How Can Objects Help Video-Language Understanding?

Supplementary Material

We elaborate on the implementation details of ObjectMLLM in Appendix A. In Appendix B, we explore the design choices for the bounding box projector, the effectiveness of object bounding boxes compared to object labels, the modality fusion strategy, and bounding box sampling rate. In Appendix C, we conduct experiments on NExT-QA and IntentQA to show that spatial cues play a minor role on these two benchmarks. Appendix D demonstrates the quality of the detected object bounding boxes and investigates how it affects the performance of ObjectMLLM. Qualitative results of our method are presented in Appendix E. Finally, Appendix F summarizes a few unsuccessful attempts.

A. Implementation Details

A.1. Object detection and tracking

In the object detection and tracking process, the video keyframes are sampled at 1 FPS. SAM 2 tracking is performed at the original frame rate of each video. We use the pre-trained YOLO-World checkpoint `YOLO-World-v2-L_CLIP-Large_800`. The employed SAM 2 checkpoint is `sam2.1_hiera_large`.

All the benchmarks (or their source datasets) in our experiments provide either manually annotated or algorithm-detected object bounding boxes. To adapt YOLO-World to the benchmarks, we fine-tune it on the training set of each benchmark individually. Fine-tuning is performed with a learning rate of $2e-4$, a weight decay of 0.05, and a batch size of 64. The number of training images, the number of training epochs, and the score thresholds used during inference are listed in Table A1.

A.2. Downsampling Rates of Bounding Boxes

As described in Section 4.2, we uniformly and temporally downsample the bounding box sequences to reduce the total number of input tokens. As the videos in CLEVRER-MC [18, 44] are 5 seconds in length, we sample one frame every one second, resulting in 6 sampled frames per video. For other benchmarks, the videos have varying lengths and numbers of objects. Therefore, we assign a separate sampling rate for each video. Specifically, we use binary search to find the maximal sampling rate for each video such that the number of bounding box tokens after downsampling is less than 1,000. We show the distributions of the sampling rates and the resulting numbers of frames in Figure A1.

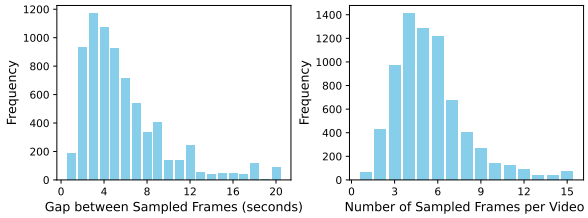
A.3. ObjectMLLM training

When fine-tuning LLaMA3-8B with LLaMA-Adapter [51], we use a batch size of 64. The learning rate is linearly warmed up to 0.0225 in the first 20% steps, after which cosine learning rate annealing is applied. The same learning rate is applied to the LLaMA-Adapter weights, bounding box embedding projector, and visual embedding projector.

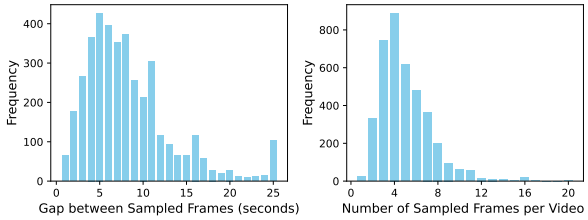
When fine-tuning VideoLLaMA2 with LoRA [8], we use a LoRA rank of 128 and a batch size of 128. The learning rate is linearly warmed up to $2e-5$ in the first 3% steps, after which cosine learning rate annealing is applied. The

Benchmark	#training images	#epochs	score threshold
CLEVRER-MC	60 k	7	$1e-3$
Perception Test	46 k	50	0.35
STAR	106 k	20	0.2
NExT-QA & IntentQA	151 k	10	0.4

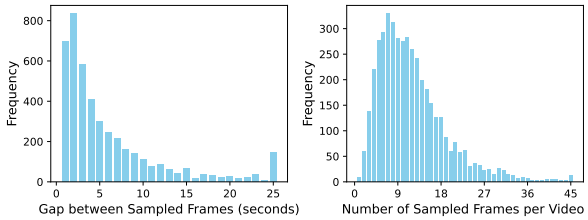
Table A1. Hyperparameters in YOLO-World fine-tuing and inference across different benchmarks.



(a) Perception Test



(b) STAR



(c) NExT-QA & IntentQA

Figure A1. Distributions of the bounding box sampling rates for Perception Test [29], STAR [41], NExT-QA [42], and IntentQA [15]. We show the gaps between the sampled frames and the numbers of the sampled frames. The last bin includes all elements greater than or equal to the corresponding x-coordinate value. IntentQA shares the same distribution as NExT-QA because it is sourced from NExT-QA.

Box adapter	Initialization	CLEVRER-MC	Perception Test
Embedding Projector	Random	42.8	57.6
	Zero	**64.9**	**60.1**

Table A2. Ablation on the initialization of the box embedding projector. The defulat random initialization in PyTorch is Kaiming uniform distribution. We find that zero-initialized linear layer as the box embedding projector always yields better performance than the default initialization.

same learning rate is applied to the LoRA weight and visual embedding projector. The pre-trained checkpoint used is `VideoLLaMA2-7B-16F`.

In both settings, the model is trained for 1 epoch on CLEVRER, 5 epochs on NExT-QA and STAR, 10 epochs on Perception Test and IntentQA, and 50 epochs on BABEL-QA. The vision encoder is kept frozen in all experiments.

