger models reason better.

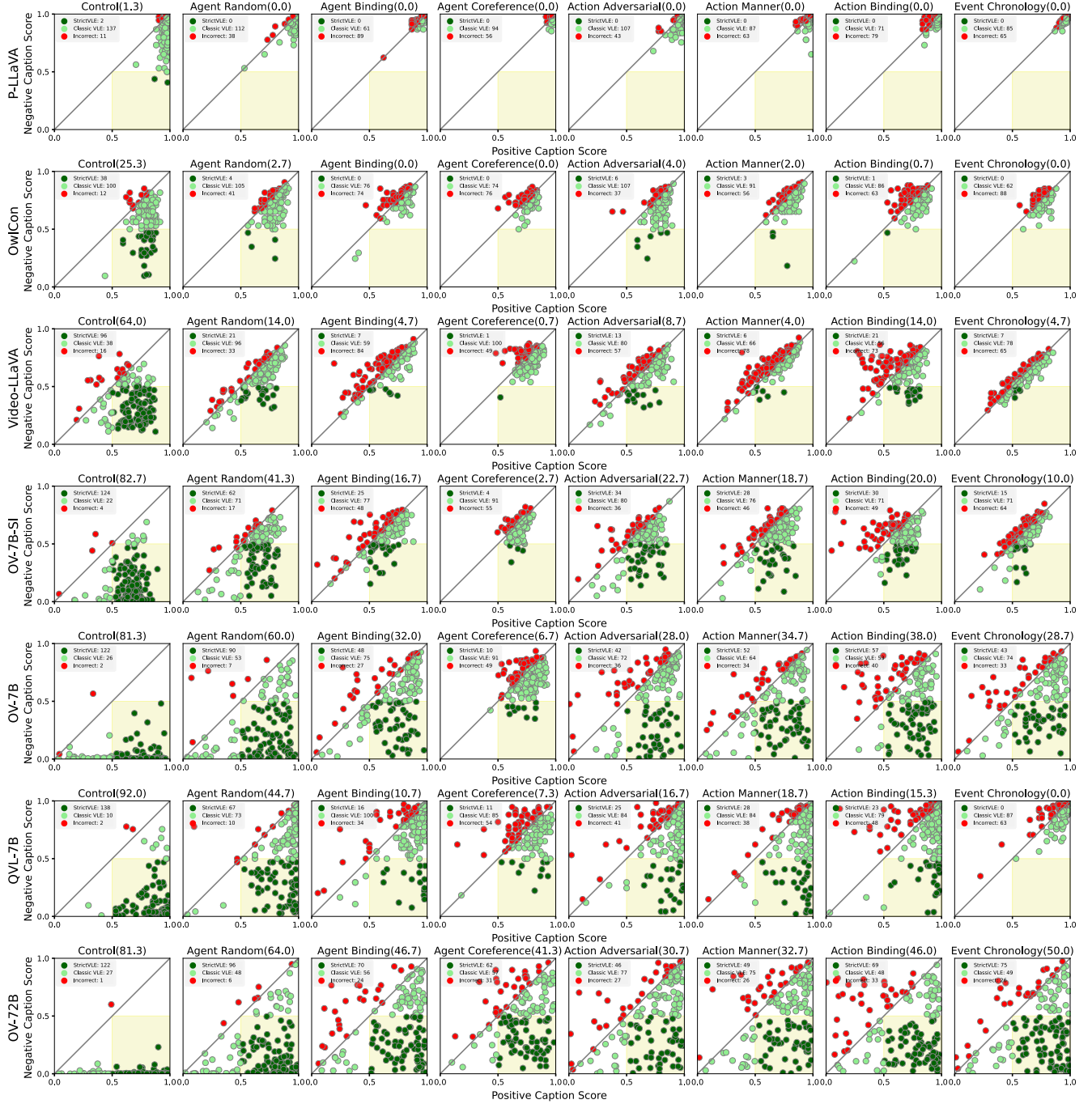


Figure 4. Scatter plot of entailment scores $e(V, C^+)$ (x-axis) and $e(V, C^-)$ (y-axis) for all tests in VELOCITY. We visualize the scores for several models indicated in the left margin. From top to bottom: P-LLaVA, OwlCon, Video-LLaVA, OV-7B-SI, OV-7B, QVL-7B, and OV-72B. ClassicVLE calls a sample correct in the region below the diagonal (light green). Instead, StrictVLE requires the dots to lie in the yellow bottom-right quadrant (dark green). Finally, samples whose points are above the diagonal are wrong for both VLE metrics (red). The legend includes the actual number of points (please zoom in). This figure is best seen in color.

Although Qwen2-VL (QVL) models achieve higher accuracy for positive captions than LLaVA-OneVision (OV) models, the negative caption accuracy is better for OV models. This indicates that QVL models are biased to say ‘Yes’ regardless of the captions, whereas OV models reason better and are less inclined to respond ‘Yes’. For QVL, specifically for the EvChr test, the positive caption accuracy is very high, but the negative caption accuracy is extremely low, indicating that the QVL

