Coarse Correspondences Boost Spatial-Temporal Reasoning in Multimodal Language Model

Benlin Liu[1]*, Yuhao Dong[2,3]⋆, Yiqin Wang[2]⋆,
Zixian Ma[1], Yansong Tang[2], Luming Tang[5], Yongming Rao[3], Wei-Chiu Ma[5,6], Ranjay Krishna[1,5]
[1]University of Washington, [2]Tsinghua University, [3]Tencent,
[4]Allen Institute for AI, [5]Cornell University
coarse-correspondence.github.io

Appendix

A. Broader Impact

Our method aims at improving the trustworthiness and reliability of deployment of MLLMs in real world application, including but not limited to Vision Pro, autonomous driving, and also humanoid robots. To have a virtual assistant like JARVIS in Marvel films, it's necassry to align the understanding of vision-language model with human's understanding, so that we can ensure safe application of these applications. Further, we are committed to reducing the carbon emissions produced by these models. By employing our coarse correspondence prompting method, we use a much smaller tracking module to reduce the number of input used as input to large GPT model. Besides, we also improve the speed and lower the cost of calling OpenAI API to understand a 3d scene. This enables democratize MLLMs so that more people and small companies can create their own real-world applications based on GPT-4V. We hope our work can make large AI models more effectively used for social good.

Still, we would like to point out that with the development of MLLMs, increased reliance on advanced MLLMs could also lead to a reduction in human skills, especially in interpreting and interacting with visual content. Overdependence on these models might erode critical thinking and analytical abilities in the long term.

B. Related Work

Video understanding. Videos carry rich information about both the 3D structure as well as temporal changes in the physical world. To perform better long-horizon reasoning, work has begun incorporating video inputs into MLLMs. Recent work [9] has improved performance on video dense captioning [8] and videoQA [5, 21]. To further advance the understanding of temporal relationships in videos, EgoSchema [12] introduced a benchmark for long video understanding, which is more challenging than previous video-language benchmarks. Meanwhile, understanding 3D spatial relationships in videos received relatively less attention. 3D-LLM [6] converts multiview images into 3D point clouds and then feeds them into LLMs, demonstrating better results on the ScanQA [1] benchmark for 3D understanding. OpenEQA [11] is also a benchmark dedicated to evaluating MLLM's understanding of 3D physical space, with outputs that are more open-vocabulary compared to ScanQA. In this paper, we propose a framework that does not require any training in modifying MLLMs; it extracts meaningful information from videos using off-the-shelf tracking models and achieves state-of-the-art performance on the benchmarks mentioned.

Visual correspondences. Visual correspondences have been a vital area of research in computer vision for a few decades. Applications such as Structure-from-Motion[17] utilize correspondences to better reconstruct 3D scenes. In the past, we relied on handcrafted features like SIFT [10] or SURF [3] to obtain good correspondence. Today, features extracted from deep models [18] can also provide increasingly accurate correspondences. Generally, people aim to achieve precise geometric and semantic correspondences at the pixel level. However, in this paper, we use coarse visual correspondence to prompt MLLMs, which can be easily obtained from off-the-shelf video tracking models [22].

C. Coarse Correspondence Implementation Details

As discussed in Method section, visualizing our proposed Coarse Correspondence on images will involve a centering algorithm. The inputs are selected instance segmentation masks that originally obtained from tracking model. A center of the instance mask needs to be determined in order to place the coarse correspondence marker. It is worth noting that the instance mask does not necessarily form a connected

*Equal contribution.

component, which makes the centering procedure worth explaining.

```
# Find center of a mask,
# May contains multiple connected components.
def find_center(mask):
    # Go through the middle column, try to find center1
    exist_y = []
    x_center = median(left_bound, right_bound)
    for y in range(upper_bound, lower_bound):
        if (x_center, y) in mask:
            exist_y.append(y)
    if exist_y is not empty:
        y_center = median(exist_y)
        center1 = (x_center, y_center)
    else:
        center1 = None

    # Go through the middle row, try to find center2 (skip)
    if avg(center1, center2) in mask:
        return avg(center1, center2)
    elif center1 in mask:
        return center1
    elif center2 in mask:
        return center2
    else:
        center_naive = ((left_bound + right_bound)//2,
                        (upper_bound + lower_bound)//2)
        return center_naive
```

Figure 1. The pseudo code of our proposed algorithm to find the center of a given object mask. The Coarse Correspondence will be further added to the object center

As shown in the pseudo code in Figure 1, firstly we calculate the medium x-index of the masked pixels and loop through this column, trying to find the first center point. Similarly, we calculate the medium y-index of the masked pixels and loop through this row, trying to find another center point. Normally we return the average location of these two centers. If either of these centers failed to be positioned in the masked area (which may happens when the mask is not a connected components), we adopt the other one. If both of them failed to deliver, we adopt a naive center by simply averaging the four boundary.

