VLM4D: Towards Spatiotemporal Awareness in Vision Language Models

Shijie Zhou[1]* Alexander Vilesov[1]* Xuehai He[2,3]* Ziyu Wan[2] Shuwang Zhang[1]
Aditya Nagachandra[1] Di Chang[4] Dongdong Chen[2] Xin Eric Wang[3] Achuta Kadambi[1]
[1]UCLA [2]Microsoft [3]UCSC [4]USC

https://vlm4d.github.io/

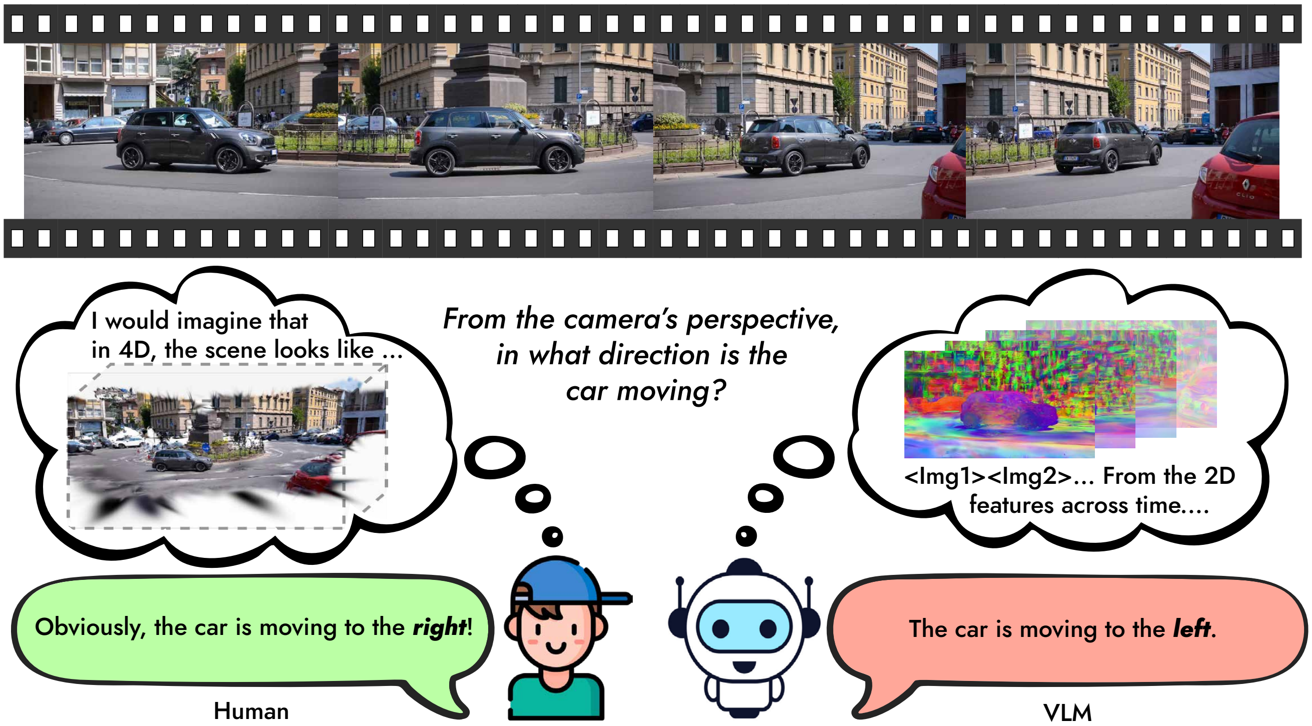

I would imagine that in 4D, the scene looks like ...
From the camera's perspective, in what direction is the car moving?
<Img1><Img2>... From the 2D features across time....
Obviously, the car is moving to the right!
The car is moving to the left.
Human
VLM

Figure 1. **Spatiotemporal (4D) Awareness**. Humans intuitively reason in 4D (3D space + time), effortlessly reconstructing the dynamic spatial trajectory of moving objects from any perspective. In contrast, current Vision Language Models (VLMs) typically rely on aggregating 2D visual features across time, leading to incorrect predictions when motion understanding and interpretation requires deeper spatiotemporal reasoning. In this example, humans correctly perceive the car moving to the right, while the VLM (GPT-4o) inaccurately predicts leftward movement, suggesting VLMs struggle to perform spatiotemporal reasoning.

Abstract

Vision-language models (VLMs) have shown remarkable capabilities in integrating linguistic and visual reasoning but remain fundamentally limited in understanding dynamic spatiotemporal interactions. Humans effortlessly track and reason about object movements, rotations, and perspective shifts—abilities essential for robust dynamic real-world understanding yet notably lacking in current VLMs. In this paper, we introduce `VLM4D`*, the first benchmark specifically designed to evaluate the spatiotemporal reasoning capabilities of VLMs. Our benchmark comprises diverse real-world and synthetic videos accompanied by carefully curated question-answer pairs emphasizing translational and rotational motions, perspective awareness, and motion continuity. Through comprehensive evaluations of state-of-the-art open and closed-source VLMs, we identify significant performance gaps compared to human baselines, highlighting fundamental deficiencies in existing models. Extensive analysis reveals that VLMs struggle particularly with integrating multiple visual cues and maintaining temporal coherence. We further explore promising directions, such as leveraging 4D feature field reconstruction and targeted*

*Equal contribution.

spatiotemporal supervised fine-tuning, demonstrating their effectiveness in enhancing spatiotemporal comprehension. Our work aims to encourage deeper exploration into improving VLMs' spatial and temporal grounding, paving the way towards more capable and reliable visual intelligence for dynamic environments.

