Supplementary Materials for V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning

Contents

A. Codes and Dataset Release

Both the evaluation code and dataset has been released, you can find it in https://github.com/V-STaR-Bench/V-STaR.

B. Additional benchmark details

Here, we provide more additional benchmark details, including the details of the automated pipeline to construct V-STaR, and more statistics on the dataset.

B.1. Pipeline Details

We provide more details of the semi-automated pipeline for generating CoT reasoning chains and questions with a coarse-to-fine granularity. As shown in Fig.2, we first automatically filter out samples where the video length is less than 15 seconds or the ratio of the video moment to the total length exceeds 50%, ensuring that the questions remain sufficiently challenging.

Then, we input a video question along with its VQA, temporal, and spatial ground truth into the GPT-4-Turbo to generate a CoT reasoning chain for the video question and sequential RSTR questions for the CoT reasoning evaluation. The full instruction prompt templates are provided to guide GPT-4-Turbo in generating the CoT reasoning chain and RSTR questions.

Finally, we manually verify the reasoning chain and the RSTR questions. We check the consistency of the generated chains to fit the VQA question, then we check if the subsequent RSTR questions are in the correct sequential logic. Ultimately, we assign the temporal and spatial labels to each sub-question to make the RSTR chains complete. We build 2 RSTR question chains for each sample to thoroughly evaluate the model's spatio-temporal reasoning ability.

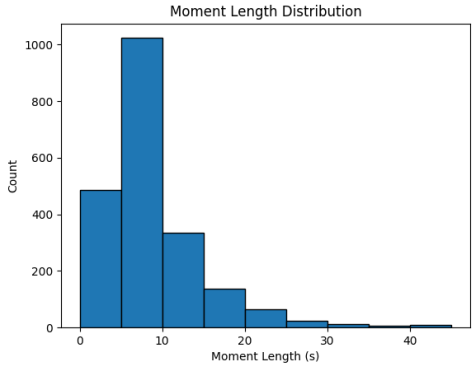

Figure 1. The distribution of moment length in V-STaR.

B.2. More Statistics on the Dataset

Here, we provide more statistics on the Moment length distribution, the ratio of box area to video resolution, and the object top-20 distribution.

Moment length distribution. The moment length distribution in the V-STaR benchmark is shown in Fig. 1. The dataset exhibits a minimum moment length of 1.70 seconds, a maximum of 47.0 seconds, and an average duration of 9.06 seconds. The majority of moments fall within the shorter duration range, particularly below 15 seconds. The ratio between moment length and video length varies from a minimum of 0.19% to a maximum of 50%, with an average of 19.29%. This ensures a reasonable level of difficulty for the temporal grounding subtask.

Ratio of box area to video resolution. The distribution of object box area to video resolution is shown in Fig. 3. The figure exhibits a right-skewed pattern where most objects

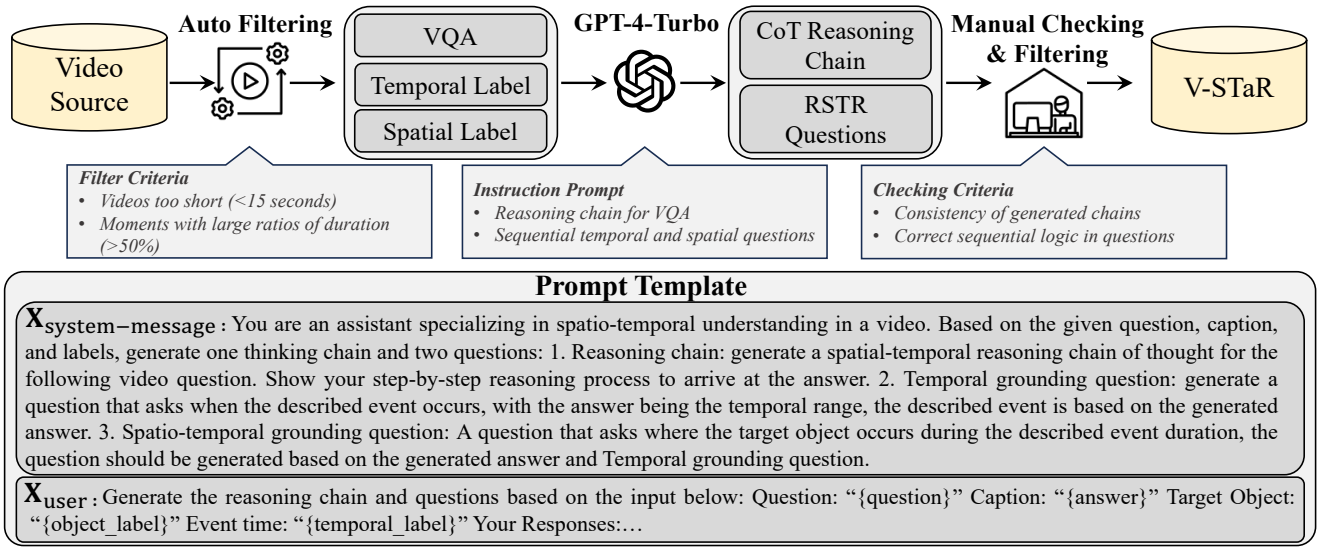

Figure 2. Illustration of the semi-automated data construction pipeline of V-STaR. We provide the full instruction prompt for GPT-4-turbo to generate a spatio-temporal reasoning CoT chain to answer VQA questions, along with a set of RSTR questions.

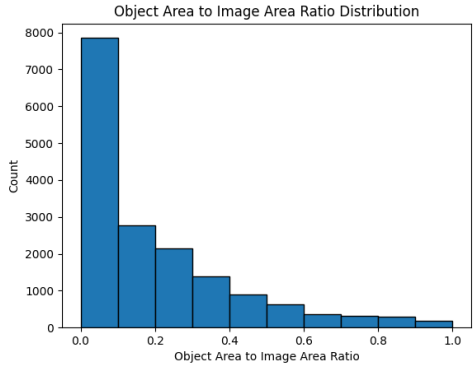

Figure 3. The distribution of moment length in V-STaR.

occupy a relatively small portion of the image. Furthermore, the average ratio of bounding box area to video resolution is 19.75%, indicating that, on average, objects occupy roughly one-fifth of the total image area. This distribution suggests that spatial grounding may need Video-LLMs' to account for a wide range of object scales. It guarantees that our spatial grounding subtask remains a challenging one.

C. Additional Implementation Details

We provide additional implementation details to describe in more detail how we evaluated the spatio-temporal reasoning capabilities of Video-LLMs using V-STaR benchmark.

C.1. Video-LLM Settings

Due to computational constraints, we evaluate only the smaller parameter versions (7B–8B) of open-source models. To ensure that models receive sufficient information for spatiotemporal reasoning, we set the video input sampling rate to 1 FPS for all models. This approach maintains a fair evaluation while providing adequate input for assessing both temporal and spatial reasoning. For cases where the number of sampled frames exceeds a model's input limit, we uniformly sample the video based on the maximum capacity of each model. Specifically, following the official configurations, we set the frame limits as follows: 128 for GPT-4o, 180 for Video-LLaMA3, 786 for Qwen2.5-VL and Qwen2-VL, 64 for InternVL-2.5 and LLaVA-Video, 16 for VideoChat2, 128 for Oryx-1.5 and VideoCCAM, 96 for TimeChat, and 128 for VTimeLLM and Trace. We do not impose a frame limit on Gemini-2 and Sa2VA, as they can process significantly longer videos.

C.2. Prompt Templates

We provide the Video-LLM prompt templates used to generate answers for the RSTR tasks in Chain 1 ("what-when-where") and Chain 2 ("what-where-when"). While the overall structure of the templates remains consistent across models, variations exist due to differences in officially provided templates and model capabilities. To illustrate our approach, we present a standardized template that captures the general framework in Fig. 4.

Prompt Templates for Chain 1. For the "what" questions in both Chain 1 and Chain 2, we use the same template. Specifically, given the input video frames, we directly pose a VQA-style "what" question.

To mitigate potential ambiguity that could prevent the model from providing a valid response—for example, GPT-4o often replies that it cannot recognize a person if the ground truth answer is a human—we include a clarification

Prompt Template for "What" in both chain

$\mathbf{X}_{\text{user}}$: Answer the question about the video: {data['question']} \n (If the answer is a person, you don't need to identify the person.)

Prompt Template for "When" in Chain 1

$\mathbf{X}_{\text{user}}$: This video is {video_length} seconds long, and {len(video_frames)} frames are uniformly sampled from it. These frames are located at {frame_time}. Answer the question about the video: {temporal_question} \n Output the start and end moment timestamps. Output the start and end moment timestamp in the format: [start_time, end_time].

Prompt Template for "Where" in Chain 1

$\mathbf{X}_{\text{user}}$: This video is {video_length} seconds long with a resolution of {w}x{h} (width x height). and {len(video_frames)} frames are uniformly sampled from it. These frames are located at {frame_time}. Please answer the question about the video: "{spatial_question}" with a series of bounding boxes in the format [x_min, y_min, x_max, y_max]. \n For each whole second within the time range {time_range} provided (inclusive of the boundaries), output a series of bounding boxes of the object in JSON format. \n In the Json, the keys should be the whole seconds (as strings), and the values should be bounding boxes in the format [x_min, y_min, x_max, y_max]. \n Example output:{{"{time_range[0]}": [x_min, y_min, x_max, y_max],...}}

Prompt Template for "Where" in Chain 2

$\mathbf{X}_{\text{user}}$: This video is {video_length} seconds long with a resolution of {w}x{h} (width x height). and {len(video_frames)} frames are uniformly sampled from it. These frames are located at {frame_time}. Please answer the question about the video: "{spatial_question}" with a series of bounding boxes in the format [x_min, y_min, x_max, y_max]. \n For each whole second that may related to the question, output a series of bounding boxes of the object in JSON format. You only need to output {len(bboxes)} bbox(es). You need to determine which frame is related to the question, and you don't need to output the bbox for the frames not related to the question. The keys should be the whole seconds (as strings), and the values should be the bounding box in [x_min, y_min, x_max, y_max]format. \n Example output: {{"0": [x_min, y_min, x_max, y_max], "1":..., ..., "{len(bboxes)}":...}} (if the frames at 0~{len(bboxes)} second are related to the questions)

Prompt Template for "When" in Chain 2

$\mathbf{X}_{\text{user}}$: This video is {video_length} seconds long, and {len(video_frames)} frames are uniformly sampled from it. These frames are located at {frame_time}. Answer the question about the video: {temporal_question} There are {len(bboxes)} bounding boxes of the key object related to the question in the video without knowing the time, which are:{bboxes}. You may find it helpful for getting the correct timestamps. \n Output the start and end moment timestamps. Output the start and end moment timestamp in the format: [start_time, end_time].

