
COARSE CORRESPONDENCES BOOST 3D SPACE-TIME UNDERSTANDING IN MULTIMODAL LANGUAGE MODEL

Anonymous authors

Paper under double-blind review

ABSTRACT

Multimodal language models (MLLMs) are increasingly being applied in real-world environments, necessitating their ability to interpret 3D spaces and comprehend temporal dynamics. Current methods often rely on specialized architectural designs or task-specific fine-tuning to achieve this. We introduce COARSE CORRESPONDENCES, a simple lightweight method which enhances MLLMs’ understanding of 3D and temporal concepts using only 2D images, without modifying the architecture or task-specific fine-tuning. Our method uses a lightweight tracking model to identify primary object correspondences between frames in a video or across different image viewpoints, and then conveys this information to MLLMs through visual prompting. We demonstrate that this simple training-free approach brings substantial gains to GPT4-V/O consistently on four benchmarks that require 3D and temporal understanding, including **+20.5%** improvement on ScanQA, **+9.7%** on OpenEQA’s episodic memory subset, **+6.0%** on the long-form video benchmark EgoSchema, and **+11%** on the R2R navigation benchmark. Additionally, we show that COARSE CORRESPONDENCES can also enhance open-source MLLMs’ understanding of 3D space (by **+6.9%** on ScanQA) when applied in both training and inference and that the improvement can generalize to unseen datasets such as SQA3D (**+3.1%**). Taken together, we show that COARSE CORRESPONDENCES effectively and efficiently boosts models’ performance on downstream tasks requiring 3D and/or temporal understanding.

1 INTRODUCTION

Intelligence is multi-faceted. While multi-modal large language models (OpenAI, 2024) have shown remarkable linguistic, logical and even mathematical intelligence, many remain doubtful about their visual and spatial intelligence. Despite their excellent performance on visual-linguistic tasks, many recent works (Majumdar et al., 2024; Singh et al., 2024) demonstrate that state-of-the-art MLLMs still struggle at 3D and long video benchmarks, performing only marginally better than blind text-only baselines. These results suggest that 3D and temporal understanding are two major bottlenecks on MLLMs’ path to general visual intelligence.

To enhance MLLMs’ 3D understanding, researchers have mainly explored three approaches: providing MLLMs with 3D data as input (Hong et al., 2024), designing specialized architectures for 3D tasks (Hong et al., 2024), or employing supervised fine-tuning with 3D data (Chen et al., 2024). Similarly, to boost MLLMs’ temporal understanding, prior works have proposed new model architectures designed for long video understanding (Papalampidi et al., 2023; Balažević et al., 2024), or adopted Socratic-based methods (Zhang et al., 2024; Kahatapitiya et al., 2024) (i.e., converting each frame of a video into text using a caption model, and then using text-only LLMs to summarize).

In contrast to prior works, we propose a simple but effective training-free visual prompting method COARSE CORRESPONDENCES to boost 3D and temporal understanding in MLLMs. COARSE CORRESPONDENCES uses a tracking model to extract object-level correspondences across multiple images, and then represent the most salient correspondence relationships on the images through visual prompting. Our method significantly boosts MLLMs’ understanding of 3D spacetime with only 2D image inputs and without any specialized architectural design or task-specific fine-tuning.