1. Introduction

Humans posses an innate ability to perceive, track and interpret motion, spatial and temporal changes, [14, 48] enable rich interpretations of complex dynamic events from both egocentric and allocentric perspectives [6]. When observing an object move, we can inherently process any changes such as lateral shifts, rotational directions and periodic or repeated actions unfolding along a specific trajectory [6]. These sophisticated perceptual abilities are a product of our spatiotemporal cognition [18], and form an essential foundation that allows us to comprehend and reason about physical phenomena, object interactions and causal relationships within our environment. [27, 33]

Vision-language models (VLMs), which can also potentially perceive the motions and spatialtemporal changes in videos, constitute a prominent class of methods designed to emulate or surpass human capabilities in integrated visual and linguistic reasoning [16, 32]. While prior work has focused on static visual understanding from mass training corpuses of language and visual data [56] or understanding video such as captioning [46] and scene understanding [7], we find that the exceptional performance in the prior mentioned tasks does not innately carry over to spatiotemporal capabilities. This limitation is notable given that contemporary state-of-the-art VLMs are typically trained on datasets comprising of billions of annotated video-text pairs [42]. In contrast, human infants naturally develop robust spatiotemporal cognition within the first few months of life [57]. Another key challenge that inhibits VLM performance in spatiotemporal tasks is the necessity to implicitly or explicitly reconstruct a four-dimensional (4D) representation of dynamic scenes and subsequently reason over such reconstruction [64]. As illustrated in Fig. 1, the car is advancing forwards and turning to the left in its own frame of reference. However, from the camera's perspective, its motion appears as a combination of heading to the right and receding into the distance despite the car being in the center of the frame due to camera view rotation. Human observers can seamlessly disentangle these complex dynamics, accurately interpreting trajectories by synthesizing diverse visual cues including camera rotation compensation, stationary scene landmarks, prior knowledge of 3D and 4D environmental structures, and perspective projections [6, 18, 33, 48]. The inability of current VLMs to similarly integrate these cues underscores an important gap. Furthermore, bridging this gap will require VLMs to develop more sophisticated mechanisms for reconstructing and reasoning over dynamic scenes, potentially drawing on insights from cognitive science and neuroscience on how humans process and integrate spatial and temporal information.

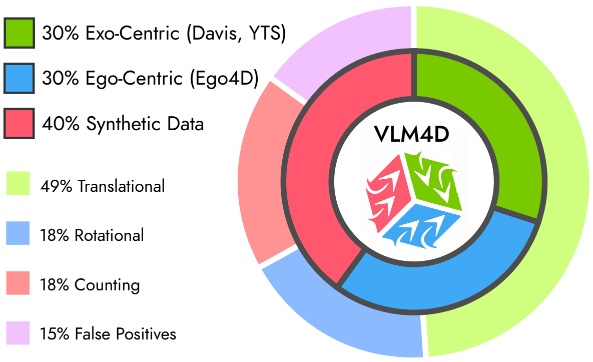


Figure 2. **Distribution of Dataset Sources and Annotations.** Breakdown of our dataset illustrating the proportions of data sourced from third-person (Davis, YouTube), first-person (Ego4D), and synthetic data, categorized by annotation types: translational, rotational, action, counting, and false positives.

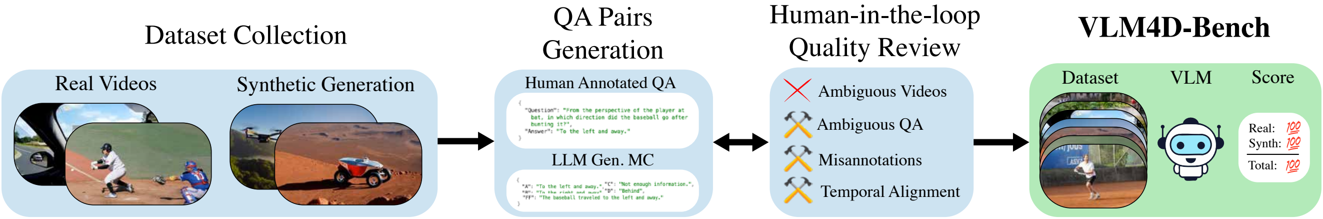
With the mentioned limitations of exisiting VLMs, to effectively characterize and challenge the existing spatiotemporal reasoning abilities of VLMs, we directly evaluate their capacity to track complex directional movements and perspective transformations over time. We introduce `VLM4D`, a rigorous benchmark specifically designed to probe the spatiotemporal grounding capabilities of current vision-language models. Through this contribution, we aim to catalyze research that addresses the critical gap in spatiotemporal understanding and reasoning within VLMs and provide a foundational analysis highlighting key deficiencies in existing models.

We summarize our main **contributions** as follows:

1. We propose the first benchmark `VLM4D` explicitly designed to rigorously evaluate the spatiotemporal reasoning capabilities of Vision-Language Models (VLMs).
2. We introduce a novel, meticulously curated dataset consisting of diverse real-world and synthetic video sequences paired with carefully crafted spatiotemporal question-answer (QA) annotations.
3. We conduct an analysis to identify critical limitations in the spatiotemporal reasoning performance of contemporary VLMs, highlighting fundamental challenges and charting clear directions for impactful future research.