Model	Ctrl	Ag Rand	Ag Bind	Ag Cref	Act Adv	Act Man	Act Bind	Ev Chrono	Avg
OV-7B	83.0 / 98.3	84.2 / 67.3	82.0 / 40.1	97.1 / 8.2	86.3 / 34.4	80.6 / 37.9	81.9 / 44.5	89.0 / 34.2	85.9 / 38.1
OV-72B	79.8 / 99.3	81.6 / 78.1	79.5 / 57.1	87.0 / 44.4	80.6 / 41.1	77.9 / 37.5	78.2 / 57.6	78.3 / 59.4	80.4 / 53.6
QVL-7B	92.4 / 91.5	92.8 / 42.1	90.7 / 14.9	97.6 / 6.6	94.3 / 18.9	93.0 / 18.8	91.0 / 18.1	98.0 / 0.4	93.9 / 17.1
VELOCITI Subset									
QVL-72B	84.0 / 98.4	85.3 / 65.6	84.0 / 34.9	77.3 / 45.7	84.0 / 35.7	86.7 / 27.7	80.7 / 43.8	98.7 / 1.4	85.2 / 36.4
OV-72B	81.3 / 100.	79.3 / 80.7	80.7 / 57.9	86.7 / 47.7	79.3 / 38.7	82.0 / 39.8	74.7 / 61.6	80.7 / 62.0	80.5 / 55.5
Gem-1.5F	93.9 / 97.8	93.3 / 60.4	93.9 / 25.4	100. / 4.7	93.3 / 35.3	95.3 / 22.7	95.3 / 26.2	99.3 / 2.8	95.8 / 25.4
Gem-1.5P	75.0 / 99.1	72.3 / 83.2	75.5 / 65.8	71.3 / 51.4	72.5 / 72.2	79.6 / 54.7	75.5 / 65.8	71.4 / 70.5	74.0 / 66.2
GPT-4o	63.3 / 100.0	58.0 / 94.3	64.0 / 69.8	62.0 / 65.6	65.8 / 83.7	65.3 / 64.3	64.7 / 83.5	71.8 / 44.9	64.5 / 72.3

Table 7. StrictVLE Analysis for various models on all tests in VELOCITI. Each cell of the table has two numbers. The first is the fraction of correctly classified positive captions. The second is the fraction of correctly classified negative captions, among samples whose positive caption is classified correctly. Refer to Appendix A.2 for a description.

Model	Ag Rand	Ag Bind	Ag Cref	Act Adv	Act Man	Act Bind	Ev Chr	Avg
OV-72B								
w/o CoT	64.0	46.7	41.3	30.7	32.7	46.0	50.0	44.5
CoT	40.0	28.0	19.3	26.7	28.0	32.0	30.0	29.1
Gemini-1.5 Pro								
w/o CoT	60.1	49.7	36.7	52.3	43.5	52.3	50.3	49.3
CoT	46.6	32.8	45.5	46.2	46.8	46.4	29.2	41.9

Table 8. Average score on VELOCITI subset: without and with CoT

models are very poor at the temporal order reasoning.

While GPT-4o achieves comparatively lower accuracy on positive captions across all tests, it consistently achieves the highest accuracy for negative captions, except in the EvChr test, where OV-72B performs best. Another surprising observation is that Gemini-1.5-Flash (Gem-1.5F), despite achieving the best accuracy for positive captions, performs worse than all other models for negative captions. This suggests that Gemini-1.5-Flash may also be responding with ‘Yes’ too often. Additionally, both Gemini-1.5-Flash and Qwen2-VL-72B exhibit very low accuracy for negative captions in the AgCref and EvChr tests.

Finally, in Sec. 5.1 of the main paper, we highlight that $AgRand > AgBind > AgCref$ – this trend is clearly observed in the negative caption accuracies presented in the Table, and explains the poor performance of some models on Agent Coreference Test (single-digit accuracies on negative captions).

A.3. Impact of Chain-of-Thought prompting

We experimented with Chain-of-Thought (CoT) prompting for Gemini-1.5-Pro and OV-72B (prompt in Fig. 5). As shown in Tab. 8, the performance reduced in both cases indicating that models are unable to reason in a step-by-step manner for such statements.

A.4. Impact of Increased Frame Rate

We explore increasing the video sampling rate to observe if more visual information aids the model to solve the tasks in VELOCITI to a greater extent. For this, we sample frames at 8 fps, amounting to 80 frames for a 10 s video. From Tab. 9 we observe that the smaller model (OV-7B) benefits with more frames resulting in improvement across most tests, an average of +3.5%. Interestingly, its larger counterpart (OV-72B) performs worse with significant drops on action tests, ActAdv and ActMan, (9-11%). This may be due to the large context size that the model is not trained for. Both models perform better on EvChr task.

A.5. Multiple-Choice (MC) Evaluation: Results on each test

In the MC setup, we provide the video along with both captions to the Video-LLM and ask it to pick the correct one (A or B). Results on the control and average over the benchmark were discussed in Sec. 5.4 of the main paper.

System Prompt

You are an AI assistant specializing in analyzing movie clips to verify captions using a Chain of Thought (CoT) approach. Given a movie clip and a corresponding caption, your task is to determine whether the caption accurately describes the events in the clip.

A caption is considered accurate (“Yes”) if **all applicable** of the following criteria are met:

1. ****Actor/Doer****: The person or entity performing the action is correctly identified.
2. ****Attributes****: The characteristics of the actors/doers and the action itself are accurately described (*e.g.* clothing color, size, speed).
3. ****Instruments/Objects****: Any tools, objects, or instruments used in the action are correctly identified.
4. ****Receiver/Patient****: The target or recipient of the action is correctly identified.
5. ****Relationships****: The relationships between the entities involved (*e.g.*, “standing next to”, “holding”) are accurately depicted.
6. ****Manner****: The way in which the action is performed (*e.g.*, “quickly”, “slowly”, “angrily”) is accurately described.
7. ****Location****: The setting or location of the scene is correctly identified.
8. ****Clarity****: There is sufficient visual information in the clip to confidently assess the correctness of the caption.
9. ****Event Order****: If the caption suggests a specific order of events, then the video should have events happening in the suggested order.

If any of the above criteria cannot be verified due to a lack of visuals, the caption should not be considered accurate.

Note that the caption is designed to represent a part of the video clip and may not explain all the events in the clip.

Follow these steps:

1. ****Analysis****: Carefully examine the provided movie clip.
2. ****Reasoning****: Analyze the caption in relation to the clip. Break down the caption into smaller parts and determine if each part meets the accuracy criteria listed above. Detail your reasoning process within ‘<thinking>’ tags.
3. ****Evaluation****: Based on your reasoning, evaluate the overall accuracy of the caption. If there is insufficient information in the clip to definitively confirm or deny the caption based on one or more criteria, explain what information is missing within ‘<reflection>’ tags.
4. ****Conclusion****: Provide a clear “Yes” or “No” answer within ‘<output>’ tags.

