Online Video Understanding: OVBench and VideoChat-Online

Supplementary Material

1. Implement Details for Ablation Study

In this section, we describe the ablation experiments, focusing on the implementation details of comparison under different conditions.

1.1. Memory Bank Design

The memory bank consists of three main modules: *temporal memory* (m_t), *main memory* (m_{main}), and *spatial memory* (m_s). Each module stores a different number of frames and processes a distinct number of tokens per frame. The configuration is as follows:

- **Temporal memory** (m_t): 12 frames, 16 tokens per frame.
- **Main memory** (m_{main}): 2 frames, 64 tokens per frame.
- **Spatial memory** (m_s): 2 frames, 256 tokens per frame.

Model	**OVBench(%)**	**VRAM Usage**
InternVL2-4B [5]	44.1	*oom*
MovieChat [14]	30.9	16.90 GB
Flash-VStream [19]	31.2	16.03 GB
Ours	**54.9**	**8.71 GB**

Table 1. Comparison of VARM evaluation results with state-of-the-art (SoTA) methods. Following the settings in Flash-VStream, we use 1000 video frames as input for VARM evaluation. Our method outperforms others in both OVBenchaccuracy and VRAM efficiency.

Dataset	OVBench
Online Data SFT	48.1
w/o dense captioning	47.0
w/o step localization	46.7
w/o spatial temporal action detection	44.8
w/o temporal grounding	45.4

Table 2. Ablation of the online instruction fine-tuning dataset by task category. For simplicity, we only use the online dataset for instruction fine-tuning for 1 epoch based on InternVL2-4B [5].

Memory Bank Capacity			OVBench
m_t	m_{main}	m_s	Overall
18	3	3	54.4
24	4	4	54.4
12	2	2	**54.9**

Table 3. The impact of Memory Bank Capacity on performance

Vision Encoder	LLM BackBone	Scale	OVBench
InternViT-300M-448px [5]	Qwen2-0.5B-Instruct [16]	1B	44.6
	InternLM2-chat-1.8b [2]	2B	43.4
	Phi-3-mini-3.8B [1]	4B	44.1
	InternLM2.5-chat-7b [2]	8B	**48.7**

Table 4. Performance of models at different scales

Total Computational Overhead: The total computational overhead for processing all memory modules is **832 tokens**, calculated as:

$$\text{Total Tokens} = (12\times16)+(2\times64)+(2\times256) = 832 \text{ tokens}.$$

This setup represents the **baseline model**. Subsequent experiments evaluate the impact of removing each memory module and redistributing the computational load to the remaining ones while maintaining the same overall computational budget.

Conditions for Removing Memory Modules:

- **w/o** m_t: Temporal memory is removed, and the load is shifted to the main memory.
- **w/o** m_{main}: Main memory is removed, and the load is redistributed to the temporal memory.
- **w/o** m_s: Spatial memory is removed, and the load is transferred to the main memory.

1.2. Memory Updating Policy

In this section, we present the implementation details of different baselines in the Update Policy.

Token merge: from MovieChat [14]: When the capacity of any memory module is full, the adjacent frames with the greatest similarity are merged.

FIFO: When any memory module is full, the frame with the earliest timestamp is evicted.

Uniform Sample: The video clips ending at the current problem timestamp are uniformly sampled, and the number of frames corresponds to the inherent capacity of each memory module.

w/o Compression: No memory compression, input all frames at fps=2.

1.3. Training Paradigm

We evaluate the training strategy under the sliding window setting without introducing a memory bank for simplicity.

Progressive training strategy. If the progressive training strategy is adopted, the online data is introduced in the second epoch for joint training with online data, otherwise, the online data is introduced in the first epoch for joint

