Learning to Reason in 4D: Dynamic Spatial Understanding for Vision Language Models

Supplementary Material

This supplementary material contains additional details of the main manuscript. Section 1 presents additional implementation details for training data, evaluation and QA generation. Section 2 complements more experiments and analysis. Section 3 illustrates some examples of QA in our DSR-Train and DSR-Bench.

1. Implementation Details

1.1. Training Data Details

For 50K QAs in DSR-Train, we present the proportion of different question types in Figure S1.

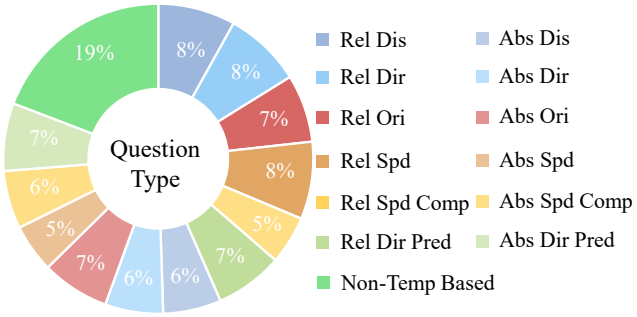


Figure S1. Proportion of questions types in our DSR-Train.

The illustration shows that DSR-Train has similar question type distribution with DSR-Bench as in Figure 3(b) of the main manuscript. For the randomly sampled 5K, 10K, 20K QAs for training in Section 5.3, we restrict their distributions to be the same as that in Figure S1.

1.2. Evaluation Details

During evaluation, we uniformly sample 32 frames from each input video as visual inputs for all VLMs. For all models except Qwen series [1, 10] and ours, which natively encode absolute timestamps, we additionally append the frame timestamps to the text input. To ensure a fair comparison, all VLMs are instructed to directly output final choice without intermediate reasoning. For evaluation on Video-MME [5], video subtitles are excluded from input.

1.3. QA Generation Details

Object Denotation. In both DSR-Train and DSR-Bench, each object is referenced using a combination of its category and bounding box coordinates either at a specific timestamp or at the beginning of the queried sub-interval. This differs from some prior benchmarks that refer to objects solely through appearance descriptions, which can be ambiguous in our setting where many objects share similar visual characteristics. When an object is used to determine the absolute viewpoint, it will be denoted as "{class} with bounding box coordinates (x_1, y_1, x_2, y_2) at $\{time_v\}$", where "{class}" is the object category and "(x_1, y_1, x_2, y_2)" correspond to the top-left and bottom-right coordinates of the bounding box in the frame sampled at the timestamp "$\{time_v\}$". Otherwise, the object is referred to as "{class} with initial bounding box coordinates (x_1, y_1, x_2, y_2), indicating that the coordinates correspond to the first frame of the queried sub-interval and can change thereafter, or as "{class} with final bounding box coordinates (x_1, y_1, x_2, y_2)" to indicate that the coordinates are taken from the last sampled video frame and the task is to predict object's moving direction.

Prompt Design. As described in Section 3.1 and 3.2 of the main manuscript, our automated QA generation pipeline uses DeepSeek-R1 [6] to filter out Koala-36M [11] videos exhibiting negligible object motion and to identify agent and non-agent object categories based on the video captions. To construct DSR-Bench, we further employ Gemini-2.5-Pro [2] to filter videos with visual content and to assign scene classes with higher accuracy. Figure S2 presents our prompts for DeepSeek-R1 and Gemini-2.5-Pro, where "{caption}" denotes the video caption, and "{agent}" and "{object}" are agent and non-agent classes in the video.

In Section 3.3 of the main manuscript, to broaden linguistic variety and reasoning patterns, we also propose to prompt DeepSeek-R1 for non-template-based QA generation. In Figure S3, we show the detailed prompt for DeepSeek-R1 for QA generation, where "{viewpoint}" is the observer to decide the viewpoint for observing the objects, "{coord}" is the 3D coordinate trajectories of all objects, "$\{time_1\}$" and "$\{time_2\}$" are the start and end timestamp of the sub-interval, "{timestamps}" is the list of timestamps at which the object 3D coordinates are collected.