Use the following format:

<thinking>

[Detailed step-by-step reasoning, referencing the accuracy criteria. This is your internal thought process.]

</thinking>

<reflection>

[Reflections on your reasoning, including any uncertainties or missing information and which criteria could not be verified. If the caption cannot be definitively verified, explain why.]

</reflection>

<output>

[Yes or No]

</output>

Evaluate the following caption for the accompanying movie clip: {caption}

Figure 5. CoT evaluation prompt.

Now, we report results across all the tests in Tab. 10. For both OV and QVL models, we see that the smaller variants have a higher choice bias and tend to prefer option B. While this bias reduces in the larger variants, it is still high. Also, as expected, the accuracy of A∧B improves for larger variants. We observe that harder tests (*e.g.* AgBind vs. AgRand) tend to have a higher bias. Among all the tests, the EvChr test has the highest bias and the lowest accuracy across all the models.

Both Gemini-1.5-Flash and GPT-4o show considerable bias. Interestingly, GPT-4o seems to prefer option A, while Gem-1.5F prefers option B.

A.6. Human Evaluation

Human evaluations were conducted in a standardized manner to establish human performance in the various tasks presented in VELOCITI. The evaluations included 3 volunteers who were assigned the subset (150 samples for each of the 7 tests). This

Model	Ag Rand	Ag Bind	Ag Cref	Act Adv	Act Man	Act Bind	Ev Chr	Avg
OV-7B								
1fps	56.7	32.9	8.0	29.7	30.6	36.4	30.5	32.1
8fps	59.3	34.7	6.0	34.7	38.3	33.3	42.0	35.6
OV-72B								
1fps	64.7	46.0	36.7	42.0	40.7	46.0	46.0	46.0
8fps	63.7	45.4	38.6	33.1	29.3	45.1	46.5	43.1

Table 9. Higher frame rate sampling results.

Model	AgRand		AgBind		AgCref		ActAdv		ActMan		ActBind		EvChr	
	Bias	A^B	Bias	A^B	Bias	A^B	Bias	A^B	Bias	A^B	Bias	A^B	Bias	A^B
QVL-7B	24.2	74.1	42.2	40.8	37.5	33.6	49.6	42.9	51.5	41.5	40.0	36.1	98.5	0.7
OV-7B	41.6	58.1	81.6	17.1	59.9	26.0	71.3	27.6	68.0	30.1	70.0	24.2	81.9	15.9
OV-72B	3.1	94.8	10.9	72.9	8.0	69.0	8.9	79.7	11.1	77.5	14.8	62.5	15.1	75.9
VELOCITI Subset														
QVL-72B	6.0	88.7	3.3	64.7	2.7	60.0	6.7	68.0	2.0	74.0	8.6	54.7	-11.3	47.3
OV-72B	2.7	95.3	11.4	75.3	11.3	67.3	9.3	76.7	7.3	77.3	16.0	61.3	16.7	72.0
Gem-1.5F	-2.8	94.4	-12.6	73.4	8.0	61.3	-12.9	72.8	-14.3	66.0	0.7	64.8	-49.6	41.4
GPT-4o	4.1	92.5	4.7	79.3	3.4	60.8	-10.1	77.0	-7.0	74.1	2.0	70.7	-60.8	25.7

Table 10. MC evaluation results on all tests. Along with the video, we provide the model with both captions A and B and ask it to pick the better-aligned one. Bias is the accuracy difference between B and A options and should be close to 0. A^B involves evaluating the model twice, once with the correct caption as A and again as B. A sample is deemed correct when it picks the correct choice in both cases. While a model’s decision should be unaffected by the order in which choices are presented, a considerable bias is observed.

amounts to a total of 2,100 video-caption pairs (7 tests \times 150 samples \times 2 captions). We use the Label Studio [43] annotation platform for this task. To ensure fair evaluations, humans are first shown a set of instructions to ensure consistency across participants. Next, we randomize and present non-overlapping video-caption pairs. An example of the annotation dashboard is shown in Fig. 6.

A.7. Qualitative Analysis

We present examples from the OV-72B model on our benchmark for three following cases: (i) Samples satisfying the StrictVLE criteria ($e(V, C^+) > 0.5 \wedge e(V, C^-) < 0.5$) are shown in Fig. 7; (ii) Samples *only* satisfying the ClassicVLE condition ($e(V, C^+) > e(V, C^-)$), but failing on the StrictVLE condition are in Fig. 8. (iii) Finally, samples classified incorrectly according to ClassicVLE ($e(V, C^+) < e(V, C^-)$), are presented in Fig. 9. Note these are also incorrect for StrictVLE. In each case, we show 10 frames from the video, the positive and negative captions, and the corresponding entailment scores. The test name is indicated in the bottom left.

Instructions

These instructions can be opened anytime by clicking 'i' on the bottom left of the panel.

You are given a video and a caption for each task.


Please watch the 10s video and select 'Yes' if the given video entails the caption, otherwise select 'No'

- The caption should provide an accurate description of the events in the video.
- The caption should correctly identify the entities (*humans, animals, objects*, etc.) and the relationships (*actions*) between them.

Note

- Ignore any spelling/grammatical errors, if any.
- You may watch the video multiple times, if needed.

Video 1



1 of 241

Here is a caption that describes the video

The woman with black hair is inserting a tongue suppressor into the girl with brown hair's mouth.

Does the given video entail the caption?

☐ Yes^[1] ☐ No^[2]

Binary-Choice, Single Select Option

Revisit Annotation Instructions

Skip Submit

Figure 6. Human Evaluation Dashboard. Instructions and interface for human evaluation for the entailment task.

