Figure 4. The prompt template used in the two chains of RSTR tasks.

in the template: "If the answer is a person, you don't need to identify the person.".

For the "when" question in Chain 1, the model is provided with the ground truth answer from the VQA task as the query in the "when" question, aiming to ground the corresponding moment in the video. In our template, we include the total video duration and timestamps of the sampled frames to aid the model in temporal reasoning. Given the "when" question, we standardize the model's output format as [$start\_time$, end_time].

For models that have been trained on temporal groundings, such as Qwen2.5-VL, Video-LLaMA3, TimeChat, VTimeLLM and Trace, we follow their official documentation for formatting outputs. For models without temporal grounding training, we instruct them to output normalized start and end times within the range [0,1]. Finally, we use regular expressions to extract the predicted timestamps from all models and apply denormalization where necessary, ensuring that all final outputs follow the [$start\_time$, end_time] format for metric computation.

For the "where" question in Chain 1, the model is provided with the ground truth (GT) from both the "what" and "when" questions. Within the given time window, the ground truth of "when", the model is tasked with spatially grounding the key object specified in the ground truth of "what". Our prompt template includes video duration and resolution and specifies a time window at a 1 FPS sampling rate, guiding the model on when to localize the key object's bounding box. Each bounding box should be formatted as [x_{min}, y_{min}, x_{max}, y_{max}]. For models that have been trained on spatial grounding, e.g. Gemini-2-Flash, Qwen2-VL, Qwen2.5-VL and Video-LLaMA3, we follow their official documentation for formatting, applying reformatting if necessary. For models without spatial grounding training, we instruct them to output normalized bounding box coordinates within [0,1]. For GPT-4o we require it to output the final coordinates directly. Finally, we require models to output the spatial grounding results in JSON format, where the keys represent timestamps and the values correspond to the bounding boxes. We use regular expressions to extract JSON-formatted spatial-temporal predictions from all models and apply denormalization where necessary, ensuring that all final outputs are in JSON format for metric computation.

We evaluate only the annotated frames. Specifically, videos sourced from VidSTG and GOT-10k provide one bounding box per second, while videos from TVQA+ offer one bounding box every two seconds.

Prompt Templates for Chain 2. For the "what" questions in Chain 2, we use the same prompt template as in Chain 1. Since the inputs are identical, we skip repeating the experiment and directly use the Video-LLM's answer of Chain 1 as the answer of Chain 2.

For "where" questions in Chain 2, the model will only be provided with the ground truth for "what". Therefore, the model will directly perform spatial grounding of the key object mentioned in the "what" GT without the given time range. The model will only be prompted about how many

bounding boxes to predict based on the length of the ground truth of spatial grounding. Specifically, in the template, we will provide the video length, resolution, and the length of the ground truth of "where", and the model must determine at which moments to localize the key object's bounding box. The model's output format remains the same as in Chain1: the model will output a series of bounding boxes in JSON format, with each box formatted as [x_{min}, y_{min}, x_{max}, y_{max}].

For "when" questions in Chain 2, the model will be provided with the GT for both "what" and "where." However, the temporal information has been removed from the "where" GT, retaining only the spatial information of the bounding boxes. Therefore, the model can utilize the given spatial information to determine the exact event timestamps. Specifically, in the template, we will provide the model with information about the video length, the sampling frame timestamps, and the bounding box information from the "where" GT to better determine the moment of occurrence. As with Chain 1, we require the model to output in a uniform format of [$start\_time$, end_time].

C.3. Metrics

we follow the commonly used temporal grounding metrics, "R@n, tIoU=m", which refers to the percentage of top-n prediction with temporal IoU score larger than m, and mean temporal IoU score (m_tIoU). Specifically, "R@n, tIoU=m" is calculate by the Eq.1 as:

$$\langle \text{R@n, tIoU@m} \rangle = \frac{1}{N_q} \sum_{i=1}^{N_q} r(n, m, q_i) \tag{1}$$

where $r(n, m, q_i) = 1$ if at least one of the top-n predictions of query q_i is higher than the tIoU threshold m, and vice versa for 0. m_tIoU is calculated by the Eq.4 as:

$$\text{m\_tIoU} = \frac{1}{N_q} \sum_{i=1}^{N_q} \text{tIoU}_i \tag{2}$$

where N_q denotes the total number of samples and $tIoU_i$ denotes the tIoU score of the i-th sample.

We follow VidSTG [16] and TVQA+ [7] to use the Average Precision score (AP@vIoU=m) and mean visual Intersection over Union (m_vIoU) as the spatial grounding metrics. We only evaluate every annotated frame. Specifically, we calculate vIoU by:

$$\text{vIoU} = \frac{1}{|\mathcal{S}_U|} \sum_{t \in \mathcal{S}_I} \text{IoU}(r^t, \hat{r}^t) \tag{3}$$

where r^t and $\hat{r}^t$ are selected and ground truth regions of frame t. AP@vIoU=m is the average precision score where the $vIoU >= m$. The m_vIoU is the average $vIoU$ of samples, calculated by:

$$\text{m\_vIoU} = \frac{1}{N_q} \sum_{i=1}^{N_q} \text{vIoU}_i \tag{4}$$

where N_q denotes the total number of samples and $vIoU_i$ denotes the vIoU score of the i-th sample.

D. In-depth Experiment Results

We provide more experimental results here, including 6 tables of experiments on different video lengths, and 18 tables of experiments on different domains, as well as 2 more qualitative results.

D.1. Performance on Different Video Length

We provide the full details of the experiments on different video lengths in short($[0, 1]min$), medium($(1, 3]min$) and long($[3, 60]min$) videos. Tab.1 and Tab.2 show the short video results. Tab.3 and Tab.4 show the short video results. Tab.5 and Tab.6 show the short video results. Generally, we can find that GPT-4o and Qwen2.5-VL perform well on short and medium video, but struggle on long video. Gemini-2-flash achieves a balanced performance overall.

D.2. Performance on Different Domains

We provide full details of experiment results on 9 domains, shown in Tab.7 to Tab.24. Generally, each model has its own domain that it works well in, and there is no model that is currently optimal in all domains.

D.3. More Qualitative Results

Here, we provide 2 more qualitative results of the cases sourced from TVQA+ and GOT-10k, shown as Fig.5 and Fig.6 respectively.

E. Limitations

Although this dataset effectively evaluates models' spatial-temporal reasoning abilities, its range of scenarios and task types remains limited, failing to cover all possible video reasoning cases, such as more domain-specific video analysis in areas like healthcare and transportation. In addition, the problem of uneven length of video data is yet to be solved. Obtaining temporal and spatial labels in real long videos instead of synthesized videos is still a challenge.

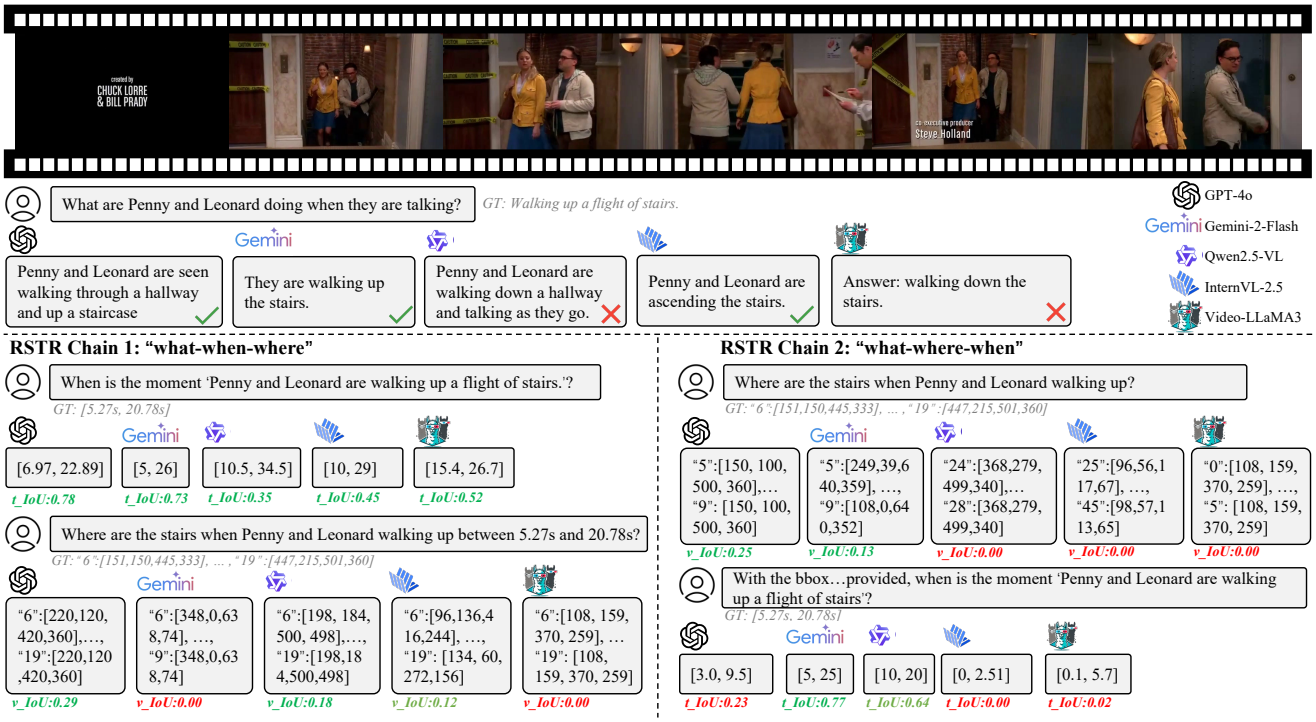

Figure 5. An example showcasing the performance of five models on the video sourced from TVQA+.