B. Ablation Studies

B.1. Embedding projector

We show that the language-based box adapter is always more performant than the embedding projector in Section 4.3. However, it is possible that the poor performance of the embedding projector is due to its design. In this section, we explore a few design choices of the embedding projector, including initialization, number of layers, and number of resulting tokens. We find that the embedding projector is inferior to the language-based representation across all design choices.

Projector initialization. When training a projector between a novel modality and the LLM backbone, previous works (*e.g.* LLaVA [22] and Vamos [36]) use the default initialization for linear layers (the Kaiming uniform distribution in PyTorch). However, we find that the default random initialization significantly impedes the training of the bounding box embedding projector. As shown in Table A2, we find that initializing the linear layer weights to zero can facilitate the learning of embedding projector. We hypothesize that the default initialization would project the bounding boxes outside the LLM word embedding space, confusing the LLM backbone at the beginning of training. In contrast, a zero-initialized linear layer can map every bounding box to a zero vector, which is extremely close to the special tokens in the LLM vocabulary. This can prevent the bounding boxes from corrupting LLM behavior.

Number of projector layers. Instead of a single linear layer, we explore using a multilayer perceptron (MLP) as the bounding box projector. In this experiment, we set the number of hidden units in each layer to match the dimension of word embeddings of the LLM (4,096 for LLaMA3-8B). We use GeLU as the activation function. As Table A3 suggests, increasing the number of MLP layers yields only marginal performance gains for the embedding projector method. More importantly, it is still dominated by the language-based representation approach.

Box adapter	#Layers	CLEVRER-MC	Perception Test
Embedding Projector	1	64.9	60.1
	2	65.0	59.7
	3	65.0	60.2
Language-based Representation	-	**77.6**	**63.5**

Table A3. Ablation on the number of layers of the box embedding projector. Enlarging the number of MLP layers does not bring significant improvement. And they are outperformed by the language-based representation box adapter.

Box adapter	#Tokens per box	CLEVRER-MC
Embedding Projector	1	64.9
	9	63.4
Language-based Representation	9	**77.6**

Table A4. Ablation on the number of resulting tokens in the embedding projector method. Increasing the number of tokens to be the same as that in the language-based representation method degrades the performance.

Number of resulting tokens. While the embedding projector maps each bounding box into only one token, the language-based representation uses nine tokens to describe one bounding box (four numbers, three spaces, and two square brackets). One might argue that the expressiveness of the bounding box embedding projector is limited by the number of tokens. To address this concern, we experiment with bounding box projectors that map each bounding box into nine tokens rather than one token. The results are in Table A4. We find that increasing the number of resulting tokens per bounding box cannot improve the performance of the embedding projector method.

B.2. Are object labels sufficient on their own?

As shown in Figure 3, object labels (i.e., object names) are also provided when we format the bounding boxes. If the object labels are hidden, the bounding boxes themselves convey much less information because the object associated with each box is unknown. Object labels can provide important information to the model; for example, color recognition in Figure 7 is improved, which is definitely not inferable from unannotated bounding boxes alone. We raise the question: are object labels sufficient on their own, or does the model still derive benefits from bounding box information?

To answer this question, we train a model with object labels provided but bounding boxes hidden. In Table A6, we find that the model consistently performs better when the

Video	Caption	Box	CLEVRER-MC	Perception Test	STAR	NExT-QA	IntentQA
✓			40.3	59.6	59.7	70.7	68.2
	✓		47.8	62.4	60.1	76.6	75.7
		✓	77.6	63.5	59.1	63.7	66.2
	✓	✓	75.5(75.8)	65.7(64.1)	64.4(63.7)	76.6(77.2)	75.6(75.4)
✓	✓	✓	75.4(29.5)	63.9(33.8)	62.9(62.8)	76.2(75.8)	75.0(73.4)

Table A5. Ablation on modality fusion strategy. Blue ones are the results of jointly training on all the modalities at once, while the others are trained in a modality-by-modality manner. The modality-by-modality fusion strategy can outperform the jointly training in most cases. And the joint training is sometimes unstable when the video inputs are involved.

Input	CLEVRER-MC	Perception Test	STAR
Obj. label	59.8	60.0	58.8
Obj. label + box	**77.6**	**63.5**	**59.1**

Table A6. Ablation on the bounding boxes versus object labels. The model indeed utilizes the boxes other than the object labels.

FPS	0.25	0.5	1	2
Accuracy	74.0	73.9	77.6	78.0

(a) CLEVRER-MC

Max Seq. Len.	600	800	1000	1200
Avg. FPS	0.13	0.20	0.26	0.31
Accuracy	62.7	63.0	63.5	63.7

(b) Perception Test

Table A7. Ablation on bounding box temporal sampling rate. Higher sampling rate usually leads to higher accuracy.

object boxes are also provided, indicating that the model makes effective use of object bounding boxes. However, the differences are more notable on CLEVRER-MC and Perception Test than on STAR, indicating the improvement made by observing bounding boxes on STAR is mainly attributed to the revealed object labels. This is reasonable because the questions in STAR are annotated based off the scene graphs. It also highlights that emphasizing the question-related objects instead of describing the scene at a high level through video frame captions helps more on video question answering.

B.3. Modality fusion strategy

Section 3.4 mentions that we fuse the modalities (captions, bounding boxes, and videos) in a modality-by-modality approach instead of joint training at once. In Table A5, we compare the two fusion strategies. We find that the modality-by-modality method outperforms joint training in the majority of cases. In addition, the joint training approach is sometimes unstable. It leads to extremely low performance on CLEVRER-MC and Perception Test when using all three input modalities. Lastly, we find that both fusion methods are ineffective at leveraging the visual embeddings. We therefore urge the development of new multimodal fusion strategies that can make visual inputs valuable.