 <p>Positive Caption</p>		 <p>Negative Caption</p>	
<p>At a police station, a policeman is hoisting a man in a green jacket up</p>		<p>Outside a home, a dog with light fur is running straight between two houses</p>	
Control		$e(V, C^+) = 0.944$	$e(V, C^-) = 0.000$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Agent Random		$e(V, C^+) = 0.841$	$e(V, C^-) = 0.468$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Agent Binding		$e(V, C^+) = 0.603$	$e(V, C^-) = 0.055$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Agent Coreference		$e(V, C^+) = 0.640$	$e(V, C^-) = 0.268$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Action Adversarial		$e(V, C^+) = 0.885$	$e(V, C^-) = 0.075$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Action Manner		$e(V, C^+) = 0.743$	$e(V, C^-) = 0.124$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Action Binding		$e(V, C^+) = 0.987$	$e(V, C^-) = 0.060$
 <p>Positive Caption</p>		 <p>Negative Caption</p>	
Event Chronology		$e(V, C^+) = 0.798$	$e(V, C^-) = 0.480$

Figure 7. VELOCITI samples where OV-72B classifies the sample correctly based on the StrictVLE criteria. In the [Agent Binding](#) example, the scene visualizes a man and a woman talking on the phone while the man drives, C^- changes the entity of the driver. The model is confidently able to identify that it is the man who is driving and not the woman, as the positive caption scores (0.603) much above the negative caption (0.055) while satisfying the StrictVLE criteria. Similarly, in [Agent Coreference](#), the scene describes two women - a woman in blue who's sitting and puts on her headphones as she begins to write, while the woman in white looks at her and eventually walks away. The C^- interchanges the roles of these two women, and the model correctly scores the positive caption (0.640) higher than the negative caption (0.268).

















	
Positive Caption	Negative Caption
In an apartment, a woman bends down to speak to a duck	At the side of the house, a woman knocks down items on the counter with her hand
Control	
$e(V, C^+) = 0.734$	$e(V, C^-) = 0.527$
	
Positive Caption	Negative Caption
A blonde woman is holding a painting out on a wall to inspect it	On a wall, the man in a jumpsuit is holding a painting out to inspect it
Agent Random	
$e(V, C^+) = 0.637$	$e(V, C^-) = 0.519$
	
Positive Caption	Negative Caption
Inside a car, a man in a red flannel shirt is coughing	Inside the car, the man in the green flannel shirt is coughing
Agent Binding	
$e(V, C^+) = 0.911$	$e(V, C^-) = 0.5$
	
Positive Caption	Negative Caption
The person who is screaming in a scared manner is also the one who is backing away	The person who is screaming in a scared manner is also the one who is wielding a knife
Agent Coreference	
$e(V, C^+) = 0.861$	$e(V, C^-) = 0.661$
	
Positive Caption	Negative Caption
In the kitchen, the man is straightening out the silverware with his hand	In a kitchen, a man is dropping the silverware
Action Adversarial	
$e(V, C^+) = 0.679$	$e(V, C^-) = 0.569$
	
Positive Caption	Negative Caption
In the kitchen, a man in a blue shirt is pushing a man in a black jacket back angrily.	In the kitchen, the man in a blue shirt gently pushes the man in a black jacket back
Action Manner	
$e(V, C^+) = 0.969$	$e(V, C^-) = 0.721$
	
Positive Caption	Negative Caption
In a restaurant, the man in a black plaid jacket is cowering.	In a restaurant, a man in a black plaid jacket is talking aggressively to the people.
Action Binding	
$e(V, C^+) = 0.870$	$e(V, C^-) = 0.531$
	
Positive Caption	Negative Caption
First, In a bar, a ghost wearing glasses is chugging alcohol from a glass. Then, In a bar, the ghost in glasses is removing his hat from his head	First, In a bar, the ghost in glasses is removing his hat from his head. Then, In a bar, a ghost wearing glasses is chugging alcohol from a glass
Event Chronology	
$e(V, C^+) = 0.845$	$e(V, C^-) = 0.644$

Figure 8. VELOCITI samples where OV-72B classifies the sample correctly based on the ClassicVLE criteria, but not on StrictVLE. In the **Action Binding** example, a man in a black plaid jacket is cowering. The negative caption (C^-) changes the action from “cowering” to “talking aggressively.” Although the model assigns a high entailment score of 0.870 to the positive caption (C^+), it also assigns a relatively high score of 0.531 to the negative caption (C^-). While this satisfies the ClassicVLE criterion, it fails to meet the StrictVLE criterion.

















	
Outside a building, a woman is reaching down with her hand to grab a bag	In a warehouse, the boy is approaching the man laying on the bed with caution
Control	
$e(V, C^+) = 0.144$	$e(V, C^-) = 0.162$
	
Outside a mansion, a woman opens the door to see who's there	Outside the mansion, the bald man is opening the door to see who's there
Agent Random	
$e(V, C^+) = 0.091$	$e(V, C^-) = 0.668$
	
Inside a restaurant, the woman in a black shirt exits the kitchen	Inside the restaurant, the man in a brown jacket is exiting the kitchen
Agent Binding	
$e(V, C^+) = 0.746$	$e(V, C^-) = 0.779$
	
The person who is aiding the man in a yellow shirt is also the one who is removing his stethoscope from his ears	The person who is aiding the man in a yellow shirt is also the one who is covering his mouth, looking distraught
Agent Coreference	
$e(V, C^+) = 0.503$	$e(V, C^-) = 0.554$
	
In a bedroom, the dog is walking forward to get out from under the covers	In a bedroom, the dog is settling into the covers
Action Adversarial	
$e(V, C^+) = 0.787$	$e(V, C^-) = 0.933$
	
At the back seat of a car, the blonde woman is picking up a bag with her left hand	The blonde woman picks up the bag from the back seat of the car with both hands
Action Manner	
$e(V, C^+) = 0.647$	$e(V, C^-) = 0.888$
	