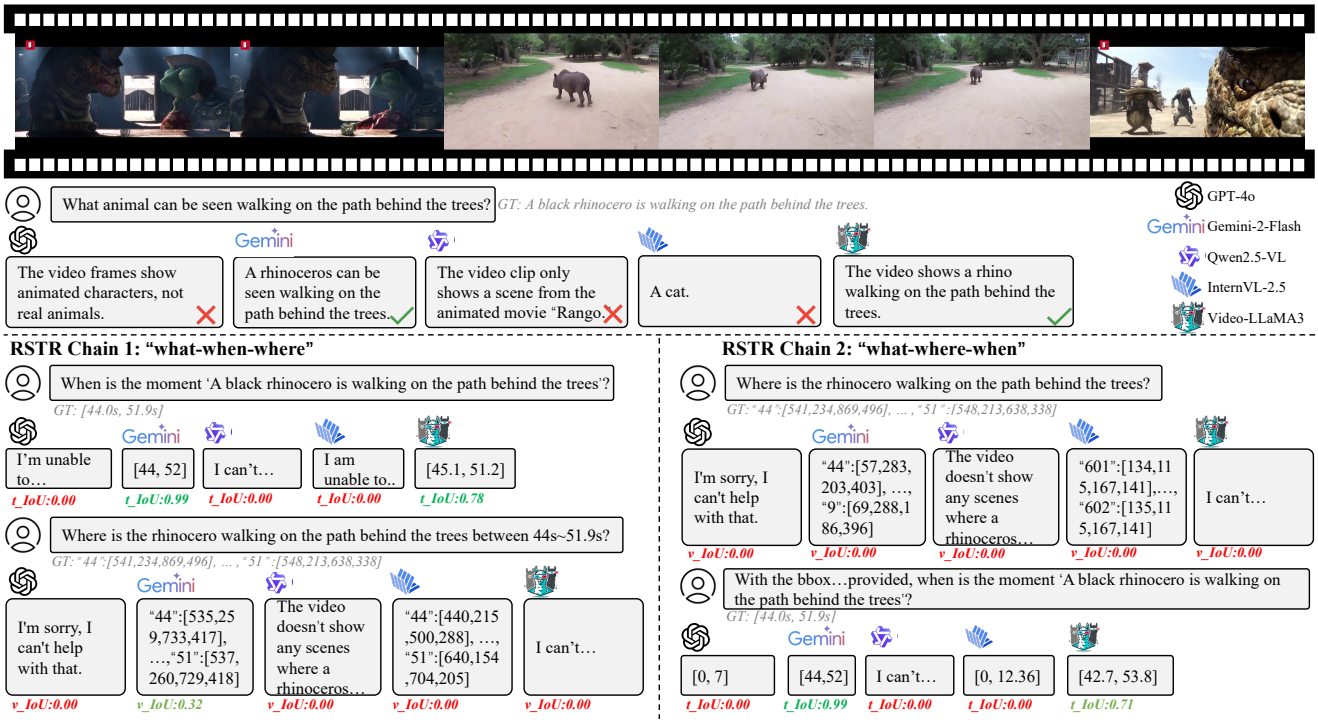

Figure 6. An example showcasing the performance of five models on the video sourced from GOT-10k.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.69	59.82	26.97	12.45	6.19	19.23	24.65	10.21	3.33	7.99	**40.28**	**29.01**
Gemini-2-Flash [4]	-	-	1.50	48.51	29.56	13.29	6.95	22.71	16.80	4.06	1.00	4.92	24.25	25.38
Video-Llama3 [14]	ArXiv'25	7B	1.33	40.49	**37.05**	**22.31**	**10.24**	**24.33**	3.46	0.80	0.09	0.95	26.91	21.92
Qwen2.5-VL [1]	ArXiv'25	7B	1.64	56.38	19.56	11.15	4.97	13.34	40.46	23.44	20.64	15.90	38.20	28.54
Qwen2-VL [12]	ArXiv'24	7B	0.98	22.38	27.81	18.64	8.63	19.03	33.29	14.52	4.44	10.74	19.27	17.39
InternVL-2.5 [2]	ArXiv'24	8B	1.48	44.61	12.91	5.35	2.44	9.40	2.19	0.36	0.04	0.71	23.22	18.24
Llava-Video [15]	ArXiv'24	7B	1.57	53.25	15.97	6.34	0.92	10.98	5.84	1.02	0.20	2.12	29.94	22.12
VideoChat2 [8]	CVPR'24	7B	1.28	37.20	20.78	12.38	6.04	13.50	11.34	1.43	0.12	2.77	21.28	17.83
Oryx-1.5 [9]	ICLR'25	7B	0.89	17.04	18.95	4.05	0.99	14.42	36.20	12.13	2.07	10.28	15.03	13.91
Video-CCAM-v1.2 [3]	ArXiv'24	7B	**1.78**	**63.48**	1.53	0.00	0.00	1.64	-	-	-	-	34.13	21.71
TimeChat [11]	CVPR'24	7B	1.10	27.04	20.09	9.63	3.59	13.34	-	-	-	-	15.28	13.46
VTimeLLM [6]	CVPR'24	7B	1.49	43.70	23.53	8.94	2.83	17.03	0.07	0.01	0.00	0.04	25.38	20.25
TRACE [5]	ICLR'25	7B	0.87	16.27	29.41	13.90	7.03	20.30	-	-	-	-	13.48	12.19
Sa2VA [13]	ArXiv'25	8B	0.67	15.81	0.15	0.00	0.00	0.16	**56.48**	**46.68**	**37.76**	**35.65**	20.49	17.21

Table 1. Performance on short videos on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.69	59.82	10.89	4.57	1.28	3.44	19.48	12.61	8.94	14.66	**36.84**	**25.97**
Gemini-2-Flash [4]	-	-	1.50	48.51	8.39	2.20	0.68	2.47	28.88	11.92	5.81	21.77	31.15	24.25
Video-Llama3 [14]	ArXiv'25	7B	1.33	40.49	0.92	0.23	0.04	0.27	**35.37**	**21.08**	**9.93**	**23.54**	26.34	21.43
Qwen2.5-VL [1]	ArXiv'25	7B	1.64	56.38	6.50	3.84	1.83	2.52	12.38	5.73	2.83	8.52	31.47	22.47
Qwen2-VL [12]	ArXiv'24	7B	0.98	22.38	7.74	4.21	1.16	2.68	24.60	16.81	7.94	17.43	15.74	14.16
InternVL-2.5 [2]	ArXiv'24	8B	1.48	44.61	0.31	0.04	0.01	0.14	11.38	4.13	1.68	8.18	23.22	18.24
Llava-Video [15]	ArXiv'24	7B	1.57	53.25	6.04	1.51	0.15	1.75	18.03	5.65	2.14	12.85	30.52	22.62
VideoChat2 [8]	CVPR'24	7B	1.28	37.20	3.78	1.21	0.46	1.22	20.02	12.91	6.88	13.55	20.77	17.32
Oryx-1.5 [9]	ICLR'25	7B	0.89	17.04	15.15	5.39	1.14	4.49	21.47	5.73	2.67	15.74	13.47	12.42
Video-CCAM-v1.2 [3]	ArXiv'24	7B	**1.78**	**63.48**	-	-	-	-	1.22	0.00	0.00	1.36	34.04	21.62
TimeChat [11]	CVPR'24	7B	1.10	27.04	-	-	-	-	22.23	9.78	3.06	14.78	15.84	13.94
VTimeLLM [6]	CVPR'24	7B	1.49	43.70	0.00	0.00	0.00	0.00	7.26	3.59	2.06	5.38	20.99	16.36
TRACE [5]	ICLR'25	7B	0.87	16.27	-	-	-	-	25.06	11.61	5.81	17.77	12.44	11.35
Sa2VA [13]	ArXiv'25	8B	0.67	15.81	**63.43**	**54.14**	**44.38**	**41.35**	0.00	0.00	0.00	0.02	23.53	19.06

Table 2. Performance on short videos on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.82	64.99	18.73	7.78	3.46	13.89	13.65	6.00	2.03	4.47	41.49	27.78
Gemini-2-Flash [4]	-	-	1.72	58.07	32.28	15.99	9.22	24.99	13.52	3.31	0.91	4.08	39.53	28.65
Video-Llama3 [14]	ArXiv'25	7B	1.44	43.23	34.01	16.28	6.48	21.02	3.03	0.77	0.17	0.90	27.04	21.71
Qwen2.5-VL [1]	ArXiv'25	7B	1.72	56.63	14.55	5.91	1.87	9.50	32.09	15.96	5.19	11.04	35.07	25.72
Qwen2-VL [12]	ArXiv'24	7B	1.10	29.83	31.12	19.02	11.38	21.57	23.38	9.47	3.37	7.78	22.60	19.72
InternVL-2.5 [2]	ArXiv'24	8B	1.48	45.24	11.82	4.61	2.45	8.55	2.32	0.15	0.04	0.59	23.25	18.13
Llava-Video [15]	ArXiv'24	7B	1.39	42.36	15.27	7.06	2.59	10.81	3.83	0.75	0.13	1.45	22.67	18.21
VideoChat2 [8]	CVPR'24	7B	1.31	36.31	22.62	16.14	8.21	15.73	8.60	1.16	0.16	2.19	21.47	18.07
Oryx-1.5 [9]	ICLR'25	7B	1.02	24.06	15.71	5.91	3.31	13.49	36.19	11.24	2.56	10.40	17.67	15.99
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.74	56.05	0.58	0.00	0.00	1.43	-	-	-	-	27.89	19.16
TimeChat [11]	CVPR'24	7B	1.03	27.67	15.71	8.07	3.75	11.03	-	-	-	-	14.69	12.90
VTimeLLM [6]	CVPR'24	7B	1.45	40.78	31.70	15.99	4.18	19.56	0.27	0.04	0.00	0.08	24.74	20.14
TRACE [5]	ICLR'25	7B	0.99	21.04	25.79	13.26	5.33	18.09	-	-	-	-	14.53	13.04
Sa2VA [13]	ArXiv'25	8B	0.81	18.88	0.00	0.00	0.00	0.03	46.79	37.29	29.02	27.63	17.76	15.51