Figure A2. LLaVA-1.5 caption vs. GPT-4o spatial-free caption.

Caption	Box	NExT-QA	IntentQA
LLaVA-1.5		76.6	75.7
LLaVA-1.5	✓	76.6	75.6
GPT-4o Spatial-free		76.1	73.6
GPT-4o Spatial-free	✓	76.3	75.6

Table A8. Performance with spatial-free GPT-4o captions.

B.4. Temporal sampling rate of bounding boxes

We conduct an ablation study on the bounding box sampling rate on CLEVRER-MC and Perception Test in Table A7. On CLEVRER-MC, all the videos have the same duration. Therefore, we directly adjust the bounding box sampling rate from 0.25 FPS to 2 FPS. On Perception Test, we vary the maximum box token length mentioned in Section 4.2 and report the corresponding average frame rates. As the results show, a higher sampling rate generally yields higher accuracy, indicating a trade-off between performance and efficiency.

C. Spatial Information Ablation on NExT-QA and IntentQA

In our experiments in Table 1 and Table 3, incorporating object bounding boxes as additional information does not improve model performance on NExT-QA and IntentQA. While these benchmarks focus on causal reasoning about events, the spatial information may not play a crucial role to answer the questions. In this section, we further verify through experiments the limited usefulness of spatial information in these two benchmarks.

As Figure A2 shows, the video frame captions generated by LLaVA-1.5 include some spatial cues in the image. To eliminate the influence of spatial cues, we feed each video frame to GPT-4o and ask it to generate a caption that does

Box	CLEVRER-MC	Perception Test	STAR	NExT-QA	IntentQA
Model-tracked	**77.6**	63.5	59.1	63.7	**66.2**
Annotation	74.9	**66.8**	**78.9**	**65.5**	65.3

Table A9. Model performance with object bounding boxes tracked by computer vision models or with those annotated by the benchmarks. Bounding box annotations in CLEVRER-MC, NExT-QA, and IntentQA are also algorithm-detected. Boxes in Perception Test and STAR are manually annotated. The experiments are in the box-only setting.

not involve any spatial information. Specifically, we use the following prompt:

> <image> Faithfully describe the content of the above image, avoid mentioning specific objects and try your best to provide a scene-level, holistic summary. No mention of any spatial information.

Figure A2 shows an example of such spatial-free captions. With the spatial-free captions, we repeat our experiments on NExT-QA and IntentQA, using ObjectMLLM with LLaMA3 as the backbone. The results are in Table A8. We find that removing the spatial cues only results in a 0.5% drop in accuracy on NExT-QA and a 2.1% drop on IntentQA. This indicates that the questions in these two datasets are still highly answerable even without spatial information. More importantly, incorporating object bounding boxes in addition to spatial-free captions can recover the performance, especially on IntentQA. This again demonstrates our method's effectiveness in conveying spatial information to MLLMs.

D. Quality of the Tracked Bounding Boxes

In this section, we assess the quality of the extracted object bounding boxes and whether it hinders the performance of our model. Because the evaluation benchmarks themselves provide object bounding box annotations (either human-annotated or algorithm-detected), we can compare the boxes obtained by our workflow with them.

Figure A4 visualizes the tracked and annotated bounding boxes across different benchmarks. All the examples are from the validation/test set of the benchmarks. We find that the detection and tracking quality on synthetic videos (CLEVRER-MC) is highly accurate. On the realistic videos (Perception Test, STAR, NExT-QA, and IntentQA), our employed tracking method can capture the main objects although some noise is present.

In Table A9, we evaluate our model with the bounding box annotations as inputs. When the annotations serve as inputs, the model is trained again with the annotated boxes before evaluation to mitigate train-test domain shift. While the performance with model-tracked bounding boxes is 2%–3% worse than that with annotations on Perception Test and NExT-QA, it is even better than the annotations on CLEVRER-MC and IntentQA. This is reasonable because the bounding box annotations provided in the CLEVRER and IntentQA benchmarks are also algorithm-detected and potentially noisier than ours. In contrast, Perception Test and STAR both provide human-annotated object bounding boxes. However, the performance gap on STAR is significantly larger than on Perception Test. We notice that the questions in STAR are generated by functional programs based on annotated object relation graphs. Because only objects of interest are annotated in STAR, using object annotations as input provides a strong prior that may bias the answers. As the tracking quality on STAR (Figure A4(c)) is fairly accurate, we hypothesize that the large performance gap is caused by the choices of objects of interest rather than the tracking precision. How we can filter the objects of interest from a video remains an open and valuable research challenge.

E. Qualitative Results

We show a qualitative result on CLEVRER-MC in Figure A5. The model can determine whether an object is moving based on its bounding boxes, which is an essential capability in this benchmark.

In Figures A6 to A9, we show qualitative results from Perception Test. In these examples, model can determine the motion of cameras, stability of objection configurations, and the number of objects taken out from bags. These questions are not answerable for the caption-only model.

In Figures A10 and A11, we show failure cases of our caption-and-box model. From the captions and object bounding boxes, the model cannot reliably infer the object states and appearances. Therefore, it fails to provide correct answers. However, visual embeddings, in principle, should capture these visual characteristics. We highlight the importance of devising MLLMs that can efficiently and effectively utilize distributed visual representations.