D. More Results on Proprietary Models

We further evaluate COARSE CORRESPONDENCES by augmenting both Gemini and Claude models. Following prior works, we adopt BLEU [14] scores, METEOR [2], ROUHE-L [15] and CIDEr[20] as our evaluation metrics. As shown in Tab. 1, COARSE CORRESPONDENCES constantly improves the performance of both models, which demonstrates the generalizability of our method.

E. Qualitative Case Study

To further demonstrate the effectiveness of our proposed Coarse Correspondence under sparse image input, we defined two challenging tasks and one qualitative case study for each task.

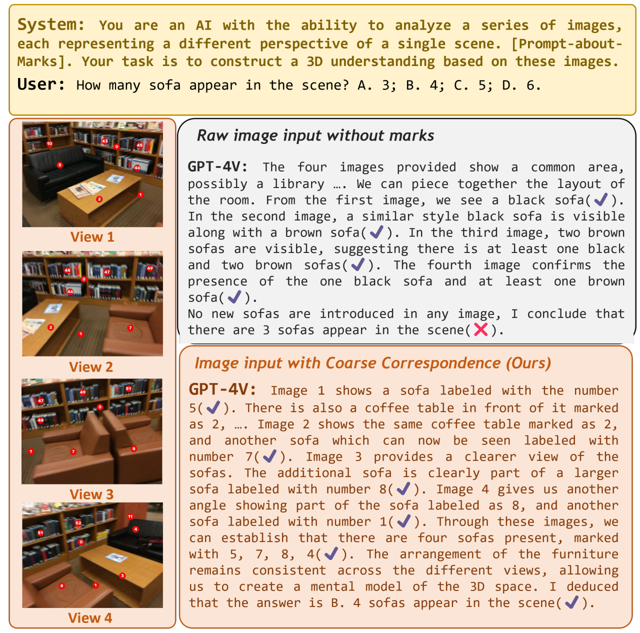


Figure 2. **Task: Duplicate Objects Counting.** There are 2 brown sofas and 2 black sofas. The brown sofas in View 2&4 are duplication of those in View 3. Only with the help of the Coarse Correspondence can GPT-4V understand duplicate objects between different views across a single 3D scene.

The results of these case studies are shown in Fig. 2 and Fig. 3. Detailed illustration of the results are provided in the figure captions. The first case study is about the task of Duplicate Objects Counting, where the model needs to count the number of objects in a 3D scene. Only equipped with coarse correspondence can GPT-4V get a comprehensive understanding of the 3D scene, excludes the duplicate objects, and give the right answer.The second case study is about the task of Relative Location Modeling, where the model needs to understand the relative location of objects in a 3D scene. It is obvious that without the correspondence markers, GPT-4V fails to response from 3D perspective with only raw 2D images.These case studies demonstrate that our proposed Coarse Correspondence can elicit MLLMs in understanding 3D scenes from sparse image inputs.

We also prove that our Coarse Correspondence method works well with hand-crafted correspondence marks as shown in Figure 4. This further proves that our proposed method are style-agnostic as long as the marks is able to deliver the spatial correspondence knowledge.

F. Analysis on Different Question Types

In Figure 5, we analyze the improvements brought by COARSE CORRESPONDENCES across different question types in OpenEQA. It can be observed that COARSE CORRESPONDENCES enhances performance for all question types,

Model	Frame	BLEU-1	BLEU-2	METEOR	ROUGE-L	CIDEr
Proprietary Multi-modal Models						
Gemini	8	24.1	13.5	11.3	35.4	68.3
Gemini+*Coarse Correspondences*	8	**25.4**	**15.7**	**12.0**	**37.1**	**75.5**
Claude	8	19.8	11.1	10.0	29.3	57.7
Claude+*Coarse Correspondences*	8	**27.1**	**23.9**	**11.7**	**33.1**	**65.7**

Table 1. **Comparison on ScanQA validation set.** We conduct experiments on the ScanQA validation set to demonstrate the effectiveness of COARSE CORRESPONDENCES with different MLLMs. Our method enables both proprietary models and open-source models to surpass all 3D-specific models.