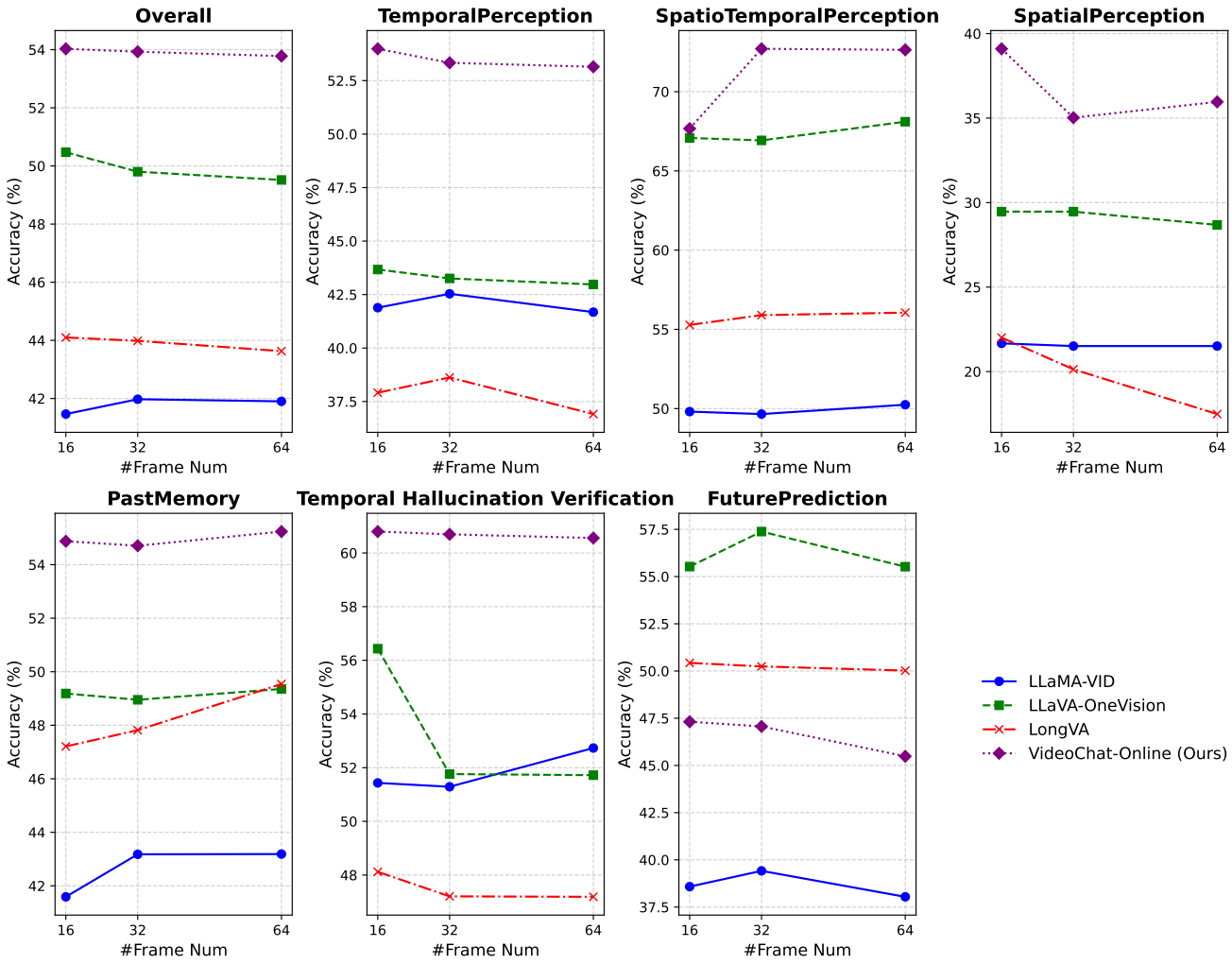


Figure 1. Models' performance in various subtasks and overall performance with varying input frames.

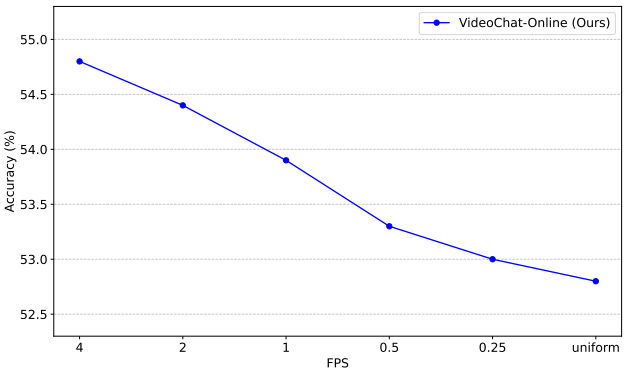


Figure 2. The impact of fps on model performance under the sliding window setting with 64 frames input.

training. Compared with the performance obtained by directly using joint training (51.84%), the performance obtained by progressive training (53.89%) is significantly improved (+2.05%).

Non-interleaved data organization. we train each query as an independent QA sample pair in the original interleaved form and keep the rest of the settings unchanged. The performance obtained by training with interleaved data (53.89%) is better than that obtained by training with non-interleaved data (52.42%), an improvement of 1.47%.

2. More Benchmark Results

For the VideoLLM-Online's evaluation, we provide more detailed results in Table 6. It cannot correctly generate answer options based on the questions, or the specific content related to the options.