2. Related Work

Spatiotemporal Understanding in Vision Language Models Vision Language Models (VLMs) have evolved rapidly by fully leveraging the significant achievements of Large Language Models (LLMs) [4, 5, 15, 55, 60, 67]



```
{
  "Question": "From the perspective of the player at
    bat, in which direction did the baseball go after
    bunting it?",
  "Answer": "To the left and away."
}
```

```
{
 "A": "To the left and away.","C": "Not enough information.",
 "B": "To the right and away","D": "Behind",
 "FF": "The baseball traveled to the left and away."
}
```

Figure 3. **Dataset Generation and Annotation Pipeline.** Our dataset was constructed by collecting real videos and generating synthetic data, followed by human-in-the-loop quality reviews to address ambiguous videos and annotations. After temporal alignment and quality assurance, human-annotated questions and answers were created, complemented by multiple-choice questions generated by large language models (LLMs). The final dataset includes real-world and synthetic video data with comprehensive VLM scoring metrics.

and large-scale visual instruction tuning datasets [13, 41, 83]. While VLMs [1, 21, 26, 35, 41, 59, 63, 83] exhibit transformative potential for applications such as embodied AI [17, 29, 58], robotics [52, 61], and world modeling [43, 77], most existing methods remain constrained to static images, focusing narrowly on spatial understanding while overlooking the dynamic temporal dimension inherent in real-world interactions. To bridge this gap, emerging research [11, 37, 47, 75, 76] has begun exploring video modality integration, aiming to equip VLMs with spatial-temporal awareness critical for tasks like video comprehension, where both contextual details and motion dynamics are essential. For example, VideoLLM-MoD [69] proposes to address the efficiency issue when processing long-term video by mixture-of-depths. [73] introduces VideoRefer to enhance the finer-level (like object-level) spatial-temporal video understanding of VLMs. Grounded-VideoLLM [62] also targets for fine-grained video understanding through incorporating an additional temporal stream. In this work, we aim to rigorously evaluate the 4D spatial-temporal reasoning capabilities of state-of-the-art VLMs, probing how and to what extent these models internalize spatial intelligence and temporal dependencies.

VLM Benchmarks Following the development trends of VLMs, benchmarking VLMs shares the similar trajectory by first evaluating vision QA on static images [24, 34, 44, 74], to align with models' early focus on 2D understanding. As VLMs evolved to tackle dynamic scenarios, benchmarks expanded to evaluate general-purpose video comprehension tasks that probe temporal coherence and event understanding [19, 28, 39, 40, 49]. Notably, MMVU [80] further proposes a knowledge-intensive benchmark to assess the expert-level reasoning ability of current video-based large models. However, while these works assess perception and semantic understanding, they largely overlook the explicit evaluation of spatial-temporal awareness, a core capability for real-world applications requiring 4D (3D space + time) reasoning. Recent efforts like [72] pioneer benchmarks for 3D visual-spatial intelligence but restrict evaluation to static 3D scene, neglecting the interplay of object motion and temporal dynamics intrinsic to videos. In this work, we introduce VLM4D, the first benchmark designed to holistically evaluate the 4D intelligence in VLMs, unifying spatial understanding, temporal continuity, and motion reasoning. By curating tasks that demand precise analysis of dynamic interactions (e.g., direction prediction, perspective anticipation, and motion reasoning), VLM4D exposes critical gaps in current models' ability to internalize spatiotemporal relationships. Our work not only advances the granularity of VLM evaluation but also shares insights and potential solutions to improve the model performance.

3. The VLM4D Benchmark

We introduce VLM4D, the first benchmark specifically designed to test the spatiotemporal reasoning abilities of VLMs. VLM4D consists of 1,000 videos paired with over 2,000 question-answer pairs, each carefully designed to assess both spatial and temporal understanding jointly. The majority of these videos are sourced from datasets with rich spatiotemporal characteristics, thus ensuring a diverse range of motion-related scenarios. We augment the dataset with synthetic videos generated by a world-foundation model, Cosmos [2], that has been modified using techniques introduced in [25] to obtain more accurate correspondence between motion-oriented prompts and the resulting generated video. Figure 2 illustrates the composition of our dataset.

3.1. Benchmark Construction

Unlike prior work that often relies heavily on LLMs and VLMs to generate first iterations of benchmarks and datasets [9] followed by human quality control - we found that existing VLMs and automated methods showed significant limitations in terms of realiability and quality. This shortcoming necessitated direct human annotations that were then followed by augmentation by LLMs to ensure a high-quality benchmark. An overview of the benchmark curation pipeline is shown in Fig. 3.

Real Video Data Collection Real-world videos were sourced from datasets with rich spatiotemporal characteristics that ensured diverse motion and perspective varia-