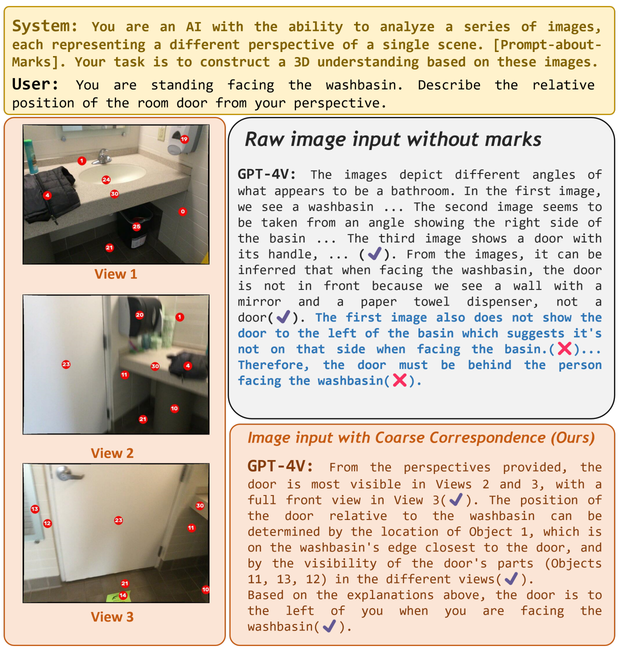


Figure 3. **Task: Relative Location Modeling.** From View 1 & 2 we can tell that the room door is on the left-hand-side when facing the washbasin. Only with the help of the Coarse Correspondence can GPT-4V understand relative location between objects appear in different views across a single 3D scene.

with the most significant improvement in spatial understanding. This demonstrates the generalizability of our approach. Furthermore, it confirms that our approach effectively boosts the spatial reasoning capabilities of MLLMs. Of course, there is still much room for improvement in enhancing the spatial reasoning capabilities of MLLMs.

G. The SOT benchmark for Spatial Orientation Test

Considering that a crucial aspect of embodied tasks like navigation is the judgment of left-right orientation, we aimed to gain a deeper understanding of how COARSE CORRESPONDENCES influences MLLMs' comprehension of left-right spatial orientation. Specifically, we focused on two key questions: 1) Are MLLMs robust to camera motion bias? Ideally, MLLMs' understanding of left-right orientation in 3D space should be independent of whether the camera is moving from left to right or right to left, meaning MLLMs should be robust to camera motion bias. 2) Do MLLMs possess spatial perspective-taking ability, i.e., the ability to imagine how an object or scene would appear from a perspective different from the current camera viewpoint? Numerous studies in humans [13, 19] have shown that this ability is closely related to the development of spatial intelligence in children.

However, current benchmarks face three issues: 1) They may have been partially used in MLLM training data, 2) Current benchmarks lack annotations regarding whether the 3D space scan was conducted from left to right or right to left, making it difficult to analyze the impact of camera motion on MLLMs, and 3) Existing benchmarks evaluate a model's 3D spatial awareness from the perspective of the camera-wielding observer.

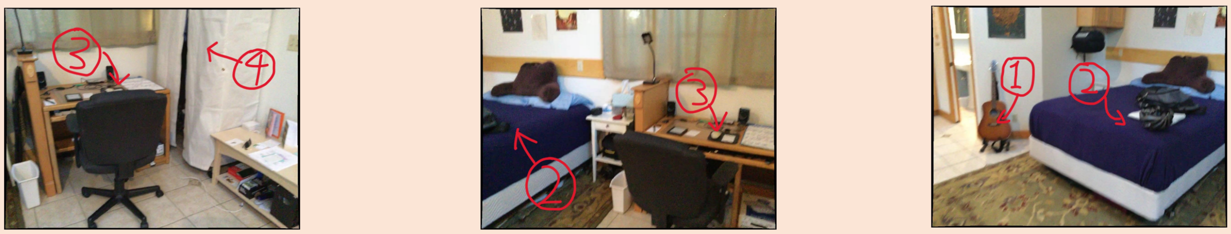
Therefore, we introduce a new diagnostic benchmark to evaluate MLLMs: Spatial Orientation Test (SOT). Once again, we show that COARSE CORRESPONDENCES improves GPT-4V,O's abilities on this new benchmark.