Table 3. Performance on medium videos on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.82	64.99	7.50	3.99	1.18	2.59	14.70	7.35	5.04	10.82	39.67	26.13
Gemini-2-Flash [4]	-	-	1.72	58.07	6.39	1.53	0.47	1.92	34.15	16.71	8.79	24.79	39.11	28.26
Video-Llama3 [14]	ArXiv'25	7B	1.44	43.23	0.14	0.08	0.00	0.06	34.73	19.16	7.49	22.58	27.42	21.96
Qwen2.5-VL [1]	ArXiv'25	7B	1.72	56.63	3.29	2.09	0.78	1.29	9.94	5.48	2.16	6.92	30.67	21.61
Qwen2-VL [12]	ArXiv'24	7B	1.10	29.83	6.88	2.79	1.26	2.23	27.38	17.58	9.94	19.61	19.83	17.22
InternVL-2.5 [2]	ArXiv'24	8B	1.48	45.24	0.68	0.02	0.00	0.18	11.10	3.60	1.59	7.85	22.86	17.76
Llava-Video [15]	ArXiv'24	7B	1.39	42.36	1.57	0.87	0.49	0.65	16.71	5.91	2.02	12.06	22.87	18.36
VideoChat2 [8]	CVPR'24	7B	1.31	36.31	2.13	0.43	0.04	0.64	16.86	12.10	5.76	12.04	19.563	16.33
Oryx-1.5 [9]	ICLR'25	7B	1.02	24.06	6.15	2.90	0.74	2.11	16.86	6.05	3.17	14.87	15.25	13.68
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.74	56.05	-	-	-	-	4.32	0.00	0.00	4.19	28.83	20.08
TimeChat [11]	CVPR'24	7B	1.03	27.67	-	-	-	-	19.45	7.20	1.73	12.98	15.43	13.55
VTimeLLM [6]	CVPR'24	7B	1.45	40.78	0.00	0.00	0.00	0.00	11.53	6.92	2.45	7.70	20.13	16.16
TRACE [5]	ICLR'25	7B	0.99	21.04	-	-	-	-	22.05	10.37	3.60	14.79	13.21	11.94
Sa2VA [13]	ArXiv'25	8B	0.81	18.88	52.75	43.71	34.92	32.52	0.00	0.00	0.00	0.00	20.08	17.13

Table 4. Performance on medium videos on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.25	43.01	2.15	0.00	0.00	1.40	0.00	0.00	0.00	0.00	19.21	14.80
Gemini-2-Flash [4]	-	-	1.75	64.52	55.91	50.54	46.24	47.02	15.82	4.33	0.16	4.56	57.27	38.70
Video-Llama3 [14]	ArXiv'25	7B	1.42	47.31	30.11	10.75	3.23	18.42	0.11	0.00	0.05	0.89	28.16	21.93
Qwen2.5-VL [1]	ArXiv'25	7B	0.17	6.54	0.00	0.00	0.00	0.26	0.10	0.00	0.00	0.03	2.32	2.24
Qwen2-VL [12]	ArXiv'24	7B	1.24	38.71	6.45	0.00	0.00	3.53	1.30	0.00	0.00	0.55	17.70	14.26
InternVL-2.5 [2]	ArXiv'24	8B	1.08	27.96	0.00	0.00	0.00	0.34	0.98	0.00	0.00	0.31	11.15	9.53
Llava-Video [15]	ArXiv'24	7B	1.55	51.61	2.15	0.00	0.00	1.82	7.24	1.11	0.18	2.55	25.67	18.66
VideoChat2 [8]	CVPR'24	7B	0.66	13.98	0.00	0.00	0.00	1.18	2.99	0.81	0.25	1.27	5.84	5.48
Oryx-1.5 [9]	ICLR'25	7B	0.95	31.18	0.00	0.00	0.00	1.47	22.34	7.01	0.62	6.18	15.08	12.95
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.46	37.63	0.00	0.00	0.00	0.01	-	-	-	-	15.74	12.55
TimeChat [11]	CVPR'24	7B	0.62	8.60	1.08	0.00	0.00	0.61	-	-	-	-	3.20	3.07
VTimeLLM [6]	CVPR'24	7B	0.92	13.98	1.08 0.00	0.00	0.36	10.96	2.69	0.73	3.63	6.37	5.99	
TRACE [5]	ICLR'25	7B	0.77	16.13	36.56	24.73	12.90	24.24	-	-	-	-	15.12	13.46
Sa2VA [13]	ArXiv'25	8B	0.31	6.45	0.00	0.00	0.00	0.00	31.38	26.67	22.37	20.10	9.70	8.85

Table 5. Performance on long videos on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.25	43.01	0.00	0.00	0.00	0.00	2.15	2.15	0.00	1.76	19.34	14.92
Gemini-2-Flash [4]	-	-	1.75	64.52	3.03	0.23	0.00	0.66	50.54	50.54	45.16	45.55	55.02	36.91
Video-Llama3 [14]	ArXiv'25	7B	1.42	47.31	0.00	0.00	0.00	0.00	34.41	20.43	11.83	21.66	29.50	22.99
Qwen2.5-VL [1]	ArXiv'25	7B	0.17	6.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.22	2.15
Qwen2-VL [12]	ArXiv'24	7B	1.24	38.71	0.00	0.00	0.00	0.00	4.30	0.00	0.00	3.07	17.36	14.09
InternVL-2.5 [2]	ArXiv'24	8B	1.08	27.96	0.00	0.00	0.00	0.00	1.08	0.00	0.00	0.93	11.24	9.63
Llava-Video [15]	ArXiv'24	7B	1.55	51.61	0.00	0.00	0.00	0.01	2.15	0.00	0.00	1.75	24.79	17.79
VideoChat2 [8]	CVPR'24	7B	0.66	13.98	0.40	0.13	0.00	0.10	0.00	0.00	0.00	1.16	5.44	5.08
Oryx-1.5 [9]	ICLR'25	7B	0.95	31.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.47	12.90	10.83
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.46	37.63	-	-	-	-	0.00	0.00	0.00	0.37	15.86	12.67
TimeChat [11]	CVPR'24	7B	0.62	8.60	-	-	-	-	2.15	1.08	1.08	1.63	3.55	3.41
VTimeLLM [6]	CVPR'24	7B	0.92	13.98	0.00	0.00	0.00	0.00	2.15	0.00	0.00	2.16	5.41	5.04
TRACE [5]	ICLR'25	7B	0.77	16.13	-	-	-	-	35.48	30.11	20.43	25.06	15.48	13.73
Sa2VA [13]	ArXiv'25	8B	0.30	6.54	31.38	26.69	22.37	20.10	0.00	0.00	0.00	0.00	9.70	8.85

Table 6. Performance on long videos on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.75	63.22	27.01	12.07	7.47	18.50	28.29	10.87	4.26	9.14	43.35	30.29
Gemini-2-Flash [4]	-	-	1.57	51.72	36.21	15.52	8.05	25.83	21.65	8.27	1.74	6.93	36.63	28.16
Video-Llama3 [14]	ArXiv'25	7B	1.32	39.08	37.36	25.29	14.37	25.11	4.90	1.58	0.24	1.57	26.68	21.92
Qwen2.5-VL [1]	ArXiv'25	7B	1.69	57.47	18.97	11.49	2.30	12.32	55.57	31.14	11.75	20.56	40.56	30.12
Qwen2-VL [12]	ArXiv'24	7B	1.03	20.69	29.89	20.11	8.62	19.65	49.63	22.22	8.89	16.49	21.03	18.94
InternVL-2.5 [2]	ArXiv'24	8B	1.57	52.87	9.77	3.45	2.30	8.26	2.60	0.68	0.31	0.97	28.28	20.70
Llava-Video [15]	ArXiv'24	7B	1.43	46.55	19.54	11.49	1.72	13.30	12.17	3.17	0.30	4.03	27.01	21.29
VideoChat2 [8]	CVPR'24	7B	1.22	40.23	21.26	14.94	5.75	14.14	6.64	1.51	0.00	1.90	22.88	18.76
Oryx-1.5 [9]	ICLR'25	7B	0.78	13.79	18.97	2.30	0.00	16.01	56.72	20.86	4.52	16.48	16.77	15.43
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.89	68.97	0.00	0.00	0.00	0.37	-	-	-	-	39.13	23.11
TimeChat [11]	CVPR'24	7B	1.37	40.80	14.37	8.05	3.45	11.38	-	-	-	-	21.50	17.39
VTimeLLM [6]	CVPR'24	7B	1.34	31.03	21.26	9.20	4.02	18.02	0.00	0.00	0.00	0.00	19.02	16.36
TRACE [5]	ICLR'25	7B	0.87	14.37	35.06	24.14	14.94	25.63	-	-	-	-	15.04	13.33
Sa2VA [13]	ArXiv'25	8B	0.67	16.09	0.00	0.00	0.00	0.00	71.28	57.79	45.68	43.86	25.09	19.98