We also examine the failure cases on NExT-QA and IntentQA, where we found that our model struggles with questions involving human actions. For example, in Figure A12, the model with bounding box inputs is aware that there are a person and a dog in the video. However, the person's action cannot be determined from the bounding boxes. In contrast, as captions provide explicit descriptions of actions, the model with caption input is better at answering these questions, contributing to the performance gap in Table 1 compared to the box-only model.

Box adapter	Perception Test
Language-based Representation	63.5
Embedding Projector	60.1
Visual Prompting	59.7

Table A10. Performance of integrating bounding boxes using visual prompting. It is significantly less performant than the language-based representation and embedding projector.

Box	Visual embedding	Perception Test Acc.
✓	✗	63.5
✓	Frame-level	62.7
✓	Object-level	63.3

Table A11. Ablation on different visual embedding levels. Object-level visual embedding works better than the frame-level embedding but still cannot bring additional improvement when symbolic boxes are used.

F. Unsuccessful Attempts

F.1. Integrating boxes via visual prompting

Beyond integrating object-centric information via structural bounding box coordinates, we explore incorporating box information through visual prompting. Inspired by prior work [33], we overlay bounding boxes directly onto video frames and extract visual embeddings from these annotated frames. Within the same video, objects are distinguished by unique colors, and the color assigned to each object remains consistent across frames to maintain temporal coherence. Figure A3 demonstrates an example of the annotated frames from Perception Test. As shown in Table A10, integrating bounding boxes using visual prompting performs worse than the embedding projector and language-based representation on Perception Test.

F.2. Integrating object-level visual embeddings

Intuitively, fine-grained object appearances like texture cannot be accurately described by video frame captions and bounding boxes. But, in principle, they can be captured by distributed visual representations like CLIP [30] embeddings. However, Table 1 illustrates that integrating frame-level visual embeddings upon captions and bounding boxes does not bring additional benefits to the performance.

Initially, we hypothesize that frame-level visual embeddings are too high-level to capture object details. To investigate this problem, we experiment with an object-level visual representation to replace the frame-level embedding. Specifically, we crop the objects from the video frames and extract their CLIP embeddings as object embeddings. Then, based on the language-based representation, we append each object embedding after its bounding box of the corresponding timestamp using the template below, where each ⟨|obj_emb|⟩ indicates an object embedding.

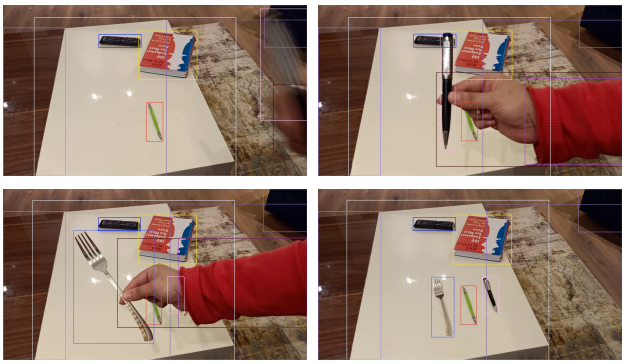
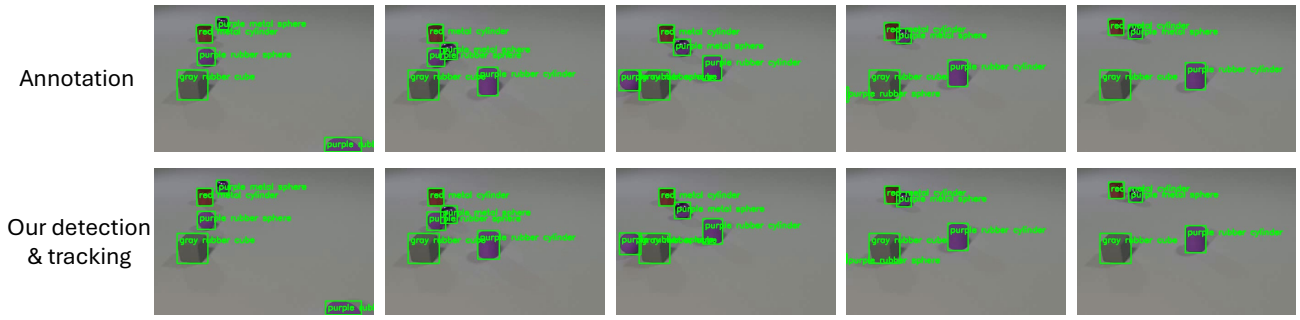


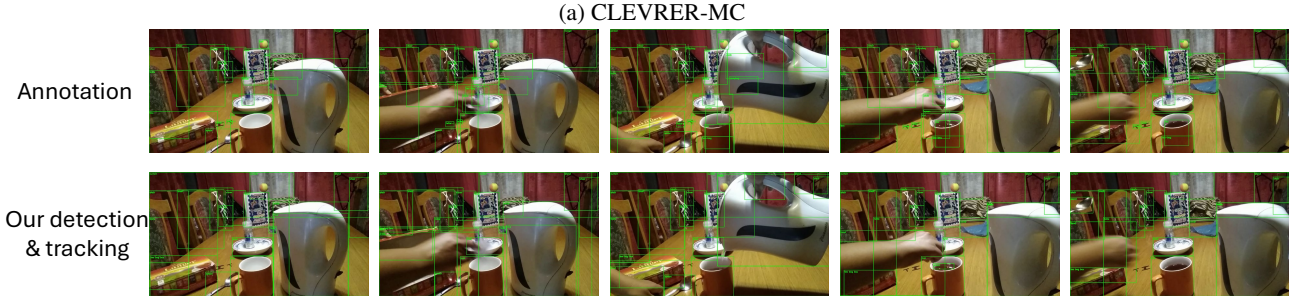
Figure A3. Examples of integrating bounding boxes via visual prompting from Perception Test.