In a backyard, the man wearing white pants is watching the man wearing a grey tank top as he holds food in his hand	In a backyard, the man wearing white pants is chopping wood
Action Binding	
$e(V, C^+) = 0.074$	$e(V, C^-) = 0.166$
	
First, A man in a green jacket is throwing an upper cut punch to a policeman's face. Then, At a police station, a policeman is scowling at a man in a green jacket	First, At a police station, a policeman is scowling at a man in a green jacket. Then, A man in a green jacket is throwing an upper cut punch to a policeman's face
Event Chronology	
$e(V, C^+) = 0.081$	$e(V, C^-) = 0.134$

Figure 9. VELOCITI samples classified incorrectly even for ClassicVLE. In **Agent Random**, the scene describes a woman opening the door for a man and hugging him. C^- replaces the person opening the door with a random person (a bald man), and the model makes a mistake - scoring the negative caption (0.668) considerably more than the positive caption (0.091). **Action Manner** has a video of two women driving into the scene where a blonde woman picks up a bag from the backseat using her left arm. The C^- modifies how the bag is picked up - with both hands, which is clearly incorrect. However, the model makes a mistake and prefers the negative caption (0.888) over the positive caption (0.647).

B. Benchmark Creation and Details

In this section, we provide details about our benchmark. In particular, we share all prompts used for creating positive captions and various tests (Appendix B.1, Appendix B.2), share our process on creating a benchmark subset for evaluating closed models (Appendix B.3), provide benchmark statistics (Appendix B.4), discuss the strategy used to manually verify and clean all the tests (Appendix B.5), and finally provide some compute and runtime details that are required to evaluate on our benchmark (Appendix B.6).

B.1. Prompt for Converting SRL Dictionary to a Positive Caption

The prompt for generating the positive caption given an SRL dictionary is shown in Fig. 10. This refers to the discussion from Sec. 3.1 in the main paper. We use a two-stage strategy that first inserts all elements of the SRL dictionary in a sentence and then refines it for proper grammatical structure.

B.2. Prompts for Creating Test Samples

The prompt above (Fig. 10) helps create the positive caption for multiple tests. Specifically, Agent Random Test, Agent Binding Test, Action Adversarial Test, Action Manner Test, and Action Binding Test, all use the above strategy, while Agent Coreference Test and Event Chronology Test adopt templates that are filled in with the complete (or partial) positive captions.

The negative prompts for Agent Random Test, Agent Binding Test, and Action Binding Test are also created in the same

System Prompt

Using the provided dictionary containing verb and argument-role pairs in the style of PropBank, follow these steps to generate two captions

Naive Caption: Generate a caption that faithfully reflects all details from the dictionary without adding or omitting any information. Ensure that every argument detail is accurately included in the Naive Caption.

Fluent Caption: If the Naive Caption is already fluent and naturally phrased, directly copy it to the Fluent Caption. If necessary, refine the Naive Caption for improved language fluency while strictly maintaining all original details and arguments from the dictionary.

Please proceed with generating the Naive Caption first, ensuring it remains comprehensive and accurate based on the provided dictionary entries. Then, if adjustments are needed to enhance fluency, refine the Naive Caption into the Fluent Caption while ensuring that no details are overlooked or omitted.

Few Shot Example 1

```
{'Verb': 'walk (walk)',  
'Arg0 (walker)': 'man in suit',  
'ArgM (direction)': 'into room',  
'ArgM (manner)': 'slowly',  
'Scene of the Event': 'Warehouse'}
```

Naive Caption: In a warehouse, a man in suit is walking slowly into the room.

Fluent Caption: In a warehouse, a man in suit is walking slowly into the room.

Few Shot example 2

```
{'Verb': 'burn (cause to be on fire)',  
'Arg0 (thing burning)': 'Wreckage',  
'ArgM (location)': 'Wreckage'}
```

Naive Caption: The wreckage is burning on the wreckage.

Fluent Caption: The wreckage is burning.

Figure 10. Prompt to generate the positive caption given an SRL dictionary.

System Prompt

Your objective is to generate a contradiction caption using the provided PropBank style “input dictionary” and the ‘Verb’ labelled as ‘source’ based on a specific “misalignment scenario” called “verb misalignment”. In this scenario, you should suggest an alternative contradictory value for the “source” and label it as “target”.

Key Requirements

1. “naive caption + verb misalignment”: should be plausible and could theoretically occur in real life.
2. The “fluent caption + verb misalignment”: If the “naive caption + verb misalignment” is already fluent and naturally phrased, directly copy it to the “fluent caption + verb misalignment”. If necessary, refine the “naive caption + verb misalignment” for improved language fluency while strictly maintaining all original details and arguments from the dictionary

Guidelines

1. The “target” should introduce a contradiction when compared to “source”, without being a mere negation.
2. The “naive caption + verb misalignment” should be clearly distinguishable from the scene described by the “input dictionary” and should be visually distinguishable.
3. Your replacements should be creative yet reasonable.
4. If adjustments are needed to enhance fluency, refine the “naive caption + verb misalignment” into the “fluent caption + verb misalignment” while ensuring that no details are overlooked or omitted.

Few Shot Example 1

```
{'Verb': 'speak (speak)',  
'Arg0 (talker)': 'a man with dark hair',  
'Arg2 (hearer)': 'old man'  
'ArgM (manner)': 'greeting him',  
'Scene of the Event': 'warehouse'}
```

Target: Ignore

Naive Caption: On the front porch, a man with dark hair is ignoring an old man, greeting him.

Fluent Caption: On the front porch, a man with dark hair is ignoring an old man.

Few Shot Example 2

```
{'Verb': 'open (open)',  
'Arg0 (opener)': 'woman with long hair',  
'Arg1 (thing opening)': 'the front door',  
'ArgM (manner)': 'slowly',  
'Scene of the Event': 'inside a house'}
```

Target: Close

Naive Caption: Inside a house, a woman with long hair is closing the front door slowly.

Fluent Caption: Inside a house, a woman with long hair is closing the front door slowly.