Table 7. Performance on Animals Domain on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.75	63.22	12.90	4.41	1.70	3.92	22.41	16.09	12.64	17.06	40.91	28.07
Gemini-2-Flash [4]	-	-	1.57	51.72	12.62	4.89	1.64	4.04	35.63	17.24	8.05	26.44	35.88	27.40
Video-Llama3 [14]	ArXiv'25	7B	1.32	39.08	1.28	0.57	0.29	0.48	33.91	21.84	10.34	23.27	25.51	20.94
Qwen2.5-VL [1]	ArXiv'25	7B	1.69	57.47	10.04	4.56	1.48	3.15	9.20	4.60	1.15	6.92	31.96	22.51
Qwen2-VL [12]	ArXiv'24	7B	1.03	20.69	12.74	6.69	2.36	4.28	27.59	17.82	8.05	17.63	15.65	14.20
InternVL-2.5 [2]	ArXiv'24	8B	1.57	52.87	0.96	0.24	0.05	0.33	15.52	6.32	2.30	10.57	28.91	21.26
Llava-Video [15]	ArXiv'24	7B	1.43	46.55	6.37	1.72	0.55	1.97	10.34	2.87	0.57	9.63	24.92	19.38
VideoChat2 [8]	CVPR'24	7B	1.22	40.23	5.51	2.07	1.53	1.84	18.97	12.64	3.45	12.63	22.28	18.23
Oryx-1.5 [9]	ICLR'25	7B	0.78	13.79	16.05	7.64	1.12	5.12	19.54	1.72	0.57	16.31	16.77	15.43
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.89	68.97	-	-	-	-	4.02	0.00	0.00	2.53	39.86	23.83
TimeChat [11]	CVPR'24	7B	1.37	40.80	-	-	-	-	24.71	9.20	3.45	16.60	23.53	19.14
VTimeLLM [6]	CVPR'24	7B	1.34	31.03	0.00	0.00	0.00	0.00	10.34	2.87	1.72	7.09	14.84	12.71
TRACE [5]	ICLR'25	7B	0.87	14.37	-	-	-	-	28.74	17.82	8.05	19.76	12.51	11.37
Sa2VA [13]	ArXiv'25	8B	0.67	16.09	78.47	67.26	53.47	50.01	0.00	0.00	0.00	0.00	28.96	22.03

Table 8. Performance on Animals Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.62	64.71	25.88	14.12	9.41	20.17	27.71	13.29	2.84	9.11	45.41	31.33
Gemini-2-Flash [4]	-	-	1.39	41.18	27.06	23.53	12.94	22.90	21.30	5.10	2.43	6.39	28.56	23.49
Video-Llama3 [14]	ArXiv'25	7B	1.47	48.24	38.82	25.88	12.94	26.32	1.99	0.86	0.21	0.70	32.36	25.09
Qwen2.5-VL [1]	ArXiv'25	7B	1.58	60.00	18.82	10.59	2.35	12.18	42.93	25.85	7.18	16.00	40.68	29.39
Qwen2-VL [12]	ArXiv'24	7B	0.86	15.29	38.82	30.59	15.29	27.81	35.10	18.69	9.43	13.02	21.04	18.71
InternVL-2.5 [2]	ArXiv'24	8B	1.39	43.53	14.12	9.41	3.53	11.12	2.58	0.08	0.00	0.69	23.21	18.44
Llava-Video [15]	ArXiv'24	7B	1.55	56.47	27.06	14.12	2.53	16.66	6.49	2.77	0.86	2.77	34.74	25.30
VideoChat2 [8]	CVPR'24	7B	1.32	42.35	29.41	14.67	9.41	18.35	6.39	0.00	0.00	1.52	25.63	20.74
Oryx-1.5 [9]	ICLR'25	7B	0.84	8.24	16.47	5.88	3.53	14.88	38.38	15.45	3.43	12.29	12.61	11.80
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.81	77.65	0.00	0.00	0.00	0.00	-	-	-	-	49.94	25.88
TimeChat [11]	CVPR'24	7B	0.94	24.71	20.00	5.88	2.35	12.96	-	-	-	-	14.09	12.56
VTimeLLM [6]	CVPR'24	7B	1.52	47.06	32.94	12.94	2.35	20.42	0.00	0.00	0.00	0.00	28.82	22.50
TRACE [5]	ICLR'25	7B	0.80	11.76	28.24	16.47	10.59	20.17	-	-	-	-	11.68	10.64
Sa2VA [13]	ArXiv'25	8B	0.40	4.71	0.00	0.00	0.00	0.00	57.57	45.30	33.52	32.73	14.82	12.48

Table 9. Performance on Nature Domain on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.62	64.71	16.66	9.79	22.4	5.87	30.59	16.47	12.94	21.67	44.87	30.75
Gemini-2-Flash [4]	-	-	1.39	41.18	12.56	3.92	0.48	3.81	29.41	21.18	11.76	25.07	28.60	23.35
Video-Llama3 [14]	ArXiv'25	7B	1.47	48.24	0.52	0.00	0.00	0.14	32.94	22.35	9.41	22.60	30.53	23.66
Qwen2.5-VL [1]	ArXiv'25	7B	1.58	60.00	7.02	3.31	1.96	2.65	14.12	9.41	5.88	11.62	35.56	24.76
Qwen2-VL [12]	ArXiv'24	7B	0.86	15.29	11.42	7.79	4.22	4.68	31.76	22.35	8.24	22.66	15.70	14.21
InternVL-2.5 [2]	ArXiv'24	8B	1.39	43.53	0.63	0.00	0.00	0.31	8.24	2.35	1.18	7.29	21.68	17.05
Llava-Video [15]	ArXiv'24	7B	1.55	56.47	6.80	4.71	3.61	3.03	14.12	3.53	3.53	12.02	33.02	23.84
VideoChat2 [8]	CVPR'24	7B	1.32	42.35	1.33	0.20	0.00	0.38	21.18	15.29	14.12	16.42	24.47	19.72
Oryx-1.5 [9]	ICLR'25	7B	0.84	8.24	13.51	6.49	2.53	5.11	20.00	10.59	5.88	17.55	11.55	10.30
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.81	77.65	-	-	-	-	3.53	0.00	0.00	1.82	50.55	26.49
TimeChat [11]	CVPR'24	7B	0.94	24.71	-	-	-	-	23.53	5.88	1.18	13.26	14.20	12.66
VTimeLLM [6]	CVPR'24	7B	1.52	47.06	0.00	0.00	0.00	0.00	8.24	7.06	4.71	7.12	23.66	18.06
TRACE [5]	ICLR'25	7B	0.80	11.76	-	-	-	-	23.53	12.94	7.06	17.98	10.78	9.91
Sa2VA [13]	ArXiv'25	8B	0.40	4.71	62.16	51.49	43.96	39.55	0.00	0.00	0.00	0.00	19.39	14.75

Table 10. Performance on Nature Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.68	61.04	15.58	6.49	2.60	16.05	27.30	7.82	2.03	7.86	39.98	28.32
Gemini-2-Flash [4]	-	-	1.57	53.25	32.47	24.68	15.58	29.80	13.25	2.22	0.05	3.32	38.26	28.79
Video-Llama3 [14]	ArXiv'25	7B	1.51	48.05	31.17	22.08	7.79	22.07	1.59	0.24	0.00	0.46	30.30	23.53
Qwen2.5-VL [1]	ArXiv'25	7B	1.60	54.55	15.58	6.49	3.90	10.51	27.92	11.97	6.61	10.40	33.64	25.15
Qwen2-VL [12]	ArXiv'24	7B	1.29	46.75	24.68	18.18	7.79	16.43	18.26	4.33	1.78	5.04	28.71	22.74
InternVL-2.5 [2]	ArXiv'24	8B	1.73	55.84	3.90	0.00	0.00	4.40	0.32	0.00	0.00	0.25	28.83	20.16
Llava-Video [15]	ArXiv'24	7B	1.64	62.34	14.29	6.49	3.90	10.73	5.92	1.20	0.35	2.30	37.11	25.12
VideoChat2 [8]	CVPR'24	7B	1.57	46.75	12.99	9.09	1.30	7.80	7.28	2.60	0.00	2.22	24.47	18.93
Oryx-1.5 [9]	ICLR'25	7B	1.17	41.56	14.29	7.79	1.30	12.63	33.32	12.05	3.02	9.44	25.71	21.21
Video-CCAM-v1.2 [3]	ArXiv'24	7B	2.17	80.52	2.60	0.00	0.00	5.54	-	-	-	-	56.42	28.69
TimeChat [11]	CVPR'24	7B	1.21	29.87	11.69	5.19	2.60	9.10	-	-	-	-	15.01	12.99
VTimeLLM [6]	CVPR'24	7B	1.83	59.74	9.09	7.79	3.90	12.01	1.30	0.72	0.29	0.67	34.82	24.14
TRACE [5]	ICLR'25	7B	1.16	37.66	29.87	15.58	6.49	21.86	-	-	-	-	23.98	19.84
Sa2VA [13]	ArXiv'25	8B	0.75	28.57	0.00	0.00	0.00	0.0	61.21	53.87	44.57	40.54	28.55	23.04