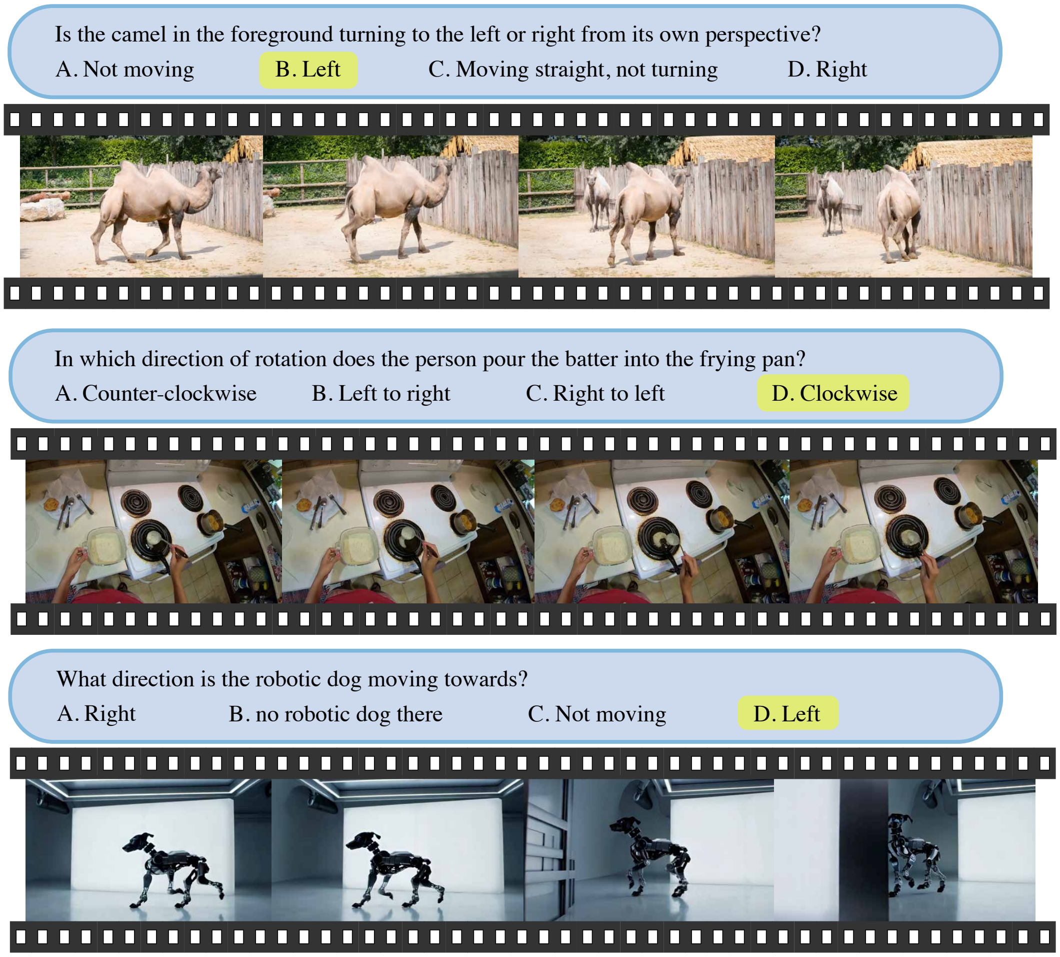

Is the camel in the foreground turning to the left or right from its own perspective?
A. Not moving
B. Left
C. Moving straight, not turning
D. Right
In which direction of rotation does the person pour the batter into the frying pan?
A. Counter-clockwise
B. Left to right
C. Right to left
D. Clockwise
What direction is the robotic dog moving towards?
A. Right
B. no robotic dog there
C. Not moving
D. Left

Figure 4. **Qualitative Examples of Dataset Annotations.** (Top) A third-person video with translational annotations ("camel turning left from its perspective"). (Middle) A first-person video with a rotational question ("clockwise rotation of ladle"). (Bottom) A synthetic scene with action recognition "robotic dog moving left").

tions. For egocentric data, we primarily relied on the Ego4D dataset [23], while most object-centric data points were collected from the Davis [54] and YouTube-VOS [70] datasets. To minimize confounders and to focus attention of VLM abilities to only spatiotemporal reasoning, we preprocessed the videos by temporally segmenting and centering them around the most relevant action thus resulting in videos with an average duration of 5-15 seconds. This ensures that the key event described in the question is clear and reduces ambiguities or confounders that would reduce VLM accuracy.

Synthetic Video Generation For synthetic video generation, we use Cosmos [2] as our video generation backbone. To ensure that the generated videos align with the intended object moving directions, we incorporate input bounding boxes as additional spatial guidance. Specifically, we follow the approach introduced in [25] modifying the diffusion forward steps to enforce object localization constraints at each timestep, ensuring consistency between the generated object direction and the user-specified trajectory. The average duration of generated synthetic videos is 5 seconds. To maintain high-quality outputs, we perform a manual verification step after generation, filtering out low-quality videos and retaining only those that accurately match the specified directions. Once a video is generated, we use an LLM (GPT-4o) to generate two types of questions for evaluation: Direct questions, which are derived directly from the textual prompt used to generate the video; Counterfactual questions, which involve querying about non-existent objects in the generated scene. Both question types follow the format: "What direction is the ⟨Object Name⟩ moving?", where the model must select one of four possible answers: "left", "right", "not moving", or "no ⟨Object Name⟩ there."

QA Generation and Quality Control Question-answer pairs are primarily constructed through human annotations. The question answer pairs are then supplemented with alternative answers by an LLM (GPT-4o) for multiple choice

(MC) questions. To ensure high-quality annotations, a rigorous human verification process was applied where ambiguous videos were filtered out and vague, misleading, or incorrect QA pairs were refined to allow for spatial and temporal alignment between the language and visual content. Figure 4 showcases some qualitative examples of annotations for different types of videos.

Assessing Human Performance To establish a human performance baseline on our benchmark, we conducted an evaluation in which participants independently answered 100 randomly sampled questions from the dataset. The accuracy of human responses was then aggregated to approximate the performance of human spatiotemporal reasoning on thedataset.