Figure 11. Prompt to generate the negative caption for Action Adversarial Test.

manner as above by first replacing the specific Verb or Arg0 in the dictionary followed by strategy above.

Finally, the prompt for generating the Action Adversarial Test negative caption is shown in Fig. 11 and for Action Manner Test negative captions in Fig. 12. Both involve generating a target replacement that seems reasonable followed by converting the SRL dictionary into a caption.

B.3. Subset Creation

We created a subset of VELOCITI with 150 samples in each test. The subset was curated through random tries such that the StrictVLE performance of the OV-72B model was comparable to the full set, allowing for fair comparisons.

System Prompt

Your objective is to generate a contradiction caption using the provided PropBank style “input dictionary” and the ‘ArgM (manner)’ labeled as ‘source’ based on a specific “misalignment scenario” called “manner misalignment”. In this scenario, you should suggest an alternative contradictory value for the “source” and label it as “target”

Key Requirements

1. “naive caption + manner misalignment”: should be plausible and could theoretically occur in real life.
2. The “fluent caption + manner misalignment”: If the “naive caption + manner misalignment” is already fluent and naturally phrased, directly copy it to the “fluent caption + manner misalignment”. If necessary, refine the “naive caption + manner misalignment” for improved language fluency while strictly maintaining all original details and arguments from the dictionary.

Guidelines

1. The “target” should introduce a contradiction when compared to “source”, without being a mere negation.
2. The “naive caption + manner misalignment” should be clearly distinguishable from the scene described by the “input dictionary.”
3. Your replacements should be creative yet reasonable.
4. If adjustments are needed to enhance fluency, refine the “naive caption + manner misalignment” into the “fluent caption + manner misalignment” while ensuring that no details are overlooked or omitted

Few Shot Example 1

```
{'Verb': 'look (vision)',  
'Arg0 (looker)': 'a man wearing all black',  
'Arg1 (thing looked at or for or on)': 'a building'  
'ArgM (direction)': 'infront of him',  
'ArgM (manner)': 'breathing heavily',  
'Scene of the Event': 'warehouse'}
```

Target: *Whistling*

Naive Caption: Outside, a man wearing all black is looking in front of him at a building while whistling.

Fluent Caption: Outside, a man wearing all black is looking at a building in front of him while whistling.

Few Shot Example 2

```
{'Verb': 'burn (cause to be on fire)',  
'Arg0 (agent, entity causing something to be suspended)': 'climbing ropes',  
'Arg1 (thing suspended)': 'woman in pink shirt',  
'Arg2 (suspended from)': 'climbing ropes',  
'ArgM (location)': 'on the face of the rocks',  
'ArgM (manner)': 'precariously'}
```

Target: *Securely*

Naive Caption: climbing ropes are hanging the woman in a pink shirt securely on the face of the rocks.

Fluent Caption: The woman in a pink shirt is hanging on the face of the rocks from the climbing ropes securely.

Figure 12. Prompt to generate the negative caption for Action Manner Test.

B.4. Benchmark Statistics

We present some statistics highlighting the diversity and nuance in the VELOCITI benchmark. Since this benchmark is a subset of VidSitu [39], we observe similar trends as presented in their work.

Videos in our benchmark are complex as there are multiple agents performing various actions. Actions in VELOCITI are fine-grained. We analyze the set using Gemini-1.5-Pro which broadly categorizes actions into 6 groups: physical action and movement, communication and expression, manipulation and physical interaction, perception and mental activity, physiological actions, and general activities and states. In general, models struggle slightly more with physiological actions (performance $\sim 10\%$ lower) as compared to the average. Some verbs from these categories are shown in Fig. 13, note that the size of the word here does *not* correspond to its frequency in the dataset.



Figure 13. Word-cloud of some actions in VELOCITI in different action categories as suggested by Gemini-1.5 Pro. Word **size does not correspond to frequency** and is assigned randomly for visualization.

Fig. 14a shows that around 87% of the videos contain 4 or more unique verbs, and Fig. 14b shows that about 85% of videos contain 2 or more unique agents (people performing actions). We evaluate binding by leveraging the fact that one agent can perform multiple actions in the video, and the richness of the SRL annotations ensure that these events are described adequately. In Fig. 14c, we observe that over 70% of the events contain 4 or more SRLs (*e.g.* agent, patient, manner, *etc.*), indicating the detail-oriented nature of the annotations. Finally, Fig. 14d shows that over 72% of agents occur twice or more in their corresponding video annotation. These agents would likely be performing two different actions, and we utilize this to create two references to the same agent in tests such as Agent Coreference Test.

B.5. Quality Control

To ensure that the data generated from the automated pipelines discussed earlier are correct, we filtered the data samples manually, following specific guidelines discussed in this section. The final count of the data samples is reported in Tab. 11.

SRL dictionary to caption. The instructions and the interface for evaluating caption quality is described in Fig. 17. For each sample, three choices were provided: positive if the caption is correct, negative if the caption is wrong, and neutral if the caption cannot be negative but contains some ambiguity due to which it could not be considered positive. Out of the 380 samples that were manually verified, 356 were marked as positive, 21 were neutral, and 3 were negative. The number of positive and neutral samples was high (99.2%).

All tests. For each sample of all tests, we perform a meticulous cleanup. The instructions and the interface are presented in Fig. 18. For each video, the green bar contains a positive caption, and the red bar contains a negative caption. Unlike human evaluations, the positive and the negative captions are known while filtering. Only the samples for which both positive and negative captions are deemed appropriate are retained.

B.6. Runtime and Compute Details

While benchmarks on long videos are interesting [13, 15], VELOCITI proposes important challenges that every Video-LLM needs to solve. The short 10 s videos enable fast evaluation and make the benchmark accessible: running OV-7B on all tests (except the Control Test) takes about 2.6 hours on a single RTX 4090 GPU (24 GB).