(Object 0) bag – frame 0 [8 0 54 93] ⟨|obj_emb|⟩ frame 90 [4 0 52 94] ⟨|obj_emb|⟩ .....

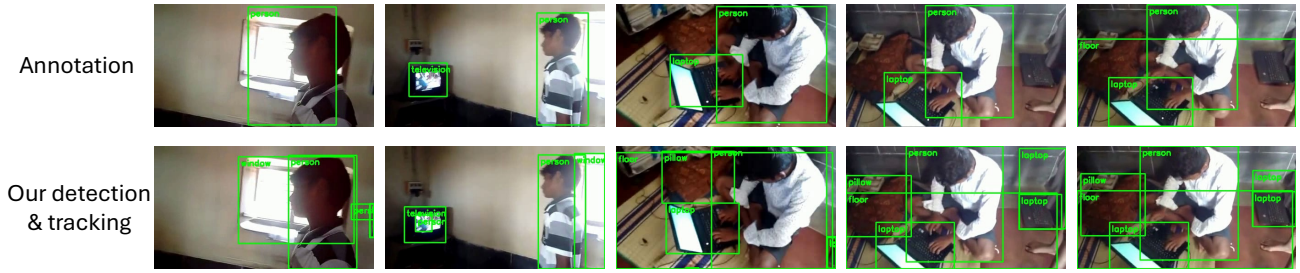
The results are in Table A11. The object-level visual embedding outperforms the frame-level embedding but still falls short of the box-only model. This experiment again highlights the difficulty of integrating distributed embedding into MLLMs in a data-efficient manner, which would be a challenging but valuable research topic.



(a) CLEVRER-MC



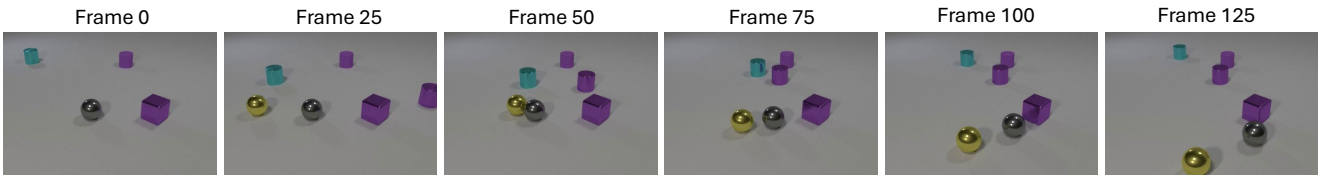
(b) Perception Test



(c) STAR

(d) NExT-QA & IntentQA

Figure A4. Visualization of the object bounding boxes across different benchmarks. IntentQA shares the same video source as NExT-QA. Our tracked bounding boxes are nearly perfect on CLEVRER-MC, while they are also fairly accurate on realistic videos.



Frame Captions

Frame 0: A white surface with three small objects placed on it…… One of the objects is a silver sphere, while the other two are cubes, one purple and the other blue…… with the silver sphere being closer to the left side of the image……
Frame 25: A white background with a variety of small, colorful objects placed on it. There are four distinct objects in the scene, each with a different color and shape…… The objects are positioned at various angles and distances from each other……
Frame 50: ……

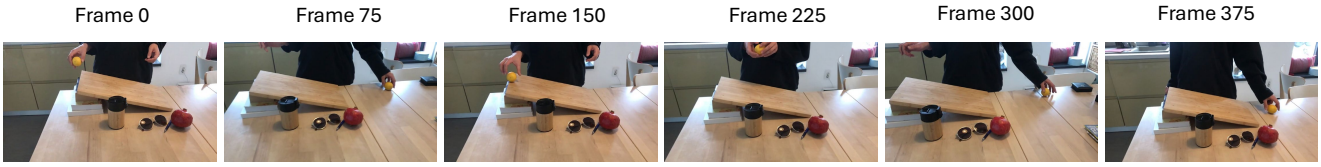
Object Bounding Boxes

(Object 0) purple metal cube - frame 0 [64 43 78 65] frame 25 [64 43 78 65] frame 50 [64 43 78 65] frame 75 [64 43 78 65]……
(Object 1) cyan metal cylinder - frame 0 [10 11 17 23] frame 25 [19 23 28 37] frame 50 [35 25 43 39] frame 75 [40 18 47 32]……
(Object 2) purple rubber cylinder - frame 0 [54 13 61 25] frame 25 [54 13 61 25] frame 50 [54 13 61 25] frame 75 [54 13 61 25]……
(Object 3) gray metal sphere - frame 0 [36 47 46 61] frame 25 [36 47 46 61] frame 50 [38 47 47 62] frame 75 [46 52 56 68]……
(Object 4) purple metal cylinder - frame 20 [99 44 100 53] frame 45 [67 28 76 44] frame 70 [50 23 58 38] frame 95 [50 23 58 37]……
(Object 5) yellow metal sphere - frame 12 [0 62 1 73] frame 37 [23 37 32 50] frame 62 [30 48 40 64] frame 87 [32 59 43 77]……

Question: How many moving metal objects are there?
Choices: (A) 2 (B) 1 (C) 3 (D) 4

Caption Model: (C)
Caption + Box Model: (D)

Figure A5. Qualitative example on CLEVRER-MC. The model can determine whether an object is moving based on its bounding boxes.