Data curation. We manually curated ten real-world scenes, both indoor and outdoor, using different mobile devices at various viewpoints. We instructed 10 human participants to take two videos in their environment from two viewpoints. When in each viewpoint, they were asked to remain in place as they laterally pan their mobile devices to scan their 3D environment. From 20 collected scenes, we filtered to and retained 10 scenes that satisfied the following four criteria: First, we could uniquely describe one viewpoint from the perspective of the other and vice-versa. For example, in Figure 6, we define the other viewpoints as 'a person stepping out of an elevator.' Second, we ensured that no single frame captured the entire 3D space, ensuring that models can not short-cut answers using any single view. Third, all scans move the camera from left to right. Fourth, to avoid privacy concerns, we ensured that no people appeared in the videos. Each scene scan lasts between 3 to 5 seconds.

For each scene, we designed **five carefully crafted questions**, each asking the model to determine if one object is to the left or to the right of another from a specific viewpoint.

System: You are an AI with the ability to analyze a series of images, each representing a different perspective of a single scene. [Prompt-about-Marks]. Your task is to construct a 3D understanding based on these images.

User: You are sitting on the chair in front of the desk. Describe the relative position of these objects from your perspective: the closet, the guitar, and the bed.



View 1 View 2 View 3

Raw image input without marks

GPT-4V: The closet is to your left, partially obscured by a white curtain or door(✗). The guitar is to your right, leaning against the wall next to the bed(✗). The bed is also to your right but behind the guitar(✗).

Image input with Hand-crafted Coarse Correspondence (Ours)

GPT-4V: The closet, labeled with the number 4, is to your right side, indicated by the positioning in the first image(✓). The guitar, labeled with the number 1, is behind you to your left. This is clear from the third image where the guitar is shown behind the bed, which is opposite the desk(✓). The bed, labeled with the number 2 , is behind you. When sitting at the desk, you would turn around to see the bed, as shown in the second and third images(✓).

Figure 4. **Hand-crafted coarse correspondence label**

The first three questions are from the observer's (camera's) perspective, while the final two describe the perspective in language, thereby, testing for a model's spatial perspective-taking ability. Human performance on these questions is 100%. We design SOT questions to have a bias towards asking about relationships between objects that appear in the first last frame of the scan, ensuring that the has to use multiple frames to answer. In total, across the 10 scenes, SOT has a modest 50 questions.

Results. As shown in Table 2, COARSE CORRESPONDENCES performs very well even on in-the-wild data. For instance, when using only the first and last frames, our method results in a 13.4 % improvement. As illustrated in Figure 6, our method enables GPT-4O to understand the 3D spatial structure represented by the images using minimal overlap, whereas GPT-4O alone performs only slightly better than random guessing.

Models	Frame	Origin	Reverse	Harmonic Mean
GPT-4O	2	58.2	50.0	53.8
GPT-4O+CC	2	**71.6**	**70.6**	**71.1**
GPT-4O	4	58.0	50.4	53.9
GPT-4O+CC	4	**71.2**	**71.2**	**71.2**

Table 2. **Comparisons on SOT.** COARSE CORRESPONDENCES shows strong capability of enhancing 3D spatial understanding of MLLMs. It can also ease the striking finding of camera motion bias of current MLLMs.