Efficiency comparison. We compared the efficiency with our baseline model InternVL2-4B in Figure 4 and the

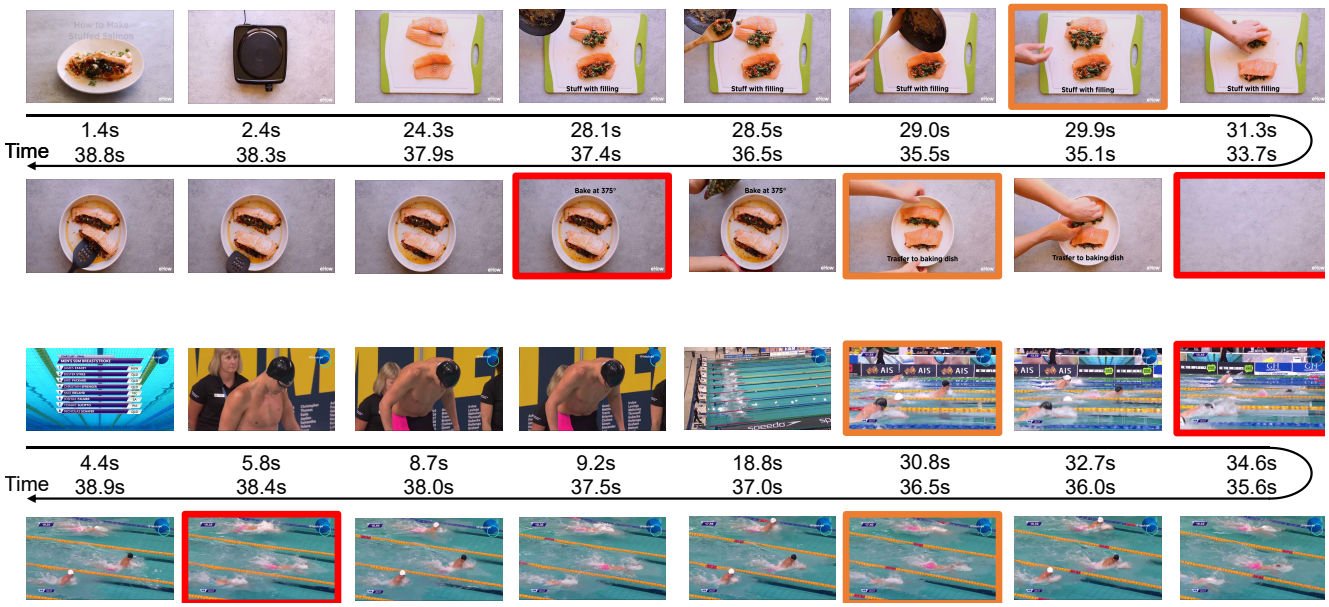


Figure 3. Visualization of the hierarchical memory bank. Frames in the m_s layer are highlighted in red, those in the main layer m_{main} are in orange, and the remaining frames belong to the m_t layer. The structure illustrates the different capacities allocated to each layer.

Task Categories	**Source**	**Domain**	**QA Generation Protocol**
• Action Discrepancy • Action Localization • Action Retrieval • Action Anticipation • Action Sequence • Action Trajectory	AVA [9]	Movie	**Question Requirements:** • Minimum 6 possible options available • Video context: max(900s, t_{query} - 120s) • Continuous frame sequences only **Answer Generation:** • Same video, different timestamps • Task-specific typical answers • Random select answers
• Step Verification • Procedure Recall • Goal/Step Prediction • Step Localization	HiREST [18] COIN [15]	Instructional Indoor Activities Open-Domain	**Question Requirements:** • Minimum 6 options available • Video context: max(0s, t_{query} - 300s) • Clear step descriptions only **Option Generation:** • Intra-video temporal alternatives • Similar topic cross-video options • Task-specific typical answers • Step Duration $\leq$ 5s • 3 $\leq$ Number of Steps $\leq$ 10
• Object Presence • Object Position • Trajectory Retrieval • Movement Prediction • Object State • Object Trajectory	TAO [6] HACS [21] ArgoVerse [3] BDD [17] LaSOT [8] AVA [9]	Road Scene Indoor Activities Outdoor Activities Open-Domain	**Question Criteria:** • Specific object class labeling • No ambiguous object class (e.g. "maybe" "unknown") **Answer Construction:** • Temporal consistency with question • Class-consistent trajectories • if use template: 3×3 grid-based position mapping • Task-appropriate typical responses