C. Model Evaluation Prompts

We present the prompts used for all open Video-LLMs, Gemini-1.5-Flash, and GPT-4o. The entailment and MC evaluation prompts for open models, such as Qwen2-VL, LLaVA-OneVision, and Gemini-1.5-Flash are provided in Fig. 15. Prompts

Test	Videos	# Samples	Subset
Ctrl	850	2635	150
AgRand	588	873	150
AgBind	615	1356	150
AgCref	183	339	150
ActAdv	355	438	150
ActMan	378	458	150
ActBind	615	1459	150
EvChr	521	1234	150

Table 11. Number of videos and samples across different tests in VELOCITI.

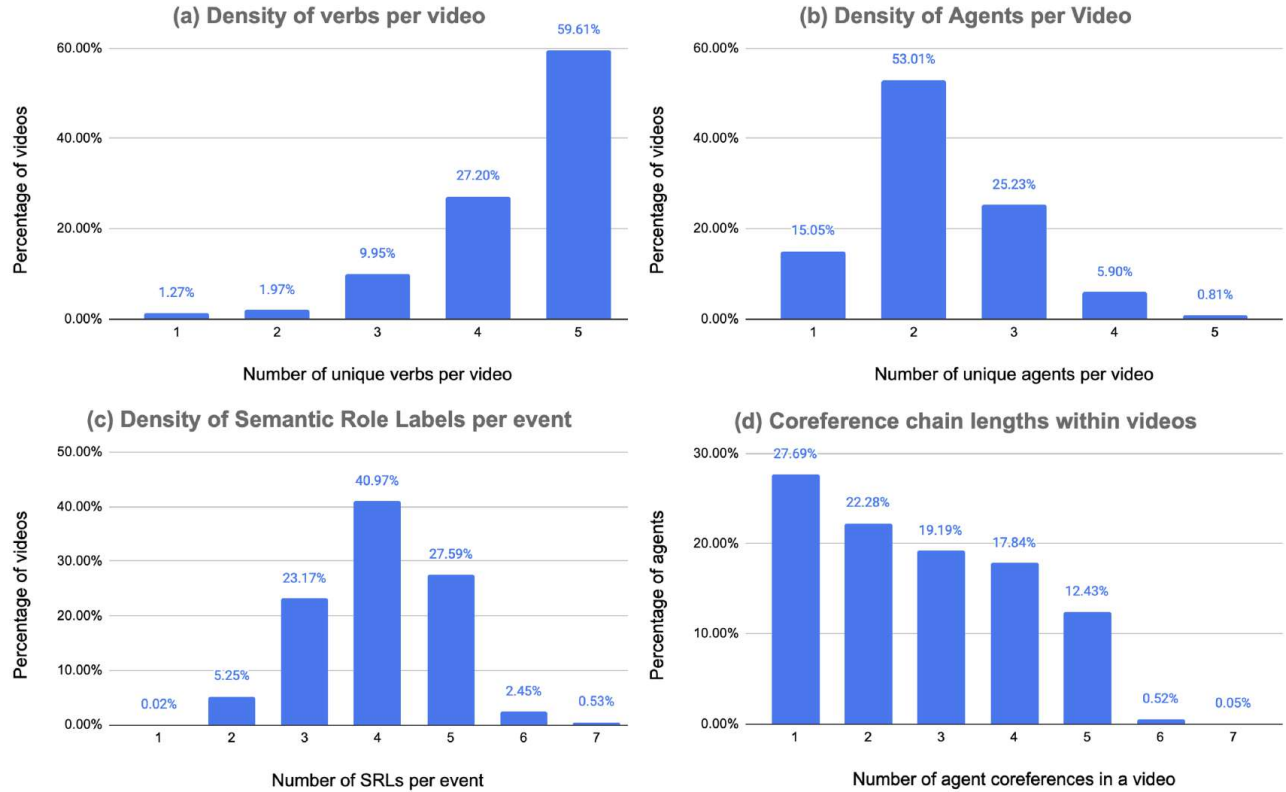


Figure 14. Statistics of various features of the VELOCITI benchmark. (a) and (b) show the distribution of verbs and agents per video, respectively. (c) shows the density of SRL annotations per event; and (d) shows the distribution of agent coreference lengths. Even with short videos, the complexity of the VidSitu annotations make the task challenging.

for GPT-4o are shown in Fig. 16. Note that GPT-4o is provided the explicit instruction of being provided frames of a video, while others are directly given a video.

Although some closed models have started optionally sharing logits, they are restricted to a limited top-K set, *e.g.* top-20 for GPT-4o. Hence, the logits for the ‘Yes’ and ‘No’ tokens may not always be included in these top-k values. To ensure the evaluation of closed models covers maximum data samples, the prompts were slightly modified to explicitly include the instruction: “Just answer with either Yes or No.”.

Entailment Prompt

Carefully watch the video and pay attention to the sequence of events, the details and actions of persons.

Here is a caption that describes the video: *Caption*

Based on your observation, does the given video entail the caption?

MC Prompt

Carefully watch the video and pay attention to the sequence of events, the details and actions of persons.

Here are two captions that describe the video.

A) *Caption₁*

B) *Caption₂*

Based on your observation, select the caption that best describes the video.

Just print either A or B.

Figure 15. Prompts for **Open Video-LLMs** and **Gemini-1.5-Flash** and **Gemini-1.5-Pro**. **Top:** Entailment evaluation prompt. **Bottom:** Multiple-choice evaluation prompt.

Entailment Prompt

You are given frames sampled sequentially from a video. Carefully watch the video frames and pay attention to the sequence of events, the details and actions of persons.

Here is a caption that describes the video: *Caption*

Based on your observation, does the given video entail the caption?

Just answer with either Yes or No.

MC Prompt

You are given frames sampled sequentially from a video. Carefully watch the video frames and pay attention to the sequence of events, the details and actions of persons.

Here are two captions that describe the video.

A) *Caption₁*

B) *Caption₂*

Based on your observation, select the caption that best describes the video.