Frame Captions

Frame 0: A wooden dining table with various items placed on it. There is a wooden cutting board, a cup, a book, a pair of sunglasses, and an apple. The person is standing near the table, holding a lemon…… The wooden cutting board is placed towards the center of the table, while the cup and the book are located closer to the left side. The sunglasses are positioned on the right side of the table……
……
Frame 300: A wooden table with various items placed on it. A wooden cutting board is the main focus, with a knife and a lemon on top of it. There is also a cup and a pair of sunglasses on the table. A person is standing near the table, possibly preparing to use the cutting board. In addition to the cutting board, there are two apples on the table, one near the center and the other towards the right side……

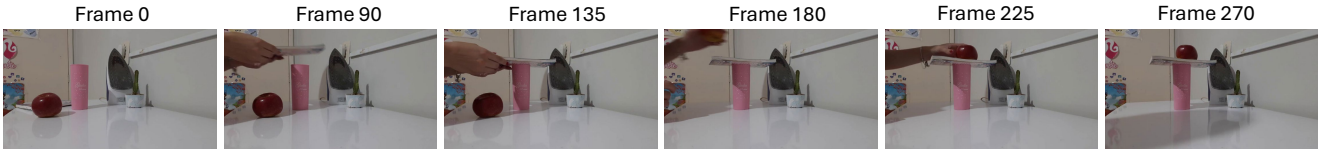
Object Bounding Boxes

(Object 0) table - frame 0 [13 36 100 100] frame 151 [5 42 94 100] frame 302 [0 42 67 100];
(Object 1) person - frame 0 [30 0 76 35] frame 151 [25 0 73 37] frame 302 [8 0 81 43];
(Object 2) wooden board - frame 0 [35 24 80 58] frame 151 [28 27 74 59] frame 302 [3 29 53 63];
(Object 3) jar - frame 0 [49 45 58 73] frame 151 [42 47 51 75] frame 302 [17 52 28 82];
……

Question: Is the camera moving or static?
Choices: (A) Moving (B) Static or shaking (C) I don't know

Caption Model: (B)
Caption + Box Model: (A)

Figure A6. Qualitative example on Perception Test. The caption+box model can determine the motion of the camera from the changing object bounding boxes.



Frame Captions

Frame 0: A white table with a variety of objects on it. There is a cup, a potted plant, and a small ironing board. The cup is placed on the left side of the table, while the ironing board is situated towards the center. The potted plant is positioned on the right side……

……

Frame 270: A white table with a pink cup sitting on top of it. The cup is filled with an apple, and a small cactus is placed nearby. Above the table, there is an ironing board with an iron on it. The scene appears to be a simple, everyday arrangement of objects in a room.

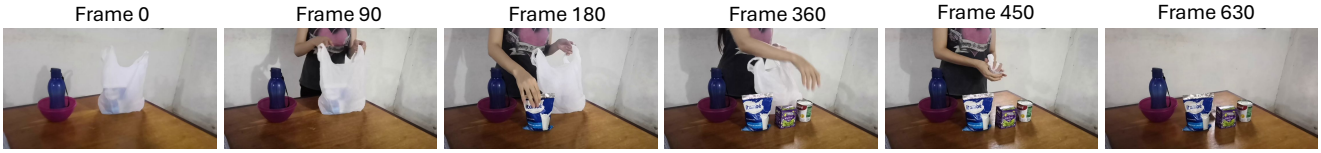
Object Bounding Boxes

(Object 0) tumbler - frame 0 [31 29 40 67] frame 90 [31 29 40 68] frame 180 [32 29 40 67] frame 270 [32 29 40 68];
(Object 5) book - frame 0 [6 59 32 68] frame 90 [24 12 47 22] frame 180 [21 23 52 32] frame 270 [21 22 52 33];
(Object 9) apple - frame 0 [13 53 28 73] frame 90 [13 53 28 74] frame 180 [16 0 28 11] frame 270 [33 13 44 49];
……

Question: Is the configuration of objects likely to be stable after placing the last object?
Choices: (A) One cannot judge the stability of this configuration.
(B) The configuration is likely to be stable.
(C) The configuration is likely to be unstable.

Caption Model: (C)
Caption + Box Model: (B)

Figure A7. Qualitative example on Perception Test. The caption+box model can predict the stability of the object configuration because it is aware of the object locations.



Frame Captions

Frame 0: A wooden dining table with a pink bowl and a white plastic bag placed on it. The pink bowl is located on the left side of the table, while the white plastic bag is situated towards the right side. The bag appears to be a grocery bag……

……

Frame 630: A wooden dining table with various food items and a bottle placed on it. There are two cans of food, one located towards the right side of the table and the other towards the left side. A box of food is also present on the table, positioned near the center. In addition to the food items, there is a bowl situated on the left side of the table, and a spoon can be seen resting inside the bowl……