Table 5. Task Categories and Question-Answer Generation Strategy

Question at 9.0s	**What is the time period the pillow [0.725, 0.483, 0.991, 0.736] appears in the video? When does it disappear?**
Options:	(A) Appears: 2.0 - 4.0s, 6.0s, 8.0s; Disappears: 5.0s, 7.0s, 9.0s. (B) Appears: 1.0 - 4.0s, 6.0 - 8.0s; Disappears: 5.0s, 9.0s. (C) Appears: 1.0s; Disappears: 2.0 - 9.0s. (D) Appears: 7.0 - 9.0s.
Answer:	**Response:** Appears: 2.0 - 4.0s, 6.0s, 8.0s; Disappears: 5.0s, 7.0s, 9.0s. **Ground Truth:** D **Task Type:** TemporalPerception
Question at 17.0s	**When does the pillow [0.477, 0.443, 0.695, 0.61] first appear in the video? What is the position?**
Options:	(A) 8 seconds before: [0.391, 0.31, 0.587, 0.626]. (B) 2 seconds before: [0.375, 0.244, 0.472, 0.829]. (C) 10 seconds before: [0.855, 0.626, 1.0, 1.0]. (D) 25 seconds before: [0.354, 0.243, 0.691, 0.624].
Answer:	**Response:** The pillow first appears at 8.391s. **Ground Truth:** C **Task Type:** PastMemory

Table 6. More detailed information about the VideoLLM-Online test. It is not able to correctly generate answer options based on the questions, or the specific content related to the options.

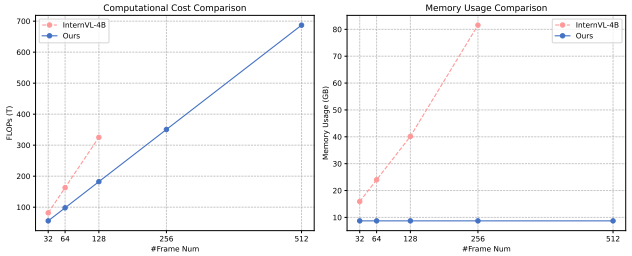


Figure 4. Comparison of computational cost and memory usage between baseline model (InternVL2-4B [5]) and our method.

existing state of art model in Table 1, highlighting the efficiency advantages of our model.

Qualitative comparison. We provide a qualitative comparison with other online models in Figure 9. Including TimeChat [13] and VTimeLLM [10], which are time-sensitive models, and Flash-VStream[19], VideoLLM-Online[4] and MovieChat[14], which can receive streaming input.

3. More Ablations

3.1. Hierarchical Memory Bank Visualization

Figure 3 provides a visualization example of the proposed hierarchical memory bank, where frames in the m_s layer are marked in red, those in the main layer are marked in orange, and the others belong to the m_t layer. The corresponding capacity of the memory bank, $[m_t, m_{\text{main}}, m_s]$, is $[12, 2, 2]$.

3.2. Online SFT Data

Table 2 shows the ablation experimental results of the online instruction fine-tuning dataset under different task categories. For simplicity, in the experiment, only 1 epoch of instruction fine-tuning was performed using the online dataset. As can be seen from the results, the complete online dataset (Online Data SFT) achieved the highest performance on OVBench(48.1). The performance of each task module decreased after removal, with the removal of w/o dense captioning and w/o step localization decreasing by 1.1 and 1.4 percentage points, respectively. However, the removal of w/o spatial-temporal action detection and w/o temporal grounding resulted in more significant performance degradation, with a decrease of 3.3 and 2.7 percentage points, respectively.

3.3. Model

Memory capacity. As shown in the Table 3, we expand the memory capacity to 3, 3, 18 and 4, 4, 24, and the existing memory capacity settings maintain optimal performance and computational efficiency.

3.4. Benchmark

The impact of fps. We fixed the sampling frame number to 64, selected fps as $\{0.25, 0.5, 1, 2, 4\}$, and uniform sampling (from the beginning of the video to the question times-