Table 11. Performance on Shows Domain on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.68	61.04	9.95	2.59	0.44	2.75	16.88	11.69	7.79	13.05	37.01	25.61
Gemini-2-Flash [4]	-	-	1.57	53.25	5.16	0.74	0.29	1.53	35.06	25.97	18.18	31.61	38.52	28.80
Video-Llama3 [14]	ArXiv'25	7B	1.51	48.05	0.00	0.00	0.00	0.00	40.26	25.97	12.99	26.17	31.94	24.74
Qwen2.5-VL [1]	ArXiv'25	7B	1.60	54.55	8.02	3.42	0.89	2.41	11.69	7.79	2.60	10.02	30.61	22.32
Qwen2-VL [12]	ArXiv'24	7B	1.29	46.75	5.51	1.75	0.40	1.56	19.48	11.69	5.19	14.56	26.78	20.96
InternVL-2.5 [2]	ArXiv'24	8B	1.73	55.84	0.43	0.00	0.00	0.14	2.60	1.30	1.30	4.12	28.70	20.03
Llava-Video [15]	ArXiv'24	7B	1.64	62.34	3.04	1.92	0.00	1.14	19.48	5.19	2.60	11.36	36.95	24.95
VideoChat2 [8]	CVPR'24	7B	1.57	46.75	6.22	2.37	0.19	1.91	14.29	11.69	2.60	9.22	24.87	19.29
Oryx-1.5 [9]	ICLR'25	7B	1.17	41.56	12.97	4.71	0.65	3.70	12.99	2.60	1.30	11.11	23.09	18.79
Video-CCAM-v1.2 [3]	ArXiv'24	7B	2.17	80.52	-	-	-	-	6.49	0.00	0.00	4.69	56.12	28.40
TimeChat [11]	CVPR'24	7B	1.21	29.87	-	-	-	-	19.48	5.19	0.00	13.66	16.72	13.75
VTimeLLM [6]	CVPR'24	7B	1.83	59.74	0.00	0.00	0.00	0.00	7.79	5.19	0.00	5.42	32.19	21.72
TRACE [5]	ICLR'25	7B	1.16	37.66	-	-	-	-	19.48	14.29	9.09	15.91	21.53	17.86
Sa2VA [13]	ArXiv'25	8B	0.75	28.57	63.01	56.97	46.61	42.57	0.00	0.00	0.00	0.00	29.70	23.72

Table 12. Performance on Shows Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.66	60.46	25.26	11.80	4.97	16.82	22.12	8.77	2.66	6.81	39.42	28.03
Gemini-2-Flash [4]	-	-	1.42	45.76	29.81	15.53	9.52	22.78	18.69	3.39	0.32	5.03	30.72	24.52
Video-Llama3 [14]	ArXiv'25	7B	1.22	33.33	32.51	20.29	10.77	22.30	4.80	1.10	0.08	1.22	22.34	18.95
Qwen2.5-VL [1]	ArXiv'25	7B	1.61	57.35	16.98	10.56	5.18	12.39	46.07	27.44	12.04	18.18	39.50	29.31
Qwen2-VL [12]	ArXiv'24	7B	0.88	14.49	22.98	15.11	8.07	16.10	39.64	18.13	4.20	12.82	15.64	14.47
InternVL-2.5 [2]	ArXiv'24	8B	1.46	44.72	11.59	4.55	2.28	8.65	3.16	0.51	0.01	0.94	23.09	18.10
Llava-Video [15]	ArXiv'24	7B	1.47	45.96	14.49	4.35	1.04	10.29	4.64	0.73	0.00	1.70	24.71	19.32
VideoChat2 [8]	CVPR'24	7B	1.27	35.20	16.98	8.70	4.76	10.81	12.59	1.88	0.32	3.07	19.31	16.36
Oryx-1.5 [9]	ICLR'25	7B	0.78	8.07	14.29	2.07	0.21	11.73	40.69	12.93	1.83	11.35	10.98	10.38
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.67	58.39	1.04	0.00	0.00	1.15	-	-	-	-	29.61	19.85
TimeChat [11]	CVPR'24	7B	1.07	25.05	18.01	6.83	3.11	11.33	-	-	-	-	13.62	12.13
VTimeLLM [6]	CVPR'24	7B	1.58	51.35	19.25	6.21	2.07	13.53	0.44	0.00	0.00	0.13	28.91	21.67
TRACE [5]	ICLR'25	7B	0.77	9.32	24.84	11.80	6.21	17.66	-	-	-	-	9.74	8.99
Sa2VA [13]	ArXiv'25	8B	0.55	7.45	0.00	0.00	0.00	0.17	66.22	57.67	48.39	44.14	22.05	17.25

Table 13. Performance on Daily Life Domain on chain "what-when-where", The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.66	60.46	11.27	5.04	1.08	3.49	15.53	9.73	6.63	11.60	36.22	25.18
Gemini-2-Flash [4]	-	-	1.42	45.76	8.55	2.12	0.74	2.52	26.50	12.63	7.45	20.54	28.90	22.94
Video-Llama3 [14]	ArXiv'25	7B	1.22	33.33	1.08	0.26	0.00	0.30	32.92	20.29	9.11	21.93	21.87	18.52
Qwen2.5-VL [1]	ArXiv'25	7B	1.61	57.35	6.09	3.79	1.47	2.44	13.25	6.42	2.48	8.90	32.33	22.90
Qwen2-VL [12]	ArXiv'24	7B	0.88	14.49	10.38	5.57	1.57	3.58	20.91	13.66	6.63	15.32	11.98	11.13
InternVL-2.5 [2]	ArXiv'24	8B	1.46	44.72	0.22	0.02	0.00	0.11	10.41	3.52	1.24	7.09	22.25	17.70
Llava-Video [15]	ArXiv'24	7B	1.47	45.96	6.52	2.20	0.07	2.05	21.12	6.63	2.07	14.54	26.45	20.85
VideoChat2 [8]	CVPR'24	7B	1.27	35.20	2.64	0.95	0.49	0.91	15.94	9.52	5.80	11.02	18.66	15.71
Oryx-1.5 [9]	ICLR'25	7B	0.78	8.07	17.39	6.82	1.67	5.32	19.46	4.76	1.86	14.37	9.80	9.26
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.67	58.39	-	-	-	-	2.07	0.00	0.00	1.86	29.85	20.08
TimeChat [11]	CVPR'24	7B	1.07	25.05	-	-	-	-	19.67	8.07	2.90	13.45	14.43	12.83
VTimeLLM [6]	CVPR'24	7B	1.58	51.35	0.00	0.00	0.00	0.00	7.25	3.52	2.07	5.64	25.95	19.00
TRACE [5]	ICLR'25	7B	0.77	9.32	-	-	-	-	21.95	9.11	5.38	15.22	8.77	8.18
Sa2VA [13]	ArXiv'25	8B	0.55	7.45	70.50	61.87	52.92	48.11	0.00	0.00	0.00	0.00	24.45	18.52

Table 14. Performance on Daily Life Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.64	56.99	25.91	10.88	4.66	19.05	15.69	5.22	1.68	4.72	36.78	26.92
Gemini-2-Flash [4]	-	-	1.52	48.19	28.50	11.92	10.36	24.42	8.93	1.85	0.24	2.55	32.11	25.05
Video-Llama3 [14]	ArXiv'25	7B	1.41	46.11	35.23	16.58	7.25	21.63	1.34	0.00	0.00	0.25	28.82	22.67
Qwen2.5-VL [1]	ArXiv'25	7B	1.50	53.89	21.76	13.99	8.81	15.28	19.79	9.22	3.57	6.75	33.66	25.30
Qwen2-VL [12]	ArXiv'24	7B	1.05	27.98	27.46	17.10	7.25	18.08	14.60	6.19	1.26	4.62	19.16	16.89
InternVL-2.5 [2]	ArXiv'24	8B	1.55	48.70	11.40	4.66	1.55	8.45	0.31	0.00	0.00	0.15	25.24	19.10
Llava-Video [15]	ArXiv'24	7B	1.53	52.33	12.95	4.66	2.07	9.64	4.29	0.31	0.00	1.48	28.57	21.15
VideoChat2 [8]	CVPR'24	7B	1.17	36.79	22.28	13.47	2.59	12.90	4.22	0.07	0.00	1.05	20.25	16.91
Oryx-1.5 [9]	ICLR'25	7B	0.98	24.35	16.58	1.04	0.00	11.84	22.65	7.69	0.74	6.32	15.68	14.17
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.72	64.77	1.04	0.00	0.00	1.58	-	-	-	-	35.30	22.12
TimeChat [11]	CVPR'24	7B	1.13	29.02	17.62	8.29	2.07	12.50	-	-	-	-	15.87	13.84
VTimeLLM [6]	CVPR'24	7B	1.36	39.90	22.80	5.70	2.07	14.20	0.82	0.18	0.00	0.24	22.15	18.11
TRACE [5]	ICLR'25	7B	0.93	23.83	31.09	14.51	9.33	23.73	-	-	-	-	18.11	15.86
Sa2VA [13]	ArXiv'25	8B	0.73	22.28	1.04	0.00	0.00	0.42	46.80	36.05	26.29	26.07	18.61	16.25

Table 15. Performance on Sports Domain on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.64	56.99	5.72	2.68	1.04	1.97	23.83	18.65	15.03	19.29	35.94	26.09
Gemini-2-Flash [4]	-	-	1.52	48.19	4.18	1.85	0.24	2.25	25.39	11.40	9.33	21.72	30.42	23.64
Video-Llama3 [14]	ArXiv'25	7B	1.41	46.11	0.00	0.00	0.00	0.00	30.57	17.10	10.88	20.35	28.20	22.16
Qwen2.5-VL [1]	ArXiv'25	7B	1.50	53.89	4.09	2.15	0.99	1.55	11.92	5.18	4.15	8.08	291.3	21.17
Qwen2-VL [12]	ArXiv'24	7B	1.05	27.98	4.60	1.76	0.39	1.39	24.35	18.65	10.36	17.99	18.02	15.79
InternVL-2.5 [2]	ArXiv'24	8B	1.55	48.70	0.33	0.08	0.00	0.08	9.84	3.11	2.07	7.07	24.72	18.61
Llava-Video [15]	ArXiv'24	7B	1.53	52.33	3.22	0.81	0.00	0.97	15.54	3.11	0.52	11.23	28.99	21.51
VideoChat2 [8]	CVPR'24	7B	1.17	36.79	1.01	0.00	0.00	0.28	22.80	12.44	5.18	13.50	20.22	16.85
Oryx-1.5 [9]	ICLR'25	7B	0.98	24.35	7.98	2.25	0.53	2.20	23.32	7.77	4.15	17.30	16.38	14.62
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.72	64.77	-	-	-	-	0.00	0.00	0.00	0.83	35.05	21.86
TimeChat [11]	CVPR'24	7B	1.13	29.02	-	-	-	-	18.13	6.74	1.04	11.27	15.41	13.43
VTimeLLM [6]	CVPR'24	7B	1.36	39.90	0.00	0.00	0.00	0.00	3.11	1.55	1.04	2.51	19.90	17.82 14.13
TRACE [5]	ICLR'25	7B	0.93	23.83	-	-	-	-	25.91	13.47	8.81	20.46	16.71	14.76
Sa2VA [13]	ArXiv'25	8B	0.73	22.28	53.24	43.71	31.93	31.01	0.00	0.00	0.00	0.00	17.78	20.76