Question Generation Template. Part of our QAs are generated based on pre-defined templates, as stated in Section 3.3 of the main manuscript. We list these templates for each type of question in Table S1. "$\{time_s\}$", "$\{time_e\}$" are the timestamps of the start, end of the queries sub-interval and "$\{time_v\}$" is the timestamp to determine the absolute viewpoint. "$\{obj_1\}$", "$\{obj_2\}$" are the target objects and "$\{obj_v\}$" is the observer to decide the viewpoint, where their denotations are in the format described in the previous **Object Denotation** section.

Answer Generation Rule. In Section 3.3 of the main

Caption-Based Motion Filter Prompt: """{caption} is the caption of a video. Think about whether the video satisfies the following condition: Regardless of camera movement and articulated motion, the world coordinates of at least 3 main subjects in the video change substantially. Only consider the objects that can be regarded as a whole and not part of other objects.

If the condition is satisfied, output “yes” in a new line. If the condition is not satisfied, output “no” in a new line."""

Caption-Based Agent/Non-Agent Extract Prompt: """{caption} is the caption of a video. In one line, list all classes of agents in the video and use dot to separate them. If there is no agent, output “None” in this line. Then in a new line, list the classes of remaining foreground objects in the video and use dot to separate them. An object should be a discrete, countable item (e.g. a bottle, a rock). The object must not be a part of other objects. If there is no object, output “None” in this line."""

(a) Caption-based motion filter and agent/non-agent extract prompt for DeepSeek-R1

Video-Based Motion Filter and Scene Classification Prompt: """The video belongs to one of following categories: Sports & Recreation, Transportation & Vehicle Operation, Art Performance, Manual Labor & Craftsmanship, Daily Activities & Hobbies, Nature & Wildlife. The agent in the video is {agent}. The object in the video is {object}. Think about whether the video satisfies the following conditions:
(1) The camera is not mainly filming the inside of one moving object;
(2) Regardless of camera movement, articulated motion and shape variance, the world coordinates of at least 3 agents (exclude hands) or object change substantially;
(3) Every object cannot occupy less than 1% of the screen;
(4) Every agent or object cannot occupy more than 50% of the screen.

Only consider the objects that can be regarded as a whole and not part of or containing other objects. Only consider the objects that are discrete, countable items (e.g. a bottle, a rock). Only consider the objects that are not in contact with each other. Only consider agents and objects belonging to classes given at the beginning. If all conditions are satisfied, output “yes” in a new line. If any condition is not satisfied, output “no” in a new line. Then in a new line, output the category of the video."""

(b) Video-based motion filter and scene classification prompt for Gemini-2.5-Pro

Figure S2. Detailed prompts leveraged for DeepSeek-R1 and Gemini-2.5-Pro for data process.

Non-Template-Based QA Generation Prompt: """ From the perspective of {viewpoint} in a video, {coord} are the 3D coordinates (x, y, z) of some objects between $\{time_1\}$ and $\{time_2\}$. The positive direction of x, y, z axis are forward, left, up respectively. The coordinates are collected at {timestamps}. According to the change of 3D relationship between objects, please generate one question, which cannot be answered only with 2D knowledge and without 3D knowledge, and 4 answers where one of them is the correct answer. The question should not include description of perspective. The question should not be quantitative and related to accurate 3D coordinates. All objects must be referred with their bounding box coordinates given at the beginning. In a new line, only output your question. Then in 4 new lines, only output your 4 answers. After that, only output the symbol of the correct answer in a new line."""

Figure S3. Detailed prompts leveraged for DeepSeek-R1 to generate non-template-based QAs.

manuscript, we describe our approach to generate answers by comparing the queried attribute every 2 adjacent frames to derive a sequence of basic choice defined in Table 2 of the main manuscript and then merging consecutive identical states. Table S2 provides the detailed derivation rules for each basic choice. After determining the correct answer, we compute the number of basic choices N it contains and generate alternative choices with random lengths within the range $[\max(1, N-3), N+3]$. To maintain a balanced distribution, the correct answer label (A, B, C, or D) is randomly assigned. We note that frames in the final second are hidden from VLMs, as they are used to determine the object's movement direction for direction prediction questions.

Table S1. Templates for different question types.