3.2. Categorizing Spatiotemporal Performance

To systematically evaluate spatiotemporal reasoning capabilities, we first categorize videos into two primary groups: egocentric (first-person) videos and exocentric (third-person) videos. Egocentric videos are sourced from the Ego-4D [23] dataset where scenes are captured from a head-mounted camera, thus offering dynamic video data that is inherently coupled with the individual's actions. Exocentric videos encompass a diverse range of recorded scenes, from sports footage to everyday scenes. Beyond this categorization, we also evaluate spatiotemporal performance across four dimensions: translational movement (TM), rotational movement (RM), spatiotemporal counting (STM), and false positives (FP). Translational movement assesses a model's ability to track linear motion within scenes, while rotation movement evaluates the understanding of changes in orientation and perspective shifts over time. Spatiotemporal counting extends these core motion-based tasks by requiring a more complex reasoning strategy to determine the number objects performing a translation or rotational movement. Lastly, the false positives category measures the model's reliability in recognizing whether any motion took place. By structuring the benchmark along these axes, we aim for a comprehensive framework for assessing spatiotemporal reasoning (Figure 5).

4. Evaluation of VLM4D Benchmark

4.1. Evaluation Setup

Benchmark Models We evaluate over 10 of the most recently released VLMs thus covering a wide range of model sizes, architectures, and training methodologies. For open-source models, we include Llama-3.2-Vision [22], DeepSeek-VL [45], InternVL2.5 [10], Pixtral [3], Aria [36], Idefics [31], H2OVL [20], Qwen2-VL [63], Qwen2.5-VL [71], VideoLLama2 [11], VideoLLama3 [75], Llava-One-Vision [35], Llava-NeXT-Video [79], InternVideo2 [65], and InternVideo 2.5 [66]. When available, we evaluate different parameter sizes for each model type, resulting overall in models ranging from 2 to 72 billion parameters. For closed-source VLMs, we evaluate GPT-4o [50], Gemini 2.0 Pro [59], and Grok-2-Vision.

Evaluation Settings The evaluations were performed in a zero-shot setting with the video or a set of sampled frames from the video followed by the prompt forming the input. For each model, we evaluate on two different inference settings. In the first setting, the model is directed to output the answer immediately without any reasoning (DO) and in the second evaluation setting, the model is directed to create intermediate reasoning steps, Chain of Thought (CoT) [68], before inferring the final answer. Additional details about the evaluation setup and prompts are provided in the Appendix.

Metrics Following prior work [72] and given the nature of our target task, we use multiple-choice questions for evaluation. The primary metric is accuracy on the multiple choice questions (MCQ). Given the two inference settings mentioned previously, we employ LLM-as-Judge following [80] to grade the VLMs' outputs. LLM-as-Judge was utilized instead of performing string or template matching as we found that especially during CoT, various VLMs may output all possible answers during the reasoning process in varying frequencies and with slight modifications to the format of the possible answer choices in MCQ. Each MCQ contains four possible answers.

4.2. Benchmark Results

VLMs Performance The evaluation results in Tab. 1 reveal several critical insights regarding the spatiotemporal reasoning capabilities of contemporary VLMs on the VLM4D benchmark. First, proprietary VLMs, particularly OpenAI's GPT-4o, consistently outperform open-source models across nearly all real-world categories, highlighting the performance gap between closed-source and publicly available VLMs. Among open-source models, InternVideo2.5-8B and Qwen2.5-VL-72B-AWQ emerge as notable contenders, with Qwen2.5-VL-72B-AWQ achieving exceptional results on synthetic data, surpassing even GPT-4o. However, all models significantly trail behind human-level performance, emphasizing substantial room for improvement, especially in nuanced spatiotemporal reasoning. These findings underscore a critical gap in current VLM architectures, reinforcing the need for further research into structured 4D scene representations and improved spatiotemporal grounding strategies. We additionally show in Fig. 5 for the top-performing models their strengths and weaknesses in the fine-grained categories mentioned in the

Organization	Model	Release	Real			Synthetic			Overall
			Ego-centric	Exo-centric	Average	Directional	FP	Average	
User Study	Human Performance		99.6	99.7	99.7	91.8	100	95.9	98.3
Random	Random Selection		24.4	23.2	23.6	25.5	24.7	25.1	24.2
Latest Proprietary VLMs									
OpenAI	GPT-4o	2024-11	**54.3**	**61.2**	**58.9**	47.8	43.0	45.4	53.9
Google	Gemini 2.0 Pro	2025-2	44.8	50.5	48.7	42.8	41.8	42.3	46.3
xAI	Grok-2-Vision	2024-12	44.1	48.8	47.3	49.0	60.5	54.8	50.0
Open-source Image VLMs									
Meta	Llama-3.2-11B-Vision	2024-9	35.2	36.1	35.8	38.3	55.8	47.0	39.9
Microsoft	Phi-3.5-Vision	2024-7	36.3	39.1	38.2	26.5	37.5	32.0	35.9
DeepSeek	DeepSeek-VL2-Tiny	2024-12	31.4	32.5	32.2	42.8	25.5	34.1	32.9
Shanghai AI Lab	InternVL2.5-38B	2024-11	42.8	53.2	49.7	37.5	55.5	46.5	48.6
	InternVL2.5-8B	2024-11	40.8	41.1	41.0	40.8	47.0	43.9	42.1
	InternVL2-8B	2024-6	33.2	38.2	36.5	34.8	58.0	46.4	40.2
Mistral AI	Pixtral-12B	2024-9	36.3	32.9	34.0	41.0	17.3	29.1	32.2
Rhymes	Aria	2024-11	42.3	44.0	43.5	35.3	56.3	45.8	44.3
HuggingFaceM4	Idefics3-8B	2024-8	34.3	36.2	35.6	33.5	47.3	40.4	37.4
H2O	H2OVL-Mississippi-2B	2024-10	37.0	33.3	34.5	27.3	41.0	34.1	34.4
Open-source Video VLMs									
Alibaba	Qwen2.5-VL-7B	2025-1	42.3	45.0	44.1	39.3	48.5	43.9	44.0
	Qwen2.5-VL-72B-AWQ	2025-1	49.9	48.7	49.1	**54.3**	**72.8**	**63.5**	**54.4**
	Qwen2-VL-7B	2024-8	36.1	38.2	37.5	38.5	40.3	39.4	38.2
	Qwen2-VL-72B-AWQ	2024-9	43.0	46.2	45.2	43.8	71.0	57.4	49.7
DAMO	VideoLLama3-2B	2025-1	48.6	43.7	45.3	29.0	69.8	49.4	46.8
	VideoLLama3-7B	2025-1	47.4	45.0	45.8	39.5	58.8	49.1	47.0
	VideoLLama2.1-7B	2024-10	43.0	36.0	38.2	31.5	40.0	35.8	37.3
	VideoLLama2-7B	2024-6	36.3	16.5	23.0	25.8	45.5	35.6	27.6
OpenGVLab	InternVideo2.5-8B	2025-1	52.8	50.1	51.0	45.3	30.0	37.6	46.1
	InternVideo2-8B	2024-8	37.2	37.9	37.6	40.5	2.8	21.6	31.7
LLaVA	LLaVA-One-Vision-7B	2024-9	32.5	33.1	32.9	32.8	36.0	34.4	33.5
	LLaVA-NeXT-Video-7B	2024-6	30.3	30.9	30.7	24.5	27.3	25.9	28.9
	LLaVA-NeXT-Video-34B	2024-6	37.2	34.9	35.7	31.5	56.3	43.9	38.7