Object Tracking	
(t=60.0s) User	Using the description "tiger walking on the dry grass," track the object's past trajectory until the current frame and then display each future frame sequentially.
(t=60.0s) Assistant	<Frames During 0.0s - 60.0s> At 0.0s, [0.439, 0.349, 0.486, 0.443]; ... At 33.0s, [0.619, 0.246, 0.672, 0.367]; At 34.0s, [0.634, 0.218, 0.69, 0.357]; ... At 60.0s, [0.397, 0.206, 0.504, 0.576].
...	
(t=71.0s) User:	<Frame at 71.0s>
(t=71.0s) Assistant:	At 71.0s, [0.403, 0.246, 0.497, 0.601].
(t=74.0s) User:	<Frames at 74.0s>
(t=74.0s) Assistant:	At 74.0s, [0.279, 0.361, 0.52, 0.657].
Dense Video Captioning	
(t=81.5s) User:	<Frames During 0.0s - 80.0s> Identify all new events that occurred and ended up to the current frame, which have not been reported before. Provide their start times, durations, and descriptions in the format: <start time>- <end time>(duration: <x>seconds), <description>.
(t=81.5s) Assistant:	28s - 54s (duration: 26.0 seconds), chop the lemon grass, ginger, onion, and chili finely. 55s - 80s (duration: 25.0 seconds), preheat oil in a bowl and add onions, chili, lemon grass, ginger, and garlic, stir, and put a lid on.
(t=161.9s) User:	<Frames During 81.5s - 161.9s> (Similar query)
(t=161.9s) Assistant:	129s - 144s (duration: 15.0 seconds), add coconut milk, fish sauce, and soy sauce into the pan.
(t=229s) User:	<Frames During 161.9s - 229.0s> (Similar query)
(t=229s) Assistant:	220s - 228s (duration: 8.0 seconds), pour the sauce over the mussel shells.
Temporal Grounding	
(t=16.9s) User:	Determine whether the event 'the person pours some water into the glass' is still happening or has already finished. If ongoing, respond with: <start time>- now (duration: <x>seconds). If completed, use: <start time>- <end time>(duration: <x>seconds).
(t=16.9s) Assistant:	The event happens from 7.3s to now (duration: 9.6 seconds).
Spatial Temporal Action Localization	
(t=1061.0s) User:	<Frames During Clip Start - 1061.0s> In the current frame, list each person's location and actions.
(t=1061.0s) Assistant:	In the current frame, person at [0.318, 0.334, 0.427, 0.504]: carry/hold (an object), walk. person at [0.421, 0.235, 0.565, 0.444]: carry/hold (an object), run/jog.
...	
(t=1134.0s) User:	What actions did the person at bbox [0.046, 0.002, 0.421, 0.996] perform now and earlier?
(t=1134.0s) Assistant:	In the current frame, the person at [0.046, 0.002, 0.421, 0.996]: stand, listen to (a person), watch (a person). At 1133.0s, the same person at [0.044, 0.004, 0.401, 0.996]: stand, carry/hold (an object), listen to (a person), watch (a person).
Step Localization	
(t=132.1s) User:	<Frames During 0.0s - 131.0s> Based on the current observation, list the events that concluded up to the current frame and were not reported in earlier outputs. Ensure your response follows the format: <start time>- <end time>(duration: <x>seconds), <description>.
(t=132.1s) Assistant:	25.0s - 37.0s (duration: 12.0 seconds), install the new faucet. 38.0s - 131.0s (duration: 93.0 seconds), connect the water pipe.
(t=182.0s) User:	<Frames During 132.0s - 179.0s> (Similar query)
(t=182.0s) Assistant:	132.0s - 179.0s (duration: 47.0 seconds), open the sluice and test the new faucet.

Table 7. Instruction template examples and formatted output answers for each task.

tamp). The impact of fps on model performance under the sliding window setting is shown in Figure 2. **Higher fps offers better performance.**

The impact of input frames (sliding window size). We fixed the fps to 2 and selected 16, 32, and 64 frames for evaluation in Figure 1. We select LongVA [20], trained exclusively on static image data, LLaMA-VID [7], which incorporates both single-image and video training data, and MLLM, an extension of LLaVA-OneVision [11] trained on single-image, multi-image, and video data, for a comprehensive comparison. Notably, the advantages of our model in handling diverse task types and achieving superior overall performance remain consistent regardless of the number of frames. This demonstrates the value of online data in