Type	Template
Rel Dis	Between {time_s} and {time_e}, following the perspective of {obj_v}, how does the distance between {obj_1} and {obj_2} change?
Rel Dir	Between {time_s} and {time_e}, following the perspective of {obj_v}, how does the direction of {obj_1} to {obj_2} change?
Rel Ori	Between {time_s} and {time_e}, following the perspective of {obj_v}, how does the orientation of {obj_1} change?
Rel Spd	Between {time_s} and {time_e}, following the perspective of {obj_v}, how does the speed of {obj_1} change?
Rel Spd Comp	Between {time_s} and {time_e}, following the perspective of {obj_v}, compare the speed between {obj_1} and {obj_2}.
Rel Dir Pred	Following the perspective of {obj_v}, predict the moving direction of {obj_1}.
Abs Dis	Between {time_s} and {time_e}, from the perspective of {obj_v} at {time_v}, how does the distance between {obj_1} and {obj_2} change?
Abs Dir	Between {time_s} and {time_e}, from the perspective of {obj_v} at {time_v}, how does the direction of {obj_1} to {obj_2} change?
Abs Ori	Between {time_s} and {time_e}, from the perspective of {obj_v} at {time_v}, how does the orientation of {obj_1} change?
Abs Spd	Between {time_s} and {time_e}, from the perspective of {obj_v} at {time_v}, how does the speed of {obj_1} change?
Abs Spd Comp	Between {time_s} and {time_e}, from the perspective of {obj_v} at {time_v}, compare the speed between {obj_1} and {obj_2}.
Abs Dir Pred	From the perspective of {obj_v} at {time_v}, predict the moving direction of {obj_1}.

Table S2. Derivation rules of different basic choices to make up answers.

Type	Basic Choice	Derivation Rule
Distance & Speed	(1) Keep nearly constant then become larger	At each timestamp within a period, except final one, the distance/speed remains within $0.8\times$ to $1.2\times$ that of the first timestamp. At the final timestamp, the distance/speed exceeds $1.2\times$ that of the first timestamp.
	(2) Keep nearly constant then become smaller	At each timestamp within a period, except final one, the distance/speed remains within $0.8\times$ to $1.2\times$ that of the first timestamp. At the final timestamp, the distance/speed is below $0.8\times$ that of the first timestamp.
	(3) Keep nearly constant	At each timestamp within a period, except final one, the distance/speed remains within $0.8\times$ to $1.2\times$ that of the first timestamp.
	(4) Become larger	At each timestamp within a period, the distance/speed exceeds $1.2\times$ that of the former one.
	(5) Become smaller	At each timestamp within a period, the distance/speed falls below $0.8\times$ that of the former one.
Direction & Orientation & Direction Prediction	(1) Front/Behind	The angle between the unit vector of direction/orientation and the forward unit vector from viewpoint is smaller/larger than 70/110 degrees.
	(2) Left/Right	The angle between the unit vector of direction/orientation and the left unit vector from viewpoint is smaller/larger than 70/110 degrees.
	(3) Above/Below	The angle between the unit vector of direction/orientation and the upward unit vector from viewpoint is smaller/larger than 70/110 degrees.
Speed Comparison	(1) Nearly the same	The speed of the former object is within $0.83\times$ to $1.20\times$ that of the latter object at one timestamp.
	(2) The former is faster	The speed of the former object exceeds $1.20\times$ that of the latter object at one timestamp.
	(3) The latter is faster	The speed of the former object falls below $0.83\times$ that of the latter object at one timestamp.

2. Complementary Experiments

In this section, we present additional experiments that were not included in the main manuscript. In Section 2.1, we provide further analysis of the results in Table 4 of main manuscript. Section 2.2 compare the performance of models trained on data with different question type distributions. In Section 2.3, we replace the original base model, *i.e.*, Qwen2.5-VL-7B, with Qwen3-VL-8B and report performance across different benchmarks to demonstrate the effectiveness of GSM and DSR-Train on different models. Section 2.4 incorporates DSR-Train with question–answer pairs for spatial reasoning in static scenes to train Qwen2.5-VL-7B with unified static and dynamic spatial reasoning capabilities. Finally, Section 2.5 extends the DSR-Train fine-tuned model fine-tuned to MineDojo [3], demonstrating its broader applicability to downstream agent tasks requiring dynamic spatial reasoning.