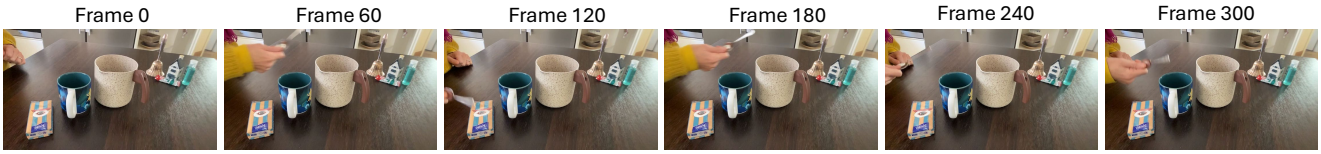
Object Bounding Boxes

(Object 3) bag - frame 0 [44 15 68 71] frame 90 [44 13 66 72] frame 180 [40 14 66 71] frame 270 [42 17 66 71] frame 360 [39 13 65 71] frame 450 [48 21 55 38] frame 540 [50 59 52 69];
(Object 5) tea bag box - frame 237 [50 27 58 33] frame 327 [52 63 62 83] frame 417 [52 64 62 83] frame 507 [52 63 62 83] frame 597 [52 63 62 83];
(Object 6) milk tetrapack - frame 124 [55 20 58 30] frame 214 [36 54 51 85] frame 304 [36 54 51 85] frame 394 [36 54 51 85] frame 484 [36 54 51 85] frame 574 [36 54 51 85];
(Object 7) box - frame 308 [53 21 57 31] frame 398 [62 59 70 78] frame 488 [62 59 69 78] frame 578 [62 59 69 78];
……

Question: How many objects did the person take out of the bag?
Choices: (A) 3 (B) 2 (C) 4

Caption Model: (C)
Caption + Box Model: (A)

Figure A8. Qualitative example on Perception Test. The caption+box model can determine the number of objects taken out from the bag with the aid of object bounding boxes.



Frame Captions

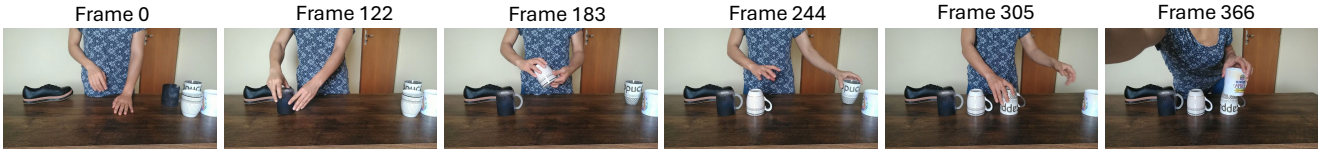
Frame 0: A wooden dining table with a cup and a mug placed on it. The cup is positioned towards the left side of the table, while the mug is situated closer to the center. The mug is larger than the cup and has a handle, making it a more functional and comfortable choice……
……
Frame 300: A wooden dining table with a variety of items placed on it. There are two cups, one of which is a coffee mug, and the other is a cream pitcher. The coffee mug is positioned towards the left side of the table…… A spoon can also be seen on the table……

Object Bounding Boxes

(Object 0) cup - frame 0 [43 21 68 65] frame 60 [43 21 68 65] frame 120 [43 21 68 65] frame 180 [43 21 68 65] frame 240 [43 21 68 65] frame 300 [43 21 68 65];
(Object 1) cup - frame 0 [25 35 41 74] frame 60 [25 35 41 74] frame 120 [25 35 41 74] frame 180 [25 35 41 74] frame 240 [25 35 41 74] frame 300 [25 35 41 74];
(Object 5) box - frame 0 [11 59 23 91] frame 60 [11 59 23 91] frame 120 [11 59 23 91] frame 180 [11 59 23 91] frame 240 [10 59 23 91] frame 300 [10 59 23 91];
(Object 9) spoon - frame 16 [0 21 2 24] frame 76 [14 28 31 39] frame 136 [1 9 12 35] frame 196 [17 21 32 32] frame 256 [30 4 47 18] frame 316 [0 23 3 29];
……

Question: What object does the person use to hit other objects? Caption Model: (A)
Choices: (A) pen (B) fork (C) spoon Caption + Box Model: (C)

Figure A9. Qualitative example on Perception Test. From the boudning box cooridnates, the caption+box model can observe that the spoon is moved to hit the other objects.



Frame Captions

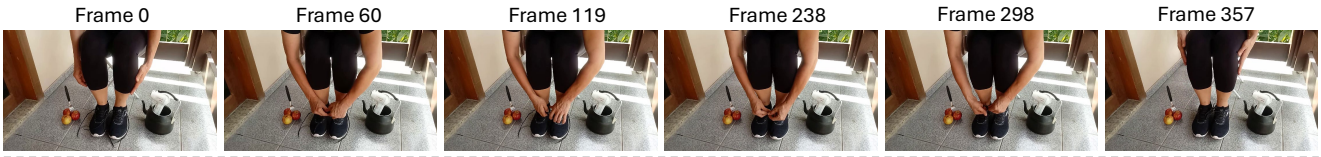
Frame 0: A person standing in front of a wooden dining table. The person is wearing a blue shirt and is positioned near the left side of the table. On the table, there is a cup placed towards the right side, and a pair of black shoes can be seen on the left side……
……
Frame 366: A wooden dining table with a variety of coffee mugs and cups placed on it. There are three coffee mugs, one of which is a tall mug, and two smaller cups. A person is standing near the table, holding a coffee mug, possibly preparing to pour coffee……