Temporal Context	Spatial Context	Query Examples
Temporal Hallucination Verification	Action Discrepancy	1) Is the person in the [0.168, 0.193, 0.846, 0.996] location in the current frame performing the walk?
	Step Verify	1) Is the person in the current frame still performing the 'install the motherboard'?
	Object Presence	1) Is the umbrella [0.507, 0.606, 0.612, 0.868] still in the screen 3.0 seconds before? 2) How many markers are there on the screen 14.0 seconds before? Does the number increase or decrease compared with the past screen?
SpatialPerception	Action Location	1) What action is the person at the location [0.024, 0.122, 0.624, 0.979] currently performing? 2) How many people in the current frame are performing the action: carry/hold (an object) ? 3) Where is the person currently performing the talk to (e.g., self, a person, a group) located in the picture?
	Object Position	1) Based on visible information, which option most accurately describes the location of the blankets on the screen? (Note: Positions with counts, e.g., 'left-middle (2) ', indicate multiple objects in the same area.) 2) Which option most accurately describes the relative positions of other sheep with respect to the reference position [0.388, 0.288, 0.509, 0.51] on the screen?
PastMemory	Action Retrieval	1) Where was the person currently performing the talk to (e.g., self, a person, a group) in the scene 8 seconds ago? 2) How many people were performing the watch (a person) in the scene 60 seconds ago?
	Procedure Recall	1) What goal was achieved in this video? 2) Did the person follow the correct procedure to achieve the 'wash dish'? 3) What actions did the person perform in sequence in the last 90 seconds? 4) What steps did the person not perform in the last 15 seconds? 5) How long has the person been performing the 'drive the car backward' in the last 90 seconds? 6) Which action did the person perform for the longest duration in the last 15 seconds? 7) What actions was the person performing before the last 30 seconds?
	Trajectory Retrieval	1) Where is the location of the monkey [0.516, 0.49, 0.679, 0.804] on the screen 17.0 seconds before? 2) When does the sheep [0.491, 0.386, 0.584, 0.615] in the current screen first appear in the video? Give the corresponding position when it first appears.
FuturePrediction	Action Anticipation	1) What action is the person currently in the [0.328, 0.211, 0.436, 0.809] location likely to do next? 2) What location in the frame is the person currently in the [0.485, 0.386, 0.578, 0.7] location likely to move to next?
	Goal/Step Prediction	1) My goal is 'make flower crown'. What are the next steps I should take? 2) Based on the series of actions performed by the person in the video, what is the ultimate goal?
	Movement Prediction	1) What direction do you think the baby [0.0, 0.062, 0.526, 0.903] may move towards in the next second?
TemporalPerception	Action Sequence	1) What is the sequence of actions the person in the scene has performed recently?
	Step Localization	1) How long has the person in the scene been performing the 'restore the fixed battery components and the back cover'?
	Object Existence State	1) What is the time period the turtle [0.459, 0.518, 0.501, 0.556] in the current screen appears in the video? And what is the time period in which it disappeared?
SpatioTemporalPerception	Action Trajectory	1) What is the sequence of actions and the corresponding movement trajectory of the person currently in the [0.383, 0.304, 0.642, 0.991] location?
	Object Trajectory	1) What is the trajectory of the object among car [0.482, 0.518, 0.485, 0.531], car [0.561, 0.51, 0.616, 0.577] in the past 5 seconds, which moves the shortest distance? If an object disappears in the middle, calculate the distance based on the time period it last appears. 2) In the video, what is the trajectory of the person [0.049, 0.103, 1.0, 1.0] in the past 2 seconds? Also, point out the period it disappears. 3) Compared with 5 seconds ago, are the person [0.295, 0.614, 0.372, 1.0] and the guitar [0.299, 0.712, 0.419, 0.847] closer or farther apart? 4) What is the trajectory of the object among person [0.315, 0.258, 0.671, 1.0], nutcracker [0.322, 0.768, 0.487, 1.0] in the past 3 seconds, which moves the shortest distance? If an object disappears in the middle, calculate the distance based on the time period it last appears.

Table 8. Task Hierarchy and Question Templates: Overview of task categories, their subcategories, and corresponding example question templates. Each task is designed to probe specific spatiotemporal reasoning capabilities in video understanding, ranging from hallucination detection to future action prediction.

enhancing performance in real-time scenarios, while minimizing computational overhead, which expands deployment possibilities.

The impact of model size. We use InternVL2 [5] family as the research object as it has a wide variety of models of different scales: {1B, 2B, 4B, 8B}.

As shown in Table 4, it can be seen that the performance of models 1, 2, and 4B is almost the same, but there has been significant improvement in performance for the 8B model. It is crucial to deploy larger-scale models in online scenarios effectively.

4. Benchmark Details

4.1. Video and Query Length Distributions

Total 1,463 videos. The distributions of video lengths and query lengths are illustrated in Figure 5.

4.2. Details of QA Generation

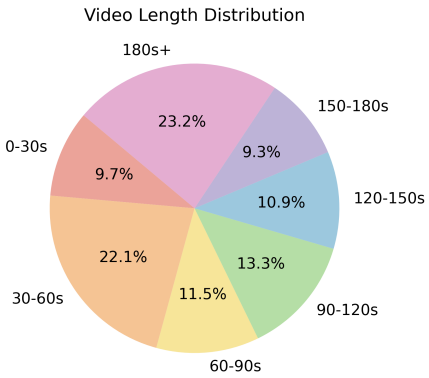
The QA template for OVBench is shown in Table 8. For each task type with different detailed spatiotemporal annotations, we have taken specific measures in Table 5 to ensure the diversity and difficulty of the problem and option generation.

4.3. Data Examples

One visual example for each task type, as shown in Figure 6, 7, and 8.

5. Training and Inference Hyper-parameters

The hyperparameters used in training and the memory bank fps and capacity settings during inference are shown in the table 9 and table 10.