2.1. Additional Result Analysis

In addition to the conclusions in Section 5.2, Table 4 of the main manuscript further reveals that spatial reasoning models, *i.e.*, VLM-3R [4] and VG-LLM [15], achieve even better performance than proprietary models, despite being trained only for static spatial reasoning. Therefore, 3D relevant data is necessary to improve spatial reasoning capability. Moreover, compared with its base model LLaVA-Video-7B, the marginal performance gain of VideoRefer [14], which is explicitly trained to understand location information for object reference, suggests that the unsatisfactory performance on DSR-Bench is not attributable to our use of bounding-box–based object references. Instead, it highlights the insufficient capability of current models to perform dynamic spatial reasoning.

2.2. Effect of Training Question Distribution

For all experiments in the main manuscript, our model is trained on QAs with the type distribution shown in Figure S1. In this section, we fix the total number of training QAs to be 20K and vary the proportion of template-based and non-template-based QAs under the following configurations: (1) 20K template-based QAs; (2) 16K template-based QAs and 4K non-template-based QAs (the setting adopted in the main manuscript experiments); (3) 10K template-based QAs and 10K non-template-based QAs; (4) 20K non-template-based QAs. Table S3 reports the performance on the template-based subset, non-template-based subset, the full DSR-Bench and Video-MME under each configuration. The results indicate that both QA types are essential for achieving strong overall performance, and that template-based QAs should constitute the majority.

Table S3. Comparison between different training data settings.

Settings	Benchmarks			
	DSR-Bench (Temp)	DSR-Bench (Non-Temp)	DSR-Bench	Video-MME
Baseline	22.1	30.1	23.5	60.2
1	61.5	32.3	56.2	59.6
2	60.4	43.7	57.4	59.9
3	58.1	46.7	56.1	60.1
4	35.2	49.1	37.6	60.5

Table S4. Comparison between GSM and other methods with Qwen3-VL-8B-Instruct as the base model.

Methods	Benchmarks				
	DSR-Bench	VLM4D	STI-Bench	Video-MME	Avg.
Baseline	28.7	46.3	36.3	64.9	44.0
SFT	56.8	47.8	37.3	64.7	51.6
Addition	59.0	49.5	37.8	53.6	49.9
GSM	58.6	49.2	37.9	64.4	52.5

2.3. Ablation on Different Base Model

In the main manuscript experiments, we adopt Qwen2.5-VL-7B as the base model, integrate our proposed GSM into it and train it on DSR-Train. To demonstrate the effectiveness of GSM and DSR-Train independent of the underlying base model, we further replace Qwen2.5-VL-7B with Qwen3-VL-8B-Instruct, incorporate GSM, and train it on the same 20K samples from DSR-Train as used in Section 5.3 of the main manuscript. We also compare its performance with different training methods when adopting the same compared training paradigm in Section 5.3 of main manuscript, *i.e.*, SFT and Addition. The results on DSR-Bench, STI-Bench [8], VLM4D [16] and Video-MME [5] listed in Table S4 show that our DSR-Train improves the dynamic spatial reasoning performance of Qwen3-VL-8B-Instruct as well. Compared with SFT, GSM further improves performance by integrating explicit geometric priors. In contrast to Addition, GSM maintains the general video-understanding ability of Qwen3-VL-8B-Instruct, preventing degradation of broader visual reasoning skills.

2.4. Performance with Data Mixture

In all previous experimental settings, the base model Qwen2.5-VL-7B is fine-tuned exclusively on our DSR-Train to perform spatial reasoning in dynamic scenes. In this section, we further mix QAs for static spatial reasoning with DSR-Train to fine-tune Qwen2.5-VL-7B, improving both static and dynamic spatial reasoning ability. Specifically, we construct a total of 800K static spatial reasoning QAs using data from VLM-3R [4] and Cambrian-S [13], and adopt VSI-Bench [12] as the benchmark for evaluating

(a) hunt a cow in plains with a iron sword, shield, and a full suite of iron armors

(b) combat a spider in night plains with a diamond sword, shield, and a full suite of leather armors

Figure S4. Examples of agent results on MineDojo.

Table S5. Performance analysis of models trained with different spatial reasoning scenes.

Training Scene	Benchmarks		
	DSR-Bench	VSI-Bench	Video-MME
Baseline	23.5	33.4	60.2
Static	24.1	55.3	58.9
Static+Dynamic	60.2	56.1	59.2

Table S6. Success rate comparison on MineDojo between agents adopted from VLM trained with different spatial reasoning scenes.