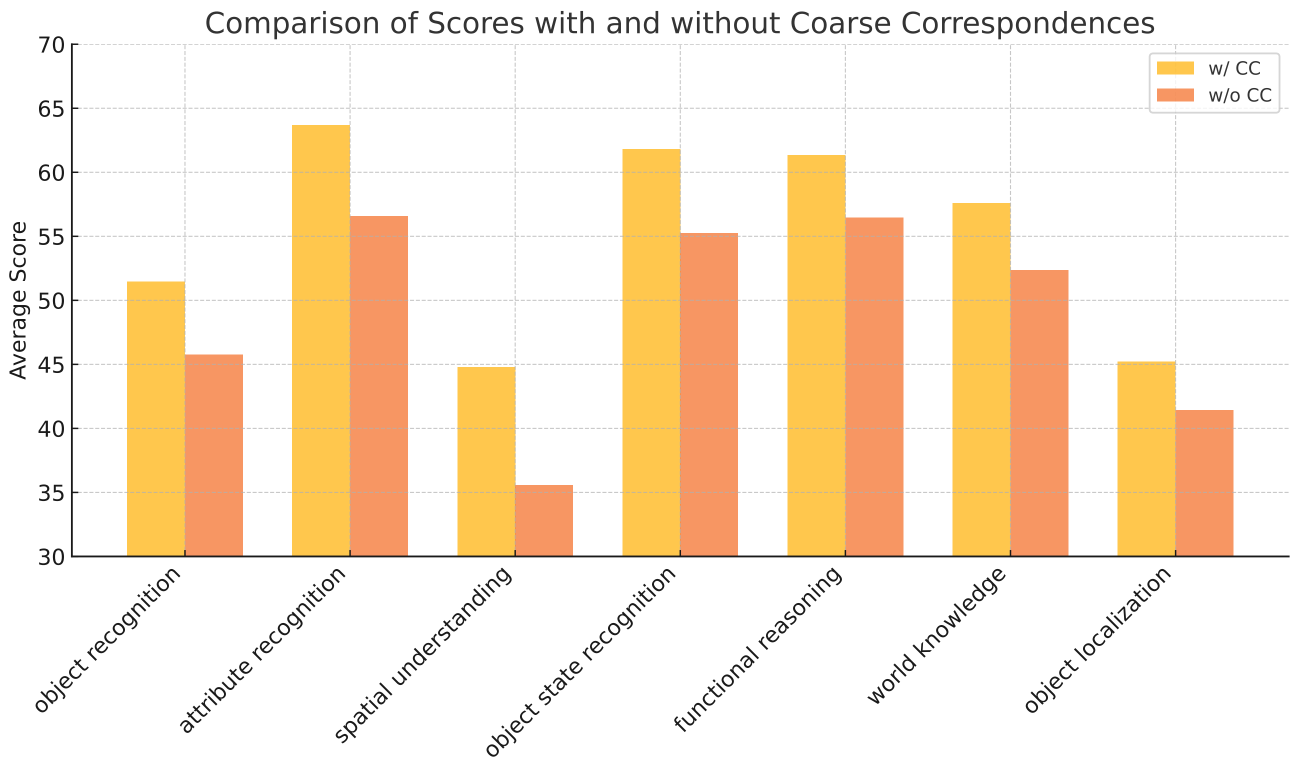


Figure 5. **Breakdown Analysis of OpenEQA Results.** We provide a detailed analysis of the improvements across different question types after applying COARSE CORRESPONDENCES.

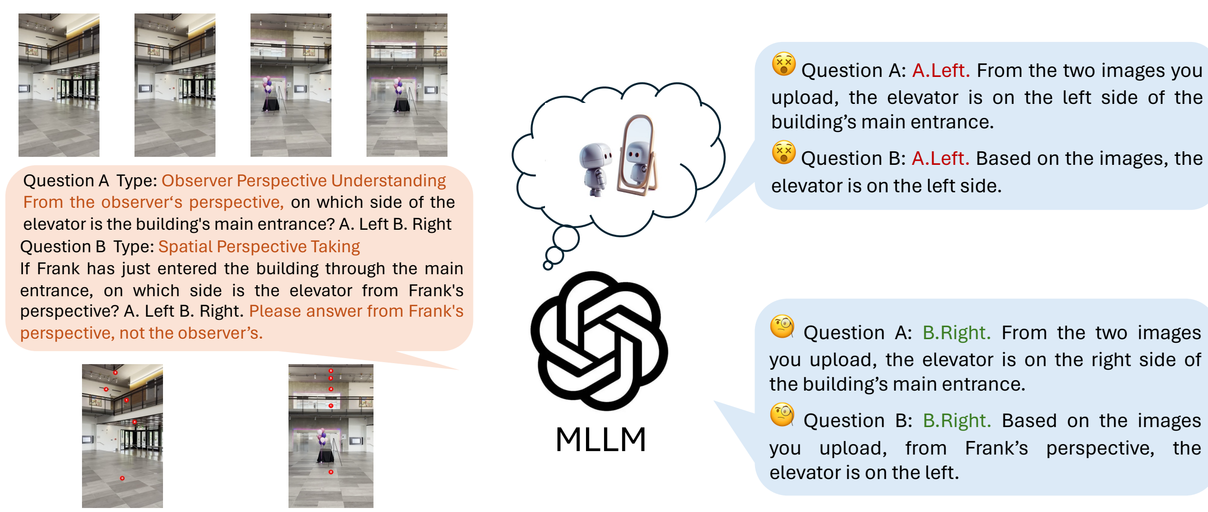


Figure 6. **Illustration of our SOT dataset.** We mention two types of questions: Observer perspective understanding and spatial-perspective taking. COARSE CORRESPONDENCES demonstrates superior effectiveness on the dataset.