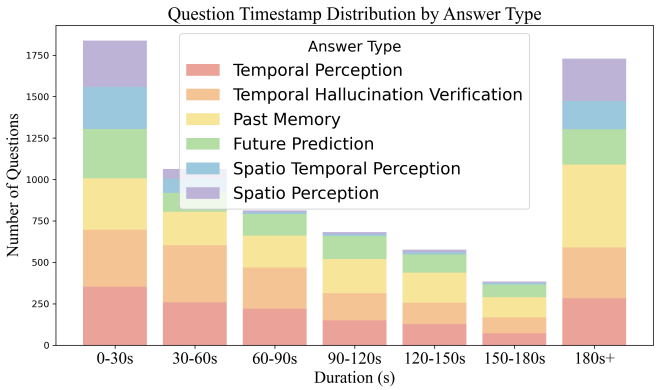


Figure 5. Distributions of video and query lengths. The left figure represents the video length distribution, while the right figure shows the query length distribution.

Hyper-parameter	**Value**
Visual Encoder	
Frame Sampling Rate	1 FPS
Max Frames	64
Preprocessing	Center Crop
Input Resolution	448 × 448
Patch Size	14 × 14
Trainable?	False
Frame Compressor	
Pixel shuffle scale factor	0.5
AvgPool2d Output Size	{16×16, 8×8, 4×4}
MLP Projector	
Number of Layers	2
Hidden Size	4096
Output Size	3072
Trainable?	True
Large Language Model	
Architecture	Phi-3 [1]
Trainable?	True
Model Training	
Offline Training Epochs	1
Online Joint Training Epochs	1
Batch Size	1024
Learning Rate	1e-4
Weight Decay	0.05
Warmup Ratio	0.03
LR Scheduler Type	Cosine
Optimizer	AdamW [12]
AdamW β_1, β_2	(0.9, 0.999)

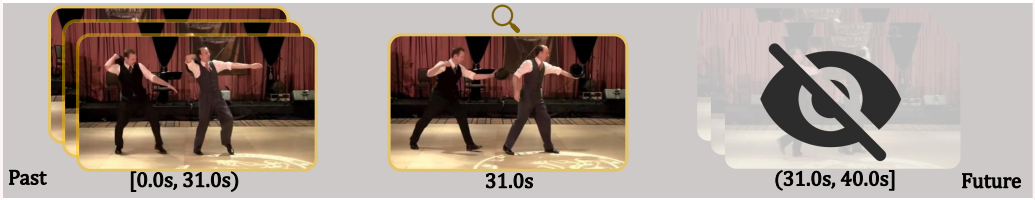
Table 9. Hyper-parameters for fine-tuning.

Memory Bank	**Value**
Frame Sampling Rate	
m_s	1 FPS
m_{main}	2 FPS
m_t	8 FPS
Capacity for Online Benchmark (Token Per Frame × Frames)	
m_s	256 tokens × 2 Frames
m_{main}	64 tokens × 2 Frames
m_t	16 tokens × 12 Frames
Total Tokens	832 tokens
Capacity for Offline Benchmark	
m_s	256 tokens × 24 Frames
m_{main}	64 tokens × 24 Frames
m_t	16 tokens × 144 Frames
Total Tokens	9984 tokens

Table 10. Pyramid Memory Bank Hyper-parameters for Inference.

[SP] Spatial Perception

Question: Based on visible information, which option most accurately describes **the location of the hats** on the screen? (Note: Positions with counts, e.g., 'left-middle (2)', indicate multiple objects in the same area.)?
Timestamp: 31.0s



Options:
(A) right-middle, right-middle, left-bottom, left-middle (2)
(B) center, left-top, left-bottom (2)
(C) left-middle, left-middle (2), left-top
(D) right-top, top
Correct Answer: (D)

[STP] Spatial Temporal Perception

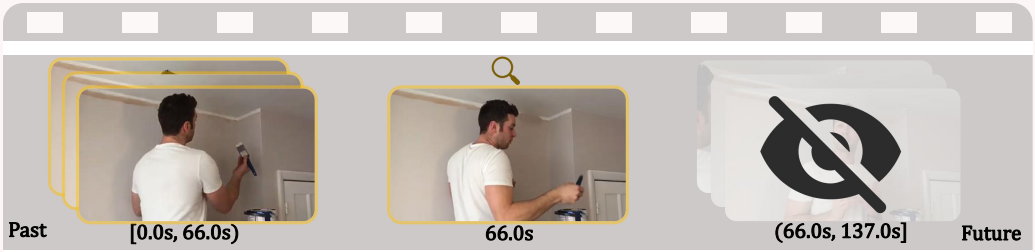
Question: Compared with 34 seconds ago, are the calf [0.177, 0.289, 0.409, 0.6] and the calf [0.635, 0.629, 1.0, 1.0] closer or farther apart?
Timestamp: 38.0s