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

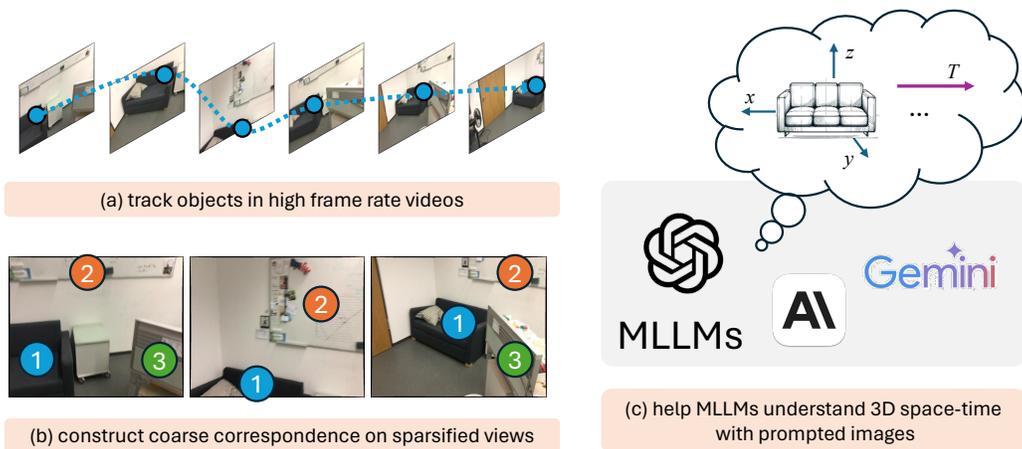


Figure 1: We combined light-weight video tracking models and multimodal LLMs to achieve a better understanding of 3D spacetime. (a) We use a tracking model at a high frame rate to obtain instance segmentation masks for each frame. (b) Then, we sequentially sparsify input frames, select prominent coarse correspondences, and visualize the constructed coarse correspondences on the images. (c) Finally, we enable MLLMs to better understand 3D spacetime from the prompted images.

We have demonstrated substantial performance gains of COARSE CORRESPONDENCES through extensive experiments with both open-source and closed-source models across 6 benchmarks on 3D spatial understanding and temporal understanding. For closed-source models, we apply COARSE CORRESPONDENCES on GPT4-V/O during inference and achieve compelling gains. First, on 3D understanding, we show that our method significantly surpasses state-of-the-art models by 20.5% and 9.7% on ScanQA (Azuma et al., 2022) and OpenEQA (Majumdar et al., 2024) respectively. Second, for long video understanding, our method leads to a 6% gain in performance on the EgoSchema benchmark (Mangalam et al., 2023). Notably, our method uses much fewer input images and, in a zero-shot manner, outperforms many fine-tuned models that use far more images. For example, on EgoSchema, COARSE CORRESPONDENCES surpasses state-of-the-art results with just 8 uniformly sampled frames from a 3-minute video, greatly reducing the computational costs of MLLMs compared to existing methods. In addition to 3D and video QA tasks, we further demonstrate that our method enhances models’ performance on embodied tasks such as navigation (Krantz et al., 2020), which require strong spatial and temporal understanding, by 11% in success rate on R2R. These results suggest that COARSE CORRESPONDENCES boosts MLLMs’ spatial and temporal understanding both effectively and efficiently. Last but not least, we experiment with open-source MLLMs (Liu et al., 2024a) by applying COARSE CORRESPONDENCES in both instruction tuning and inference; again, our method shows performance gains against the baseline (by 6.9% on ScanQA), and the improvement even generalizes to unseen datasets such as SQA3D (+3.1%). These results suggest that COARSE CORRESPONDENCES works well universally with any model – both closed-source and open-source – that can take in multiple images and understand visual markers.

To further understand why our simple method brings significant improvements on spatial and temporal understanding, we conducted additional investigations with a small diagnostic benchmark on spatial orientation, as orientation is a key component of spatial understanding and important to navigation tasks. We manually curated a benchmark called SOT to test how well MLLMs understand left-right relationships from different viewpoints on images taken by cameras with different motions. Our experiments reveal that (1) even GPT-4o struggles with understanding images taken by a camera moving right to left (instead of the more common motion left to right) and (2) that COARSE CORRESPONDENCES leads to improvements on this benchmark. These results indicate that COARSE CORRESPONDENCES mitigates the camera motion bias in MLLMs – their tendency to understand images better when the camera moves in a particular direction – when processing visual signals and helps them learn a more equivariant internal visual representation.

Overall, we want to highlight with this work that, despite its simplicity and being underestimated for semantic tasks in deep learning, visual correspondence can still bring significant utility to spatial and temporal understanding in MLLMs, just as it has long contributed to 3D reconstruction Schonberger & Frahm (2016). Although MLLMs still exhibit some non-negligible shortcomings in spatial

and temporal understanding, even with COARSE CORRESPONDENCES —such as the ability to perform spatial perspective-taking—we hope our work demonstrates the potential of leveraging visual correspondences to help MLLMs better understand our physical world.

2 METHOD

We introduce COARSE CORRESPONDENCES, a visual prompting method that allows MLLMs to reason about 3D space and time.

Problem formulation. Given a question \mathcal{Q} and a sequence or set of observations in an environment $[I_1, \dots, I_n]$, our aim is to design a visual prompt $\mathcal{P}(\dots)$ that modifies the input image set. These image inputs don’t have to be a video. They can also represent a set of images of a specific scene from multiple viewpoints. We evaluate the prompt by measuring its utility in prompting an MLLM \mathcal{M} :

$$[I'_1, \dots, I'_n] = \mathcal{P}([I_1, \dots, I_n])$$

$$\hat{\mathcal{A}} = \mathcal{M}([I'_1, \dots, I'_n], \mathcal{Q})$$

We compare the generated answer $\hat{\mathcal{A}}$ with the ground truth \mathcal{A} .