Training Scene	Tasks	
	Animals	Hostiles
Baseline	15.6±3.4	10.2±2.4
Static	16.3±2.3	12.4±3.1
Dynamic	26.5±1.7	22.3±2.3

static spatial reasoning. The performance of model trained only on static spatial reasoning or mixed QAs on DSR-Bench, VSI-Bench and Video-MME are listed in Table S5. From the results, when the model is trained only with static spatial reasoning QAs, it is still unable to perform dynamic spatial reasoning. With both static and dynamic spatial reasoning data, the trained model obtains the best performance simultaneously on DSR-Bench and VSI-Bench while general understanding capability is still preserved.

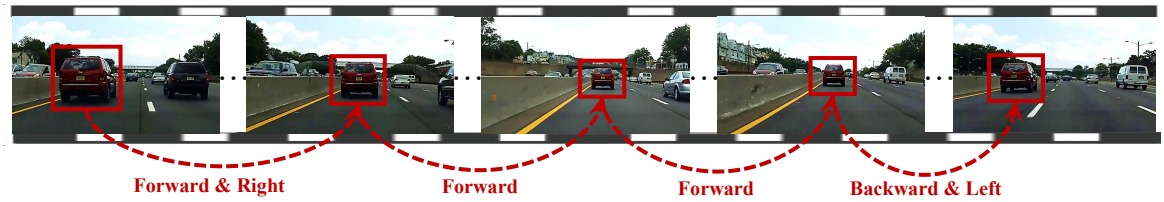
2.5. Extension to Agent Tasks

In this section, we also explore the downstream performance on agent tasks [3, 9], showing the spatial capability on real-time working agent. We further extend the Qwen2.5-VL-7B fine-tuned with DSR-Train to MineDojo [3] Benchmark, which simulates a Minecraft game agent to solve different tasks, to show the application of dynamic spatial reasoning models in downstream agent tasks. Since the training dataset of MineDojo only provides videos and task instructions without intermediate action labels, we instead leverage the data from MineRL [7] for training by mapping its action space to that of MineDojo. For evaluation, we select part of tasks from MineDojo that are related to the following two sets of objects: (1) Animals including cow, sheep, pig and chicken, whose related tasks are targeted at combating or harvesting them; (2) Hostiles including spider, zombie, pigman and enderman, whose related tasks aim to combat them. Since these tasks require interacting with dynamic objects, they command dynamic spatial reasoning ability of agents. For compared models, we train original Qwen2.5-VL-7B and that fine-tuned with static spatial reasoning data as stated in Section 2.4 on the same MineRL data. Each task is tested on 3 random seeds and with 200 episodes. The average and variance of success rate are listed in Table S6, showing the advantage of our model in downstream agent tasks. Some task results of the agent adopted from model trained with DSR-Train are illustrated in Figure S4.

3. Visualization

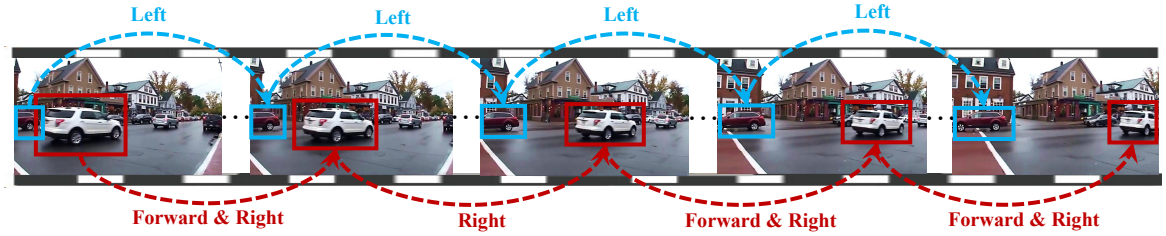
In this section, we illustrate the video and QA examples of different categories in our DSR-Train and DSR-Bench in Figure S5, S6 and S7, which shows our in-the-wild video source and comprehensive evaluation aspects including multi-object interaction, viewpoint transformation and fine-grained temporal reasoning. Note that the arrows in the figures indicate the movement from the viewpoint of camera, which can be different when the viewpoint changes.

Transportation & Vehicle Operation



Question (Relative Distance): Between 13s and 35s, following the perspective of camera, how does the distance between the car with initial bounding box coordinates [124, 116, 263, 224] and camera change?