Options:
(A) Their distance almost remains unchanged
(B) They are getting farther apart.
(C) They are getting closer
Correct Answer: (B)

Figure 6. Task examples in OVBench: Spatial Perception&Spatial Temporal Perception

[TP] Temporal Perception

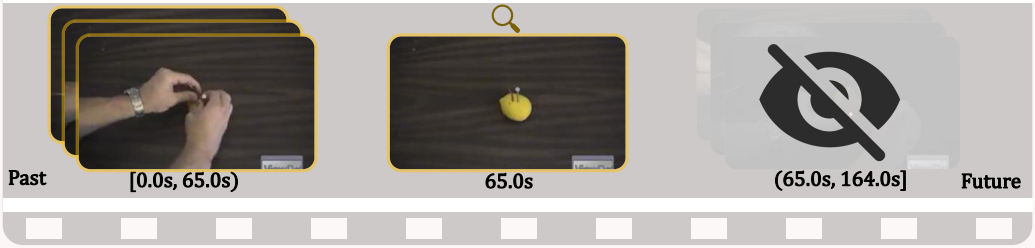
Question: How long has the person in the scene been performing the 'put up to the corner'?
Timestamp: 66.0s



Options:
(A) 27~28s (B) 31~32s
(C) 5~6s (D) 10~11s
Correct Answer: (D)

Hallucination

Question: Is the person in the current frame still performing the 'put stick on lemon'?
Timestamp: 65.0s



Options:
(A) Yes (B) No
Correct Answer: (B)

Figure 7. Task examples in OVBench: Temporal Perception&Temporal Hallucination Verification

[PM] Past Memory

Question: What actions did the person not perform in the last 150 seconds?
Timestamp: 205.0s

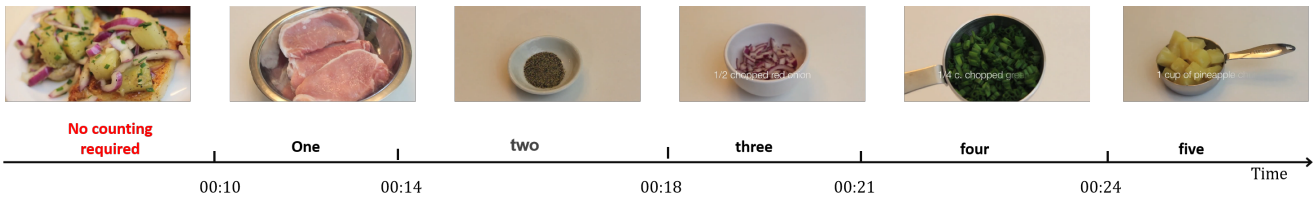
Options:
(A) remove the peel
(B) cut in half
(C) cut both ends and remove fruit seeds
(D) slice the pulp
Correct Answer: (B)

[FP] Future Prediction

Question: My goal is 'paste Window Decal'. What are the next steps I should take?
Timestamp: 91.0s

Options:
(A) apply soap water to the glass surface
(B) press the decal
(C) tear off the other side of the decal
(D) align the position
Correct Answer: (B)

Figure 8. Task examples in OVBench: Past Memory&Future Prediction



User Question at 25.0s: How many ingredients are shown in the past 15 seconds?

FlashVstream The video shows a total of 4 ingredients in the past 15 seconds. ☹

VideoLLM − Online (Video Time = 25.0s) Assistant: You've added flour to the bowl. That's one ingredient so far! ☹

MovieChat I see a white bowl of green onions on a table in the first 10 seconds, a white bowl of green peas on a table in the next second, and a white bowl of green and red onions on a white table in the next second. ☹

TimeChat 7. ☹

VTimeLLM The video shows a person cutting up a pepper and a tomato, as well as a plate of chicken. Therefore, there are three ingredients shown in the past 15 seconds. ☹

VideoChat − Online Five ingredients are shown in the past 15 seconds.

Figure 9. Qualitative Results

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 1, 7

[2] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024. 1

[3] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[4] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18407–18418, 2024. 4

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 4, 6

[6] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, page 436–454, Berlin, Heidelberg, 2020. Springer-Verlag. 3

[7] Yanwei Li et al. Llama-vid: An image is worth 2 tokens in llms. In *ECCV*, 2024. 5

[8] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. *CVPR*, 2017. 3

[10] 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. 4

[11] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[13] 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 (CVPR)*, pages 14313–14323, 2024. 4

[14] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv:2307.16449*, 2023. 1, 4

[15] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216, 2019. 3

[16] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. Qwen2 technical report. *ArXiv*, abs/2407.10671, 2024. 1

[17] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[18] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23056–23065, 2023. 3

[19] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024. 1, 4

[20] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 5

[21] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3