Table 1. **Evaluation on VLM4D Benchmark** across various proprietary and open-source VLMs. Top three performers in each category are highlighted from dark (highest) to light (third highest). Human and random selection baselines are included for reference.

previous section. As expected, translational motion performs best, followed by rotational motion and spatiotemporal counting.

Human Level Performance We use **Prolific**, an online platform designed to connect academic researchers with user research participants for human-level performance evaluation. The participants are English-speaking random users verified by this platform without prior knowledge of computer vision. We asked 51 candidates to answer the spatial awareness questions in our benchmark. Each question has four choices, and the user may select only one correct answer. We collect their answers and report the average precision in Table. 1

5. Analysis: Why VLMs Don't Work Well?

5.1. Limited Spatiotemporal Cognition

Despite significant advances in VLMs, their ability to understand and reason about motion, spatial relationships, and temporal coherence remains fundamentally underdeveloped [8, 51]. Chain of Thought (CoT) [68] is widely employed as a method to improve accuracy through step-by-step reasoning. We showcase a comparison between CoT and DO

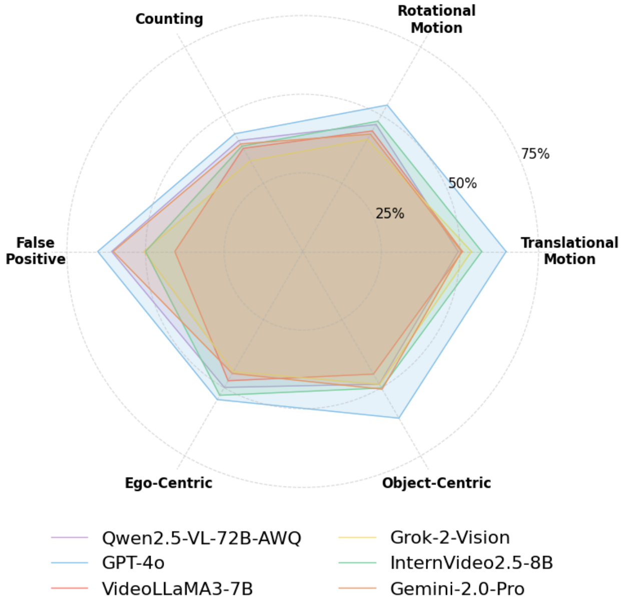


Figure 5. Model Accuracy Across Real Scene Question Categories of top-performing VLMs.

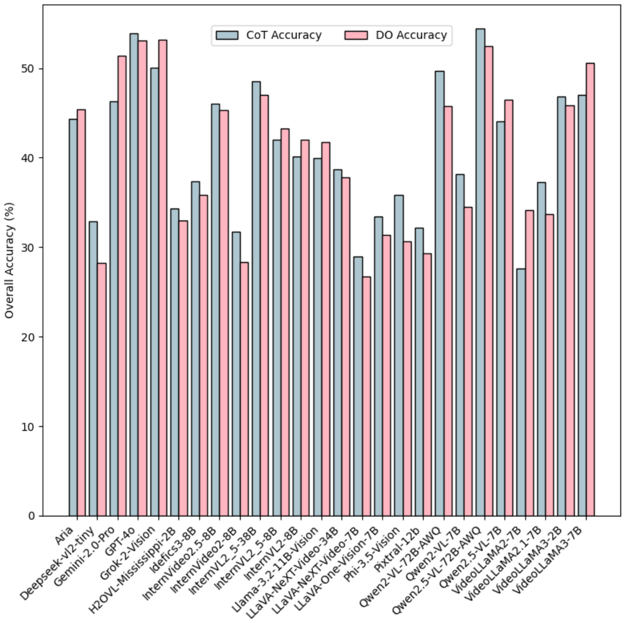


Figure 6. **Comparison of CoT and DO Accuracy Across Models.** Accuracy comparison between Chain-of-Thought (CoT) and Direct Output (DO) prompting across VLMs.

in Fig. 6. Overall, there is no indication of a large advantage of CoT over all evaluated models. Upon deeper exploration of the CoT reasoning of some models, we observe that the reasoning process was primarily flawed in the following ways: irrelevant information and arriving at conclusions that are inconsistent with the reasoning process. Larger models exhibited strategies that would be similar to how a human processes spatiotemporal information, but the resulting execution falls short of human performance. This demonstrates a disconnect between its visual and linguistic knowledge. We provide examples of this behavior in the supplement.