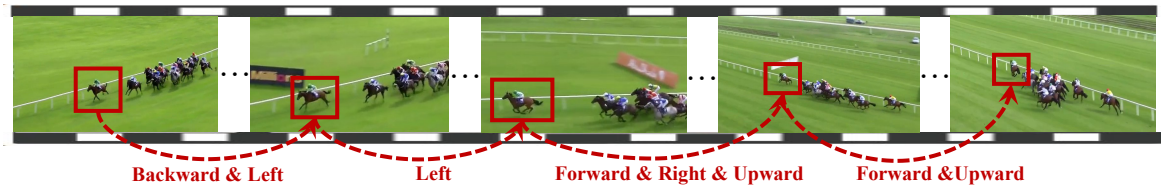
Choices: A. (1) Become smaller. (2) Become larger. B. (1) Keep nearly constant then become larger. (2) Become larger. (3) Become smaller.
C. (1) Become larger. (2) Keep nearly constant then become smaller. (3) Become smaller. D. (1) Keep nearly constant.



Question (Relative Direction): Between 16s and 18s, following the perspective of the car with initial bounding box coordinates [78, 156, 304, 284] , how does the direction of the car with initial bounding box coordinates [0, 174, 67, 234] to the car with initial bounding box coordinates [78, 156, 304, 284] change?

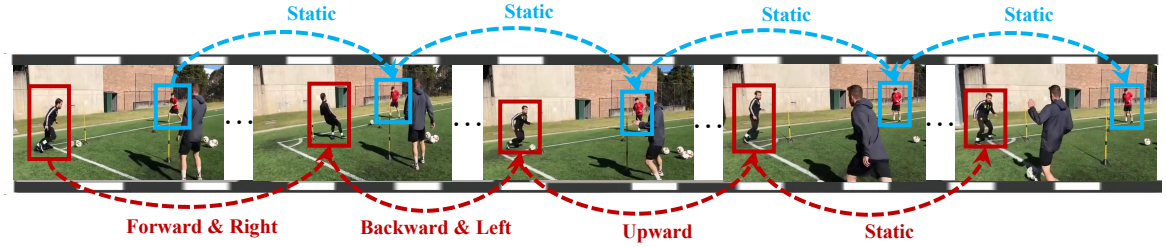
Choices: A. (1) Front|Right|Above. (2) Behind|Left. B. (1) Left. (2) Behind|Left C. (1) Right|Below. D. (1) Left. (2) Left|Above.

Sports & Recreation



Question (Relative Orientation): Between 0s and 10s, following the perspective of camera, how does the orientation of the horse with initial bounding box coordinates [204, 182, 236, 223] change?

Choices: A. (1) Behind|Left|Below. (2) Behind|Left. (3) Left. (4) Left|Above. (5) Front|Left|Above
B. (1) Left. (2) Left|Below. (3) Front|Right|Above. (4) Left|Above. (5) Front|Left|Above. (6) Left|Above
C. (1) Right. (2) Front|Left. (3) Left|Above. (4) Front|Left|Below
D. (1) Left|Below. (2) Behind|Left. (3) Left|Below

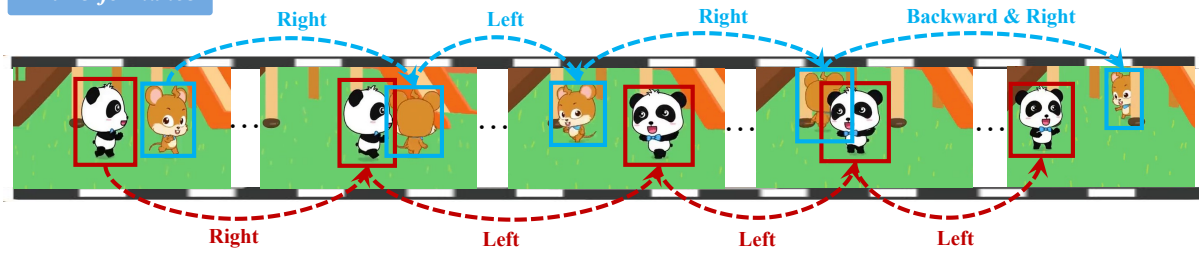


Question (Relative Speed): Between 3s and 10s, following the perspective of the man with initial bounding box coordinates [44, 96, 79, 275] , how does the speed of the man with initial bounding box coordinates [413, 87, 436, 162] change?