Table 16. Performance on Sports Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.98	71.09	18.36	9.38	4.49	14.18	11.44	5.17	1.13	3.71	47.73	29.66
Gemini-2-Flash [4]	-	-	1.98	70.31	38.09	17.19	8.98	28.21	7.38	1.54	0.27	2.35	52.32	33.62
Video-Llama3 [14]	ArXiv'25	7B	1.51	46.48	46.48	20.70	5.08	26.99	1.41	0.58	0.09	0.50	31.49	24.66
Qwen2.5-VL [1]	ArXiv'25	7B	1.65	50.78	15.43	6.25	1.37	9.67	18.89	8.61	2.65	6.21	29.16	22.22
Qwen2-VL [12]	ArXiv'24	7B	1.21	35.35	39.65	25.00	13.67	26.83	8.82	1.86	0.23	2.25	25.81	21.58
InternVL-2.5 [2]	ArXiv'24	8B	1.35	37.30	12.89	5.86	3.12	8.80	1.85	0.17	0.00	0.39	18.76	15.50
Llava-Video [15]	ArXiv'24	7B	1.63	50.78	15.04	7.23	1.56	9.68	2.13	0.28	0.00	0.89	27.32	20.45
VideoChat2 [8]	CVPR'24	7B	1.33	33.79	28.32	20.31	12.30	20.58	10.66	1.02	0.07	2.58	22.29	18.98
Oryx-1.5 [9]	ICLR'25	7B	1.17	33.59	12.50	6.05	2.15	10.64	23.13	6.52	1.14	6.18	19.52	16.81
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.66	47.46	0.39	0.00	0.00	0.94	-	-	-	-	21.77	16.13
TimeChat [11]	CVPR'24	7B	0.89	19.73	21.68	13.48	5.27	14.50	-	-	-	-	12.55	11.41
VTimeLLM [6]	CVPR'24	7B	1.36	32.62	37.89	17.38	1.76	22.37	0.55	0.23	0.09	0.20	21.67	18.40
TRACE [5]	ICLR'25	7B	1.04	22.85	32.62	14.25	3.52	20.20	-	-	-	-	16.17	14.35
Sa2VA [13]	ArXiv'25	8B	1.00	24.02	0.00	0.00	0.00	0.00	32.26	23.84	17.33	17.38	14.28	12.75

Table 17. Performance on Entertainments Domain on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.98	71.09	5.35	2.25	0.26	1.66	14.06	5.08	1.56	9.07	45.10	27.27
Gemini-2-Flash [4]	-	-	1.98	70.31	5.08	1.07	0.00	1.36	40.43	18.36	8.98	27.56	51.69	33.08
Video-Llama3 [14]	ArXiv'25	7B	1.51	46.48	0.21	0.02	0.00	0.00	45.70	23.63	7.23	27.68	31.67	24.75
Qwen2.5-VL [1]	ArXiv'25	7B	1.65	50.78	1.69	0.71	0.26	0.55	8.01	4.69	1.17	5.11	25.56	18.81
Qwen2-VL [12]	ArXiv'24	7B	1.21	35.35	1.61	0.28	0.06	0.42	33.98	23.63	12.30	24.10	23.83	19.92
InternVL-2.5 [2]	ArXiv'24	8B	1.35	37.30	0.48	0.00	0.00	0.12	10.94	4.69	2.15	8.09	18.42	15.34
Llava-Video [15]	ArXiv'24	7B	1.63	50.78	0.42	0.10	0.00	0.12	13.28	6.25	1.76	9.20	26.89	20.04
VideoChat2 [8]	CVPR'24	7B	1.33	33.79	4.44	1.02	0.22	1.33	24.41	18.95	9.18	17.56	20.63	17.56
Oryx-1.5 [9]	ICLR'25	7B	1.17	33.59	4.32	1.10	0.00	1.09	12.89	7.03	4.30	11.72	18.17	15.47
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.66	47.46	-	-	-	-	0.78	0.00	0.00	2.66	22.35	16.71
TimeChat [11]	CVPR'24	7B	0.89	19.73	-	-	-	-	19.53	9.38	2.34	13.08	12.00	10.94
VTimeLLM [6]	CVPR'24	7B	1.36	32.62	0.00	0.00	0.00	0.00	10.74	5.08	0.59	6.11	15.26	12.91
TRACE [5]	ICLR'25	7B	1.04	22.85	-	-	-	-	27.93	12.89	4.30	18.01	15.27	13.62
Sa2VA [13]	ArXiv'25	8B	1.00	24.02	32.26	23.84	17.33	17.38	0.00	0.00	0.00	0.00	15.52	13.80

Table 18. Performance on Entertainments Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.66	58.42	26.73	15.84	7.92	20.49	20.55	9.16	3.05	7.08	39.34	28.66
Gemini-2-Flash [4]	-	-	1.43	47.52	27.72	17.82	8.91	25.77	15.63	3.38	1.98	4.84	33.08	26.05
Video-Llama3 [14]	ArXiv'25	7B	1.40	44.55	39.60	23.76	9.90	24.18	1.85	0.43	0.00	0.56	29.07	23.10
Qwen2.5-VL [1]	ArXiv'25	7B	1.58	54.46	21.78	9.90	4.95	14.98	43.36	25.46	8.57	16.52	37.64	28.56
Qwen2-VL [12]	ArXiv'24	7B	1.06	28.71	18.81	10.89	5.94	13.57	33.42	15.09	6.15	11.05	20.04	17.78
InternVL-2.5 [2]	ArXiv'24	8B	1.56	50.50	13.86	4.95	1.98	9.44	1.89	0.00	0.00	0.54	26.92	20.16
Llava-Video [15]	ArXiv'24	7B	1.54	59.41	17.82	5.94	0.00	11.26	7.27	1.24	0.50	2.37	34.83	24.35
VideoChat2 [8]	CVPR'24	7B	1.23	35.64	18.81	12.87	9.90	11.93	10.85	0.69	0.00	2.50	19.77	16.69
Oryx-1.5 [9]	ICLR'25	7B	0.94	20.79	17.82	5.94	1.98	14.58	33.00	9.46	1.13	8.96	16.15	14.87
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.93	74.26	2.97	0.00	0.00	3.46	-	-	-	-	46.41	25.91
TimeChat [11]	CVPR'24	7B	1.23	33.66	14.85	7.92	3.96	12.83	-	-	-	-	18.26	15.50
VTimeLLM [6]	CVPR'24	7B	1.48	45.54	22.77	7.92	3.96	14.65	0.00	0.00	0.00	0.07	25.56	20.09
TRACE [5]	ICLR'25	7B	0.93	20.79	24.75	13.86	13.86	18.46	-	-	-	-	14.57	13.09
Sa2VA [13]	ArXiv'25	8B	0.79	24.75	0.00	0.00	0.00	0.00	53.37	46.12	37.35	34.20	23.43	19.65

Table 19. Performance on Vehicle Domains on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.66	58.42	9.58	5.61	2.29	3.68	15.84	8.91	8.91	20.49	35.25	25.12
Gemini-2-Flash [4]	-	-	1.43	47.52	6.93	0.80	0.00	1.67	21.78	11.88	5.94	21.45	30.10	23.55
Video-Llama3 [14]	ArXiv'25	7B	1.40	44.55	0.00	0.00	0.00	0.07	40.59	20.79	11.88	24.99	29.27	23.20
Qwen2.5-VL [1]	ArXiv'25	7B	1.58	54.46	2.23	1.90	1.19	1.22	11.88	2.97	2.97	7.98	29.40	21.22
Qwen2-VL [12]	ArXiv'24	7B	1.06	28.71	6.96	2.85	1.12	2.28	18.81	14.85	8.91	14.15	17.14	15.05
InternVL-2.5 [2]	ArXiv'24	8B	1.56	50.50	0.31	0.00	0.00	0.08	17.82	5.94	0.99	10.84	27.29	20.47
Llava-Video [15]	ArXiv'24	7B	1.54	59.41	5.37	0.00	0.00		0.96	11.88	4.95	1.98	9.15	33.57
23.17														
VideoChat2 [8]	CVPR'24	7B	1.23	35.64	1.20	0.00	0.00	0.28	15.84	7.92	5.94	9.54	18.09	15.12
Oryx-1.5 [9]	ICLR'25	7B	0.94	20.79	14.36	4.03	0.99	4.03	17.82	8.91	1.98	15.41	14.72	13.41
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.93	74.26	-	-	-	-	1.98	0.00	0.00	1.56	45.76	25.27
TimeChat [11]	CVPR'24	7B	1.23	33.66	-	-	-	-	17.82	4.95	0.99	12.62	18.18	15.43
VTimeLLM [6]	CVPR'24	7B	1.48	45.54	0.00	0.00	0.00	0.00	1.98	0.99	0.99	2.88	21.23	16.14
TRACE [5]	ICLR'25	7B	0.93	20.79	-	-	-	-	24.75	13.86	6.93	17.19	14.06	12.66
Sa2VA [13]	ArXiv'25	8B	0.79	24.75	68.39	59.57	48.11	44.45	0.00	0.00	0.00	0.00	29.08	23.07