5.2. Deficiencies in Spatiotemporal Labeling

Another avenue of exploration we undertook is to understand the richness of spatiotemporal labels in popular SFT VLM datasets. Typically, video captioning occurs at the 'scene' level, lacking fine-grained temporal, spatial, and object-level details. We performed an extensive analysis, encompassing over 2 million samples [9, 12, 30, 38, 78]. We performed this analysis through string-matching of spatiotemporal descriptors related to directionality, translational motion, rotation, and perspective shifts and provide the overall results in Fig. 7. We then performed a manual finegrained evaluation of the ShareGPT4Video dataset [9] which we found had the highest density of spatiotemporal datasets. We found that from a sample of 100 labels that were detected as spatiotemporal, less than 10% of them were judged as accurate upon human evaluation. This result underscores the inadequacy of current dense captioning approaches, which frequently generate spatiotemporal descriptors without capturing precise motion dynamics. We provide more detailed analysis and explanations in the supplement.

6. Probing Future Solutions

To probe promising future solutions for enhancing spatiotemporal video understanding, we propose two ap-

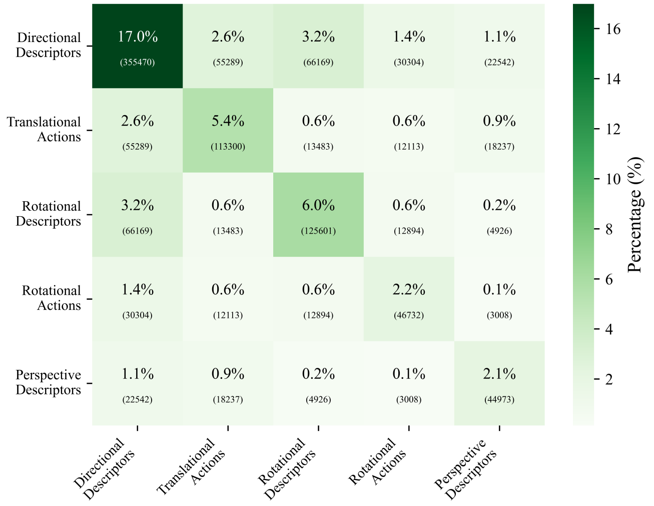


Figure 7. **Heatmap of Occurances of Spatial-Temporal Terms in popular video SFT datasets.**

proaches that address some of the shortcomings of current state-of-the-art VLMs: fine-tuning a VLM on data-rich in spatiotemporal actions and the other leveraging 4D reconstruction and feature fields jointly with a VLM. SFT refines the model's abilities by training on datasets that contain temporally and spatially rich actions and interactions. By integrating structured visual representations and targeted fine-tuning, these approaches enhance video-language models' ability to interpret motion. The second method lifts the feature space of VLMs into a temporally coherent 4D feature field, providing structured scene representations that improve motion and spatial reasoning in the stage of decoding and inference.

Spatial-Temporal SFT We evaluate on a subset split of the real dataset by splitting the real-world dataset into a training and testing split (80% / 20%) and we try settings using synthetic/real/both for training. We conducted the experiments using Qwen 2VL (7B) and Qwen 2.5VL (7B) through LLama-Factory [81], and compared the performance before and after supervised fine-tuning in Tab. 2. The results demonstrated an improvement in accuracy in spatiotemporal reasoning, suggesting that performance gains can be obtained through targeted training. However, the addition of synthetic data does not necessarily increase performance over using real data alone, suggesting the importance of synthetic data quality.

4D Feature Fields Reconstruction Recent advances in 3D/4D reconstruction methods, such as Feature4X [82], have significantly enhanced Vision-Language Model (VLM) performance on visual question answering (VQA) tasks by integrating structured 4D scene representations into the model's inference stage. Inspired by these promising results, we investigate incorporating spatiotemporal awareness into the InternVideo2-8B model [65], employing the 4D feature lifting strategy proposed by Feature4X. To assess this approach, we evaluate performance on a subset of the VLM4D benchmark, specifically leveraging all 50 videos from the DAVIS 2016 dataset [53]. Our experimental evaluation compares the inference results across three distinct input modalities: original 2D videos, reconstructed global-view RGB videos (4D), and reconstructed global semantic feature fields. As demonstrated in Table 3, the highest accuracy consistently results from the reconstructed semantic feature fields, highlighting the clear advantages of structured 4D representations. These findings confirm that global 4D feature field reconstruction enhances contextual understanding and mitigates artifacts associated with RGB rendering during reconstruction. However, the current approach requires per-scene optimization as a post-processing step, limiting its generalizability and making it computationally intensive.

Model	FF	MC
Original Model		
Qwen 2VL (7B)	31.9	38.3
Qwen 2.5VL (7B)	31.6	43.4
Finetuned Model		
Qwen 2VL (7B) (R)	**50.7**	53.5
Qwen 2VL (7B) (S)	38.9	41.0
Qwen 2VL (7B) (R+S)	49.7	52.8
Qwen 2.5VL (7B) (R)	48.9	**56.3**
Qwen 2.5VL (7B) (S)	35.4	42.0
Qwen 2.5VL (7B) (R+S)	39.2	48.3

Table 2. **SFT on Spatial-Temporal Datasets.** MC and FF refer to multiple-choice and freeform accuracy, respectively. R means SFT using the real-world dataset, S denotes the synthetic dataset, R+S represents using both.