More importantly, according to Table 2, we found that current MLLMs achieve significantly higher accuracy on videos filmed from left to right compared to those filmed from right to left, indicating that even models like GPT-4O have a strong camera motion bias. Our method greatly mitigates this issue. By calculating the harmonic mean of

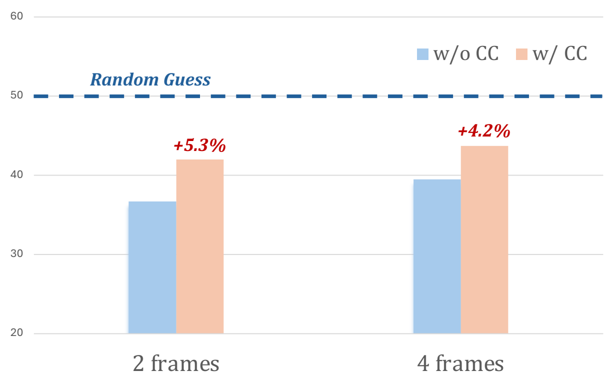


Figure 7. **Comparisons on SOT's spatial perspective-taking questions.** COARSE CORRESPONDENCES improves performance but GPT-4O still performs below random chance.

results from both left-to-right ($L- > R$) and right-to-left ($R- > L$) camera pans, we found that our method brought a 17.3 % improvement, indicating that COARSE CORRESPONDENCES helps MLLMs learn a more equivariant visual representation from image sequences.

Additionally, we isolated the performance on the two perspective-taking questions per scene in Figure 7. We discovered that current MLLMs still lack the ability for spatial perspective-taking. While COARSE CORRESPONDENCES improves GPT-4O's perspective-taking capability, the results are bittersweet, as they still perform worse than random guessing. This suggests that embodied spatial awareness has yet to emerge in MLLMs—at least for now—highlighting a potential direction for future research.

H. More Discussions

Limitations. Our method relies on off-the-shelf video tracking models to obtain instance-level correspondences. Although the performance of tracking models has significantly improved with the advent of tools like SAM [7], achieving good results on long-form in-the-wild videos remains challenging. This is particularly evident on the 180-second EgoSchema benchmark, where Track-Anything often loses track of objects after 100 seconds, leading to inconsistent instance segmentation masks between the beginning and end of the video clip. Despite observing consistent and significant improvements on EgoSchema, we believe that accurate correspondence would further enhance the benefits of our approach.

Relation to SlowFast SlowFast [4] is a framework for video recognition that includes two parallel pathways: a Slow pathway that captures motion information at a high frame rate and a Fast pathway that captures semantic information at a low frame rate. The information from both pathways is fused through lateral connections for downstream video recognition tasks. In a way, our coarse correspondence prompting can be seen as another form of SlowFast. However, unlike SlowFast, where the Slow and Fast pathways operate in parallel, our framework operates sequentially. First, it captures low-level, class-agnostic motion information at a high frame rate using a lightweight tracking model. Then, at a low frame rate, it performs recognition and reasoning requiring semantic understanding using larger MLLMs. The two stages are bridged through visual prompting. Moreover, while SlowFast learns a representation of the input video for pure vision recognition tasks such as action classification and detection, our coarse correspondence framework aims to better understand the 3D spatial structure and temporal information contained in the input video to achieve spatiotemporal perception and reasoning simultaneously.

Eulerian vs Lagrangian If deep learning-based methods represent camera or object motion in videos from an Eulerian viewpoint—i.e., expressing how features at fixed locations evolve over time through a multi-dimensional tensor—then our framework adds a Lagrangian viewpoint to this representation. The Lagrangian viewpoint describes the trajectories of entities moving through space and time in the video. Previously, the Lagrangian viewpoint in video descriptions has been shown to better aid human action recognition [16]. Here, we demonstrate that it can more generally help MLLMs understand the 4D spatiotemporal context represented in videos.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 1

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 2

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 1

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 6

[5] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 1

[6] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 6

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos, 2017. 1

[9] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1

[11] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[12] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. 1

[13] Nora Newcombe. The development of spatial perspective taking. *Advances in child development and behavior*, 22: 203–247, 1989. 3

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2

[15] Zhaopeng Qiu, Xian Wu, and Wei Fan. Automatic distractor generation for multiple choice questions in standard tests. *arXiv preprint arXiv:2011.13100*, 2020. 2

[16] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition, 2023. 6

[17] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[18] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1

[19] Barbara Tversky and Bridgette Martin Hard. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, 110(1):124–129, 2009. 3

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[21] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. 1

[22] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 1