Choices: A. (1) Keep nearly constant then become larger. (2) Keep nearly constant.
B. (1) Become larger. (2) Become smaller. (3) Become larger. (4) Keep nearly constant
C. (1) Become smaller. (2) Become larger. (3) Keep nearly constant.
D. (1) Become larger. (2) Become smaller. (3) Keep nearly constant

Figure S5. Examples of QA and video in DSR-Train and DSR-Bench.

Art Performance



Question (Relative Speed Comparison): Between 0s and 6s, following the perspective of camera, compare the speed between the animal with initial bounding box coordinates [253, 44, 326, 208] and the animal with initial bounding box coordinates [384, 43, 447, 196].

Choices: A. (1) The former is faster. (2) The latter is faster. (3)The former is faster. (4) The latter is faster.
B. (1) Nearly the same. (2) The former is faster. (3) The latter is faster. (4) Nearly the same. (5) The latter is faster.
C. (1) The latter is faster. (2) Nearly the same . (3) The former is faster. (4) Nearly the same.
D. (1) The former is faster. (2) Nearly the same. (3) The former is faster.

Invisible to VLM

Question (Relative Direction Prediction): Following the perspective of camera, predict the moving direction of the man with final bounding box coordinates [143, 57, 164, 306]?

Choices: A. (1) Behind|Left B. (1) Right|Below C. (1) Right D. (1) Behind|Right|Above

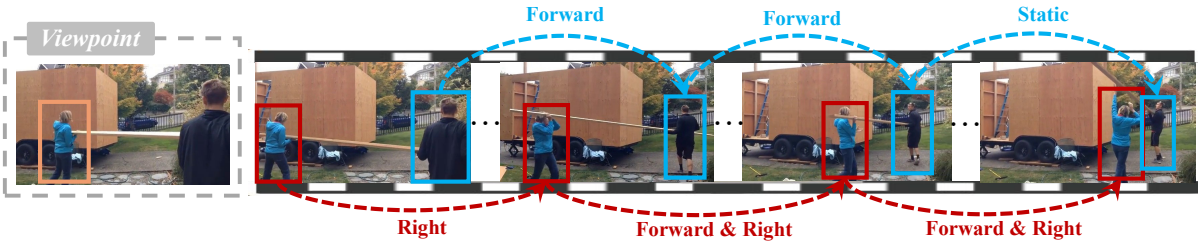
Manual Labor & Craftsmanship

Viewpoint

Question (Absolute Distance): Between 21s and 25s, from the perspective of camera at 18s, how does the distance between the man with initial bounding box coordinates [125, 86, 324, 359] and the camera change?

Choices: A. (1) Become larger. (2) Become smaller.
C. (1) Keep nearly constant then become larger. (2) Become larger.
B. (1) Become smaller
D. (1) Become larger.

Viewpoint



Question (Absolute Direction): Between 16s and 21s, from the perspective of the woman with bounding box coordinates [134, 156, 163, 359] at 14s, how does the direction of the man with initial bounding box coordinates [547, 136, 639, 359] to the woman with initial bounding box coordinates [47, 153, 82, 359] change?

Choices: A. (1) Behind|Left. (2) Front|Right. (3) Left. (4) Left|Above.
C. (1) Right. (2) Front|Right. (3) Left|Above. (4) Front|Left
B. (1) Front|Right. (2) Front. (3) Front|Left
D. (1) Left|Below. (2) Behind|Left.

Figure S6. Examples of QA and video in DSR-Train and DSR-Bench.