Input Modality	Accuracy
Chain of Thought Response	
Original 2D Video	36.0
Global View Video	32.7
Global Feature Field	**37.4**
Direct Output Response	
Original 2D Video	24.3
Global View Video	23.8
Global Feature Field	**29.0**

Table 3. **InternVideo2 Accuracy with 4D Reconstruction.** Comparison of InternVideo2 accuracy given different input modalities from the same dataset.

7. Conclusion

Through the construction of the VLM4D benchmark, we evaluate the spatiotemporal reasoning capabilities of various Vision-Language Models (both open-source and proprietary). While more recently released models demonstrate improved performance over their counterparts, they remain significantly behind human proficiency. Overall, our work questions whether VLMs posses spatiotemporal reasoning abilities that are imperative to have for more sophisticated visual agents in fields ranging from robotics to interactive AI systems that require a deep understanding of dynamic visual environments. We hope to inspire future work to explore novel approaches for integrating spatiotemporal grounding, thereby enhancing their spatiotemporal reasoning capabilities and facilitating robust deployment.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. 3

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3, 4

[3] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 5

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[6] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10(12):551–557, 2006. 2

[7] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18030–18040, 2022. 2

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 6

[9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 3, 7

[10] 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. 5

[11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3, 5

[12] Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2024. 7

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[14] Julian De Freitas, Nicholas E. Myers, and Anna C. Nobre. Tracking the changing feature of a moving object. *Journal of Vision*, 16(3):22, 2016. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[17] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023. 3

[18] Jennifer J. Freyd and Ronald A. Finke. Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):126–132, 1984. 2

[19] 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. *arXiv preprint arXiv:2405.21075*, 2024. 3

[20] Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. H2ovl-mississippi vision language models technical report. *arXiv preprint arXiv:2410.13611*, 2024. 5

[21] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping

Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 3

[22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 4, 5

[24] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024. 3

[25] Xuehai He, Shuohang Wang, Jianwei Yang, Xiaoxia Wu, Yiping Wang, Kuan Wang, Zheng Zhan, Olatunji Ruwase, Yelong Shen, and Xin Eric Wang. Mojito: Motion trajectory and intensity control for video generation. *arXiv preprint arXiv: 2412.08948*, 2024. 3, 4

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[27] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. 2

[28] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms. *arXiv preprint arXiv:2405.03690*, 2024. 3

[29] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3

[30] Yi Wang Yizhuo Li Wenhai Wang Ping Luo-Yali Wang Limin Wang KunChang Li, Yinan He and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 7

[31] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 5

[32] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 2

[33] Alan M. Leslie. Spatiotemporal continuity and the perception of causality in infants. *Perception*, 13(3):287–305, 1984. 2

[34] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 3

[35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 5

[36] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 5

[37] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3

[38] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023. 7

[39] 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. 3

[40] Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. Videoeval: Comprehensive benchmark suite for low-cost evaluation of video foundation model. *arXiv preprint arXiv:2407.06491*, 2024. 3

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[42] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024. 2

[43] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 3

[44] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 3

[45] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 5

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*, 2019. 2

[47] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[48] D. Marr and S. Ullman. Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 211(1183):151–180, 1981. 2

[49] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 3

[50] OpenAI. Hello gpt-4o. Technical report, 2024. 5

[51] Yann Dubois Nikhil Mehta Tong Xiao Philippe Hansen-Estruch Licheng Yu Xiaofang Wang Felix Juefei-Xu Ning Zhang Serena Yeung-Levy Orr Zohar, Xiaohan Wang and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024. 6

[52] Shivansh Patel, Xinchen Yin, Wenlong Huang, Shubham Garg, Hooshang Nayyeri, Li Fei-Fei, Svetlana Lazebnik, and Yunzhu Li. A real-to-sim-to-real approach to robotic manipulation with vlm-generated iterative keypoint rewards. *arXiv preprint arXiv:2502.08643*, 2025. 3

[53] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 8

[54] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4

[55] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[57] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007. 2

[58] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding. *arXiv preprint arXiv:2406.13807*, 2024. 3

[59] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3, 5

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[61] Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. Vlm see, robot do: Human demo video to robot action plan via vision language model. *arXiv preprint arXiv:2410.08792*, 2024. 3

[62] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024. 3

[63] 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. 3, 5

[64] Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4d dynamic scenes understanding with physics priors for video question answering. *arXiv preprint arXiv:2406.00622*, 2024. 2

[65] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 5, 8

[66] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 5

[67] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2

[68] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5, 6

[69] Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Advances in Neural Information Processing Systems*, 37:109922–109947, 2024. 3

[70] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 4

[71] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 5

[72] 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. *arXiv preprint arXiv:2412.14171*, 2024. 3, 5

[73] 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. *arXiv preprint arXiv:2501.00599*, 2024. 3

[74] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3

[75] 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. 3, 5

[76] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[77] Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Yilun Du, and Chuang Gan. Combo: compositional world models for embodied multi-agent cooperation. *arXiv preprint arXiv:2404.10775*, 2024. 3

[78] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 7

[79] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 5

[80] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025. 3, 5

[81] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 8

[82] Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan, Suya You, Zhangyang Wang, Leonidas Guibas, and Achuta Kadambi. Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 8

[83] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3