Just print either A or B.

Figure 16. Prompts for **GPT-4o**. **Top**: Entailment evaluation prompt. **Bottom**: Multiple-choice evaluation prompt.

D. Limitations

We discuss some limitations of our work.

1. One of the shortcomings is the limited ability to scale the benchmark. VELOCITI relies on SRLs, which are obtained from careful (and costly) human annotations [39]. Further, we use LLMs to generate captions from the SRL dictionary and to create several tests (Appendix B.1, Appendix B.2). However, LLMs are prone to hallucinations, and hence, we do a round of human verification to confirm that the captions are appropriate. Thus, costly human intervention is required from SRL curation to verification of individual test samples.
2. VELOCITI is not intended as a one-stop benchmark to evaluate all abilities of Video-LLMs. Instead, it evaluates Video-LLMs for facets of compositionality, a fundamental aspect of visio-linguistic reasoning. Also, as VELOCITI is derived from VidSitu, a person-centric dataset, our benchmark focuses on people and their actions/interactions.
3. Lastly, our proposed StrictVLE metric cannot be used to evaluate contrastive models, as these models do not provide a direct ‘Yes’ probability. When the alignment score is used as a proxy to the entailment score (similar to [25]), we show that contrastive CLIP-based models do not perform well even with ClassicVLE and are therefore unlikely to be competitive at a stricter entailment.

Instructions

These instruction can be opened anytime by clicking 'i' on the bottom left of the panel

Your objective is to mark whether the provided positive caption is an “*accurate*” description of the dictionary contents.

What does an “*accurate*”, positive-caption mean?

- The generated caption must include all ideas *inferred* from the dictionary, even though it may miss some exact phrases.
- Ideally, the caption should include everything from the dictionary, but if the caption misses some value of an argument (for example, *direction*), then the caption is correct only if the missing value is *implied* from the caption.
- It needs to be grammatically correct, even though it may sound uncommon in conversational English.

Positive, Neutral, Negative

- Select *positive*, when the caption clearly meets the above requirements.
- Select *neutral*, when the caption partially meets the above requirement (not fully correct).
- Select *negative*, when the caption does NOT meet the above requirements.

Other Rules

- You are only required to look at the provided dictionary, and not the video for this task.
 - Captions should NOT be marked incorrect because of noisy annotations in the dictionary.
 - Captions should NOT be marked incorrect because of abrupt capitalization inside the sentence.
-




#107092418
101 of 101 < >

Compositionality Benchmark Human Evaluation


Name	Value
event	Ev2
video	v_bky6BtAbTU8_seg_85_95

Name	Value
.Verb	disbelieve (not believe)
Arg0 (non-believer)	girl in a gray hat
Scene of the Event	in the woods


In the woods, a girl in a gray hat disbelieves.

☐ Positive  ^[1] ☐ Neutral  ^[2] ☐ Negative  ^[3]






Info Comments History

Add a comment 

Regions Relations

Manual By Time 

Regions not added

Submit

Figure 17. Instructions and interface to verify the quality of captions generated from LLaMA-3-70B.

Instructions

These instructions can be opened anytime by clicking 'i' on the bottom left of the panel

You are given 1 video and 2 captions for each task, one **correct caption** and a **negative caption**.

Please watch the video and *verify* if the positive and the negative captions are “*logically correct*”.

What is a “*logically-correct*”, **positive caption**?

- Caption that provides a *correct* description of the event in the video.
- It should correctly identify the entities (*humans, animals, objects, etc.*) and the relationships (*action*) between them.
- Spelling/grammatical errors, if any, shall be ignored.

What is “*logically-correct*”, **negative caption**?

- Caption that provides an *incorrect* description of events in the video.

Note

- You may watch the video multiple times, if required.
- Careful and precise judgement is requirement, as point-of-difference between the **positive** and the **negative** caption, may be subtle.

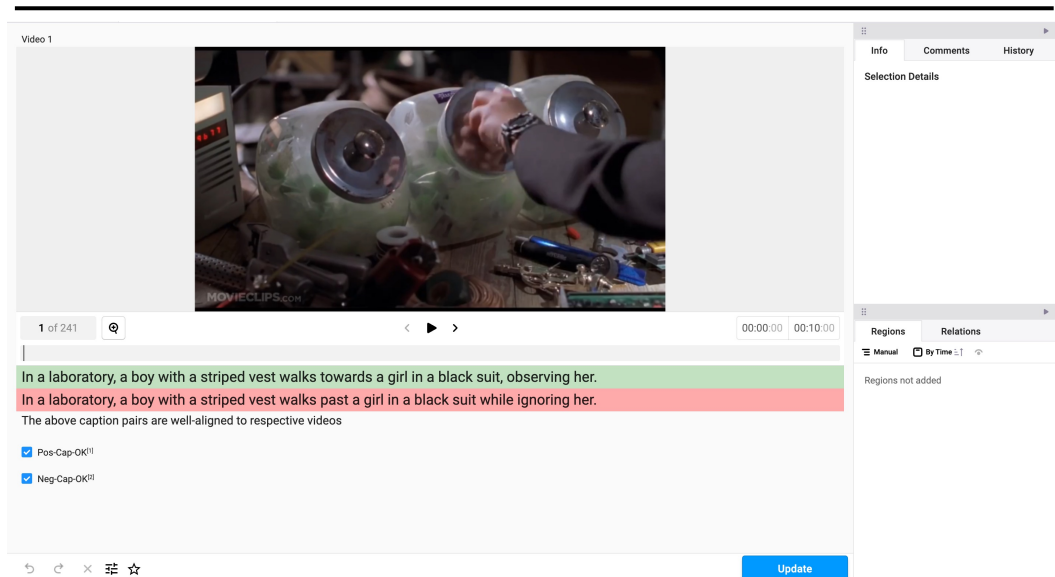


Figure 18. Data cleaning instruction for all the tests.