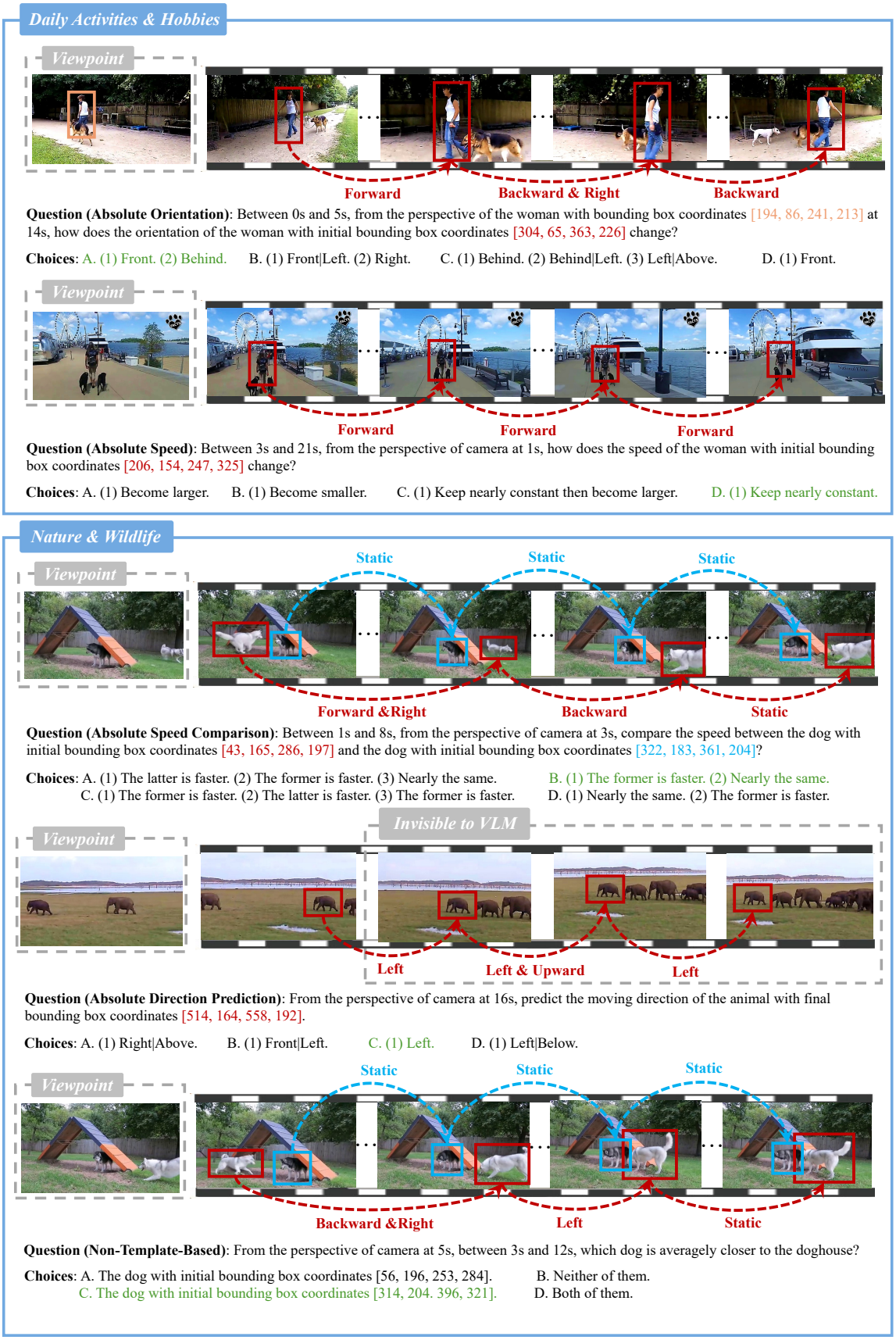


Figure S7. Examples of QA and video in DSR-Train and DSR-Bench.

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. 1

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[3] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35: 18343–18362, 2022. 4, 5

[4] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 4

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 1, 4

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

[7] William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: a large-scale dataset of minecraft demonstrations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2442–2448, 2019. 5

[8] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 4

[9] Mingxian Lin, Wei Huang, Yitang Li, Chengjie Jiang, Kui Wu, Fangwei Zhong, Shengju Qian, Xin Wang, and Xiaojuan Qi. Embrace-3k: Embodied reasoning and action in complex environments. *arXiv preprint arXiv:2507.10548*, 2025. 5

[10] QwenTeam. Qwen3-vl: Sharper vision, deeper thought, broader action, 2025. Accessed: 2025-09-23. 1

[11] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8428–8437, 2025. 1

[12] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10632–10643, 2025. 4

[13] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025. 4

[14] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18970–18980, 2025. 4

[15] Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *Advances in Neural Information Processing Systems*, 2025. 4

[16] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8600–8612, 2025. 4