Object Bounding Boxes

(Object 0) cup - frame 0 [92 51 100 73] frame 61 [92 51 100 73] frame 122 [92 51 100 73] frame 183 [92 51 100 73] frame 244 [92 51 100 73] frame 305 [92 51 100 73] frame 366 [55 33 67 56];
(Object 2) glass - frame 0 [74 46 83 65] frame 61 [65 43 73 62] frame 122 [25 49 33 71] frame 183 [24 51 37 73] frame 244 [24 51 37 73] frame 305 [24 51 37 73] frame 366 [24 51 37 73];
(Object 6) cup - frame 0 [83 51 93 74] frame 61 [83 51 93 74] frame 122 [83 51 93 74] frame 183 [42 30 53 50] frame 244 [38 50 50 72] frame 305 [38 50 50 72] frame 366 [38 50 50 72];
(Object 7) cup - frame 0 [84 43 93 52] frame 61 [84 43 93 52] frame 122 [84 43 93 52] frame 183 [84 43 93 63] frame 244 [83 42 92 63] frame 305 [54 49 66 72] frame 366 [54 53 66 72];
……

Question: Did the person place all the containers facing upwards or downwards? Caption Model: (C)
Choices: (A) upwards (B) downwards (C) mixed Caption + Box Model: (C)

Figure A10. Failure case on Perception Test. The model cannot see the state of the mugs from either the captions or the bounding boxes. So it does not whether the mugs are upwards or downwards.



Frame Captions

Frame 0: A person sitting on a chair with their legs crossed. The person is wearing a pair of black shoes and appears to be in the process of putting on socks. A knife is placed nearby, possibly for cutting the socks. There are two apples in the scene……
……
Frame 238: A person sitting on a chair, wearing a pair of sneakers. They are in the process of tying their shoelaces, with a fork and a knife nearby. The person is surrounded by a few apples, with one placed close to the left side of the chair, another on the right side……
……
Frame 357: A woman sitting on a bench with her legs crossed. She is wearing a pair of sneakers and appears to be tying her shoelaces. There are a few apples placed on the ground near her, and a fork is also visible in the scene. A kettle can be seen in the background……

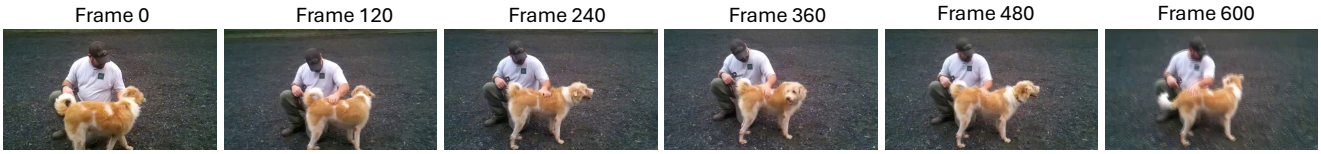
Object Bounding Boxes

(Object 2) shoe lace - frame 0 [35 65 51 89] frame 119 [35 66 43 88] frame 238 [50 65 53 72] frame 357 [47 61 51 74];
(Object 9) shoe lace - frame 0 [41 64 50 83] frame 119 [42 68 48 77] frame 238 [43 70 49 78] frame 357 [42 64 50 77];
(Object 10) shoe - frame 0 [49 63 58 90] frame 119 [49 70 58 90] frame 238 [49 70 58 91] frame 357 [49 62 58 90];
(Object 12) shoe - frame 0 [40 59 50 88] frame 119 [40 59 50 88] frame 238 [41 59 50 89] frame 357 [41 59 50 88];
……

Question: Is there something unusual about the way the person ties the shoe laces?
Choices: (A) The person ties correctly the left shoe lace, but not the right shoe lace.
(B) The person ties the shoe laces normally.
(C) The person ties the lace of the left shoe to the lace of the right shoe.

Caption Model: (B)
Caption + Box Model: (B)

Figure A11. Failure case on Perception Test. Both the captions and the bounding boxes cannot tell if the shoe laces are tied normally. This suggests that our model has difficulty in recognizing the appearance of the objects.



Frame Captions

Frame 0: A man kneeling down on the ground next to a brown and white dog. The man appears to be petting the dog, showing affection and care for the animal. The dog is positioned to the left of the man, and they are both situated on a dirt surface……
……
Frame 600: A man kneeling down on the ground next to a brown dog. The man is petting the dog, showing affection and care. The dog is positioned to the left of the man, and both of them are on a dirt surface. The man appears to be wearing a hat……

Object Bounding Boxes

(Object 0) adult - frame 0 [22 10 57 94] frame 30 [20 1 56 86] frame 60 [26 2 60 84] frame 90 [27 7 61 89] frame 120 [26 16 59 88] frame 150 [25 11 54 83] frame 180 [21 6 52 78] frame 210 [18 7 48 80] frame 240 [18 9 50 80] frame 270 [20 10 52 80] frame 300 [21 9 52 79]……
(Object 1) dog - frame 0 [24 47 67 100] frame 30 [27 34 66 100] frame 60 [31 38 68 100] frame 90 [34 40 70 100] frame 120 [37 46 71 100] frame 150 [35 44 69 100] frame 180 [33 39 69 98] frame 210 [29 39 68 98] frame 240 [29 44 70 100] frame 270 [31 44 70 99] ……

Question: Why is the man kneeling down on the floor?
Choices: (A) feed the dog (B) crawling around (C) let kids walk through
(D) fell down (E) pet the dog

Caption Model: (E)
Box Model: (A)

Figure A12. Failure case on NExT-QA. Although the detection and tracking algorithm can tell that there are an adult and a dog in the video, their actions cannot be inferred from the object bounding boxes. The captioning model can capture the person's action so that the model with captions as inputs correctly answers this question.