Table 20. Performance on Vehicle Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.55	50.58	25.17	8.55	4.16	17.39	24.75	11.75	4.93	8.71	32.90	22.56
Gemini-2-Flash [4]	-	-	1.39	42.96	25.17	9.93	5.08	18.91	21.97	6.29	2.05	6.73	28.02	22.87
Video-Llama3 [14]	ArXiv'25	7B	1.32	40.18	25.40	15.24	8.55	18.26	4.64	0.81	0.20	1.27	24.28	19.90
Qwen2.5-VL [1]	ArXiv'25	7B	1.68	56.58	16.40	7.62	3.46	10.78	45.27	25.88	13.01	18.02	38.24	28.46
Qwen2-VL [12]	ArXiv'24	7B	0.93	23.09	21.25	12.70	6.47	15.02	39.12	17.02	6.07	12.95	18.80	17.02
InternVL-2.5 [2]	ArXiv'24	8B	1.49	43.88	14.09	5.08	2.31	9.83	2.63	0.26	0.05	0.85	22.99	18.19
Llava-Video [15]	ArXiv'24	7B	1.38	45.27	12.93	4.85	1.15	10.10	6.26	0.93	0.40	2.35	24.43	29.24
VideoChat2 [8]	CVPR'24	7B	1.23	35.57	15.01	9.01	3.70	10.19	11.82	1.56	0.18	2.92	19.22	16.22
Oryx-1.5 [9]	ICLR'25	7B	0.84	15.47	26.79	6.93	4.16	19.42	43.80	14.33	3.37	12.99	17.44	15.96
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.81	61.20	2.31	0.00	0.00	2.22	-	-	-	-	32.31	21.14
TimeChat [11]	CVPR'24	7B	1.07	27.94	17.09	7.62	3.00	10.98	-	-	-	-	14.80	12.97
VTimeLLM [6]	CVPR'24	7B	1.46	42.03	23.09	12.93	6.24	17.89	0.04	0.00	0.00	0.04	24.76	19.99
TRACE [5]	ICLR'25	7B	0.86	17.09	23.56	10.39	3.93	16.64	-	-	-	-	12.31	11.24
Sa2VA [13]	ArXiv'25	8B	0.55	13.39	0.00	0.00	0.00	0.16	58.71	48.40	39.63	37.73	20.64	17.10

Table 21. Performance on Indoor on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.55	50.58	11.05	5.28	2.05	3.74	16.40	11.09	8.08	13.05	29.42	22.45
Gemini-2-Flash [4]	-	-	1.39	42.96	8.05	2.24	1.11	2.63	27.94	9.47	3.70	19.51	26.83	21.70
Video-Llama3 [14]	ArXiv'25	7B	1.32	40.18	0.94	0.29	0.00	0.26	26.10	16.86	9.70	19.85	24.59	24.43
Qwen2.5-VL [1]	ArXiv'25	7B	1.68	56.58	6.90	4.31	3.00	3.11	12.47	5.31	3.23	8.58	31.85	22.76
Qwen2-VL [12]	ArXiv'24	7B	0.93	23.09	8.94	4.68	1.43	3.11	19.17	10.16	5.54	13.35	14.59	13.19
InternVL-2.5 [2]	ArXiv'24	8B	1.49	43.88	0.42	0.00	0.00	0.17	11.09	2.77	1.15	7.73	22.00	17.72
Llava-Video [15]	ArXiv'24	7B	1.38	45.27	5.89	1.08	0.23	1.66	21.94	6.47	3.23	15.77	26.37	20.90
VideoChat2 [8]	CVPR'24	7B	1.23	35.57	2.31	0.80	0.00	0.77	12.70	7.85	4.39	9.10	18.09	15.15
Oryx-1.5 [9]	ICLR'25	7B	0.84	15.47	12.85	4.90	1.25	4.08	26.33	4.62	2.08	18.38	13.76	12.65
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.81	61.20	-	-	-	-	3.46	0.00	0.00	2.69	32.31	21.14
TimeChat [11]	CVPR'24	7B	1.07	27.94	-	-	-	-	23.09	11.09	3.70	15.34	16.48	14.43
VTimeLLM [6]	CVPR'24	7B	1.46	42.03	0.00	0.00	0.00	0.00	10.62	7.62	4.85	8.15	21.01	16.73
TRACE [5]	ICLR'25	7B	0.86	17.09	-	-	-	-	20.32	8.55	3.46	14.65	11.53	10.58
Sa2VA [13]	ArXiv'25	8B	0.55	13.39	68.83	59.30	49.50	45.87	0.00	0.00	0.00	0.00	25.28	19.78

Table 22. Performance on Indoor Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		When (Temporal Grounding)				Where (Spatial Grounding)				LGM	AM
			Score	Acc	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}		
GPT-4o [10]	-	-	1.24	47.37	5.26	0.00	0.00	3.05	0.00	0.00	0.00	0.00	22.43	16.80
Gemini-2-Flash [4]	-	-	1.76	63.16	55.26	50.00	47.37	46.31	16.41	3.32	0.38	4.50	55.55	37.99
Video-Llama3 [14]	ArXiv'25	7B	1.61	55.26	36.84	15.79	2.63	18.97	3.17	0.00	0.00	0.00	33.83	24.75
Qwen2.5-VL [1]	ArXiv'25	7B	0.32	7.89	0.00	0.00	0.00	0.37	1.20	0.00	0.00	0.29	2.96	2.85
Qwen2-VL [12]	ArXiv'24	7B	1.26	44.74	10.53	2.63	2.63	7.38	2.49	0.00	0.00	1.03	22.67	17.72
InternVL-2.5 [2]	ArXiv'24	8B	1.03	26.32	0.00	0.00	0.00	0.73	0.26	0.00	0.00	0.14	10.47	9.06
Llava-Video [15]	ArXiv'24	7B	1.61	60.53	7.89	2.63	0.00	5.25	6.33	1.17	0.00	2.20	33.52	22.66
VideoChat2 [8]	CVPR'24	7B	0.71	15.79	7.89	7.89	0.00	5.64	6.95	2.61	0.60	2.84	8.63	8.09
Oryx-1.5 [9]	ICLR'25	7B	0.97	26.32	0.00	0.00	0.00	1.86	18.79	5.85	1.30	5.79	12.80	11.32
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.61	44.74	0.00	0.00	0.00	0.0	-	-	-	-	19.77	14.91
TimeChat [11]	CVPR'24	7B	0.66	15.79	2.63	0.00	0.00	0.88	-	-	-	-	6.02	5.56
VTimeLLM [6]	CVPR'24	7B	0.95	21.05	7.89	2.63	0.00	3.73	13.68	2.03	0.00	3.86	8.66	7.79
TRACE [5]	ICLR'25	7B	0.84	18.42	42.11	26.32	10.53	26.41	-	-	-	-	17.01	14.92
Sa2VA [13]	ArXiv'25	8B	0.45	13.16	0.00	0.00	0.00	0.00	31.67	26.84	23.35	19.58	11.97	10.91

Table 23. Performance on Tutorial Domain on chain "what-when-where". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

Model	Venue	Parameters	What (VQA)		Where (Spatial Grounding)				When (Temporal Grounding)				LGM	AM
			Score	Acc	AP@0.1	AP@0.3	AP@0.5	m_{vIoU}	R1@0.3	R1@0.5	R1@0.7	m_{tIoU}		
GPT-4o[10]	-	-	1.24	47.43	0.00	0.00	0.00	0.00	2.63	2.63	0.00	2.28	22.17	16.55
Gemini-2-Flash [4]	-	-	1.76	63.16	8.00	0.57	0.00	1.82	55.26	55.26	50.00	51.01	57.68	38.66
Video-Llama3 [14]	ArXiv'25	7B	1.61	55.26	0.00	0.00	0.00	0.00	31.58	13.16	2.63	18.56	33.66	24.61
Qwen2.5-VL [1]	ArXiv'25	7B	0.32	7.89	0.48	0.00	0.00	0.10	0.00	0.00	0.00	0.00	2.78	2.67
Qwen2-VL [12]	ArXiv'24	7B	1.26	44.74	0.00	0.00	0.00	0.00	5.26	2.63	0.00	5.88	21.79	16.87
InternVL-2.5 [2]	ArXiv'24	8B	1.03	26.32	0.00	0.00	0.00	0.00	2.63	0.00	0.00	2.60	11.60	9.64
Llava-Video [15]	ArXiv'24	7B	1.61	60.53	0.00	0.00	0.00	0.00	5.26	0.00	0.00	3.22	32.11	21.28
VideoChat2 [8]	CVPR'24	7B	0.71	15.79	0.99	0.33	0.00	0.24	0.00	0.00	0.00	2.25	6.57	6.10
Oryx-1.5 [9]	ICLR'25	7B	0.97	26.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.74	10.77	9.35
Video-CCAM-v1.2 [3]	ArXiv'24	7B	1.61	44.74	-	-	-	-	0.00	0.00	0.00	0.71	20.01	15.15
TimeChat [11]	CVPR'24	7B	0.66	15.79	-	-	-	-	5.26	2.63	2.63	3.85	7.04	6.55
VTimeLLM [6]	CVPR'24	7B	0.95	21.05	0.00	0.00	0.00	0.00	5.26	0.00	0.00	2.31	8.66	7.79
TRACE [5]	ICLR'25	7B	0.84	18.42	-	-	-	-	44.74	31.58	15.79	27.87	17.68	15.43
Sa2VA [13]	ArXiv'25	8B	0.45	13.16	31.67	26.84	23.35	19.58	0.00	0.00	0.00	0.00	11.97	10.91

Table 24. Performance on Tutorial Domain on "what-where-when". The top result is highlighted in **bold**. "-" denotes a model failed to generate formatted answers. The score ranges from 0 to 4.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7, 8, 9, 10, 11, 12, 13

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[3] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[4] Google. Google, gemini-2-flash. Technical report, Google, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[5] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[6] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[7] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, 2020. 4

[8] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[9] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[10] OpenAI. Openai, gpt-40. Technical report, March 2024. 6, 7, 8, 9, 10, 11, 12, 13

[11] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[13] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 6, 7, 8, 9, 10, 11, 12, 13

[14] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 6, 7, 8, 9, 10, 11, 12, 13

[15] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 6, 7, 8, 9, 10, 11, 12, 13

[16] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 4