In our framework, the MLLMs can be any general-purpose model without requiring any special architecture or training for 3D or temporal understanding. Our aim is to develop a prompting strategy that allows models to improve such capabilities without any training (Figure 1).

COARSE CORRESPONDENCE

Our prompting method, COARSE CORRESPONDENCES, contains four steps: (1) tracking correspondences, (2) sparsify frames, (3) selecting, and (4) visualizing coarse correspondences.

(1) Tracking correspondences. Given n input images, $[I_1, \dots, I_n]$, we first use an off-the-shelf video object tracking model, such as Tracking Anything (Yang et al., 2023c). This model extracts class-agnostic instance segmentation masks (M_1, \dots, M_n) for each image. Each M_i is a $H \times W$ dimensional matrix where H and W are the height and width of the input image I_i . Each pixel location in M_i contains an instance ID, indicating which instance the pixel at that position belongs to within the image sequence.

(2) Sparsify frames. Since most MLLMs contain a large number of parameters, directly using them to process long image sequences is very computationally intensive. Additionally, proprietary MLLMs like GPT-4O can also incur significant costs if the number of image tokens that need to be processed increases. Reducing the number of input images might lose vital information necessary for MLLMs.

COARSE CORRESPONDENCES strikes a balance in this tradeoff by extracting meaningful video object tracks (a relatively cheaper operation) from high-frame-rate image sequences, and then samples a few image inputs along with the tracks, to retain—and even improve—performance while reducing the MLLM’s computation cost. From this extracted video object tracks, we perform temporal downsampling, retaining only $m \ll n$ uniformly sampled images and their corresponding masks, denoted as $[I_{s_1}, \dots, I_{s_m}]$ and $[M_{s_1}, \dots, M_{s_m}]$, where $s_i \in \{1, \dots, n\}$. This downsampling reduces the number of images we feed into \mathcal{M} .

(3) Selecting coarse correspondences. Prompting an MLLM with all the detected correspondences results in information overload. In fact, our ablations (discussed in Sec 5) find that adding all the correspondences reduces the MLLM’s performance. Therefore, we select a subset of prominent instances to retain. We select the prominent instances of the top-K objects that co-occur in the most number of frames. We first calculate the occurrence frequency and area sum of each unique instance ID in the retained m masks using the following equation:

$$Freq(\text{ID}) = \sum_{i=s_1}^{s_m} \mathbf{1}_{\{\text{ID} \in M_i\}},$$

$$Area(\text{ID}) = \sum_{i=s_1}^{s_m} \sum_{p \in M_i} \mathbf{1}_{\{\text{ID}=p\}}.$$

Model	Frame	BLEU-1	BLEU-2	METEOR	ROUGE-L	CIDEr
<i>3D-Specific Models</i>						
ScanQA (Azuma et al., 2022)	-	26.9	16.6	11.5	30	55.4
ScanRefer+MCAN (Yu et al., 2019)	-	30.2	20.4	13.1	33.3	64.9
3D-LLM (Hong et al., 2024)	-	39.3	25.2	14.5	35.7	69.4
<i>Open-source Multi-modal Models</i>						
LLaVA(Fine-tuned)	64	34.7	22.0	13.8	31.1	67.3
LLaVA+Coarse Correspondences	64	38.6	24.7	15.4	38.3	74.2
<i>Proprietary Multi-modal Models</i>						
GPT-4V	8	28.6	13.4	13.5	33.4	59.6
GPT-4V+Coarse Correspondences	8	39.7	25.5	17.4	40.8	79.2
GPT-4O	4	30.5	19.8	14.8	36.1	72.2
GPT-4O+Coarse Correspondences	4	35.4	25.5	18.0	42.6	87.0

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.

Then, we first sort all instance IDs in descending order based on $Freq(\text{ID})$. If there are ties, we further sort based on $Area(\text{ID})$. Finally, we retain the top k instance IDs as tracklets, denoted as $[T_1, \dots, T_k]$, to visualize for MLLMs.

(4) Visualizing coarse correspondences. For each set of obtained correspondence relationships, we visualize the correspondences directly in the image as a marker. Specifically, for each identified primary instance ID T_i , if it exists in the mask M_{s_j} of a retained image I_{s_j} , we overlay a mark with a fixed size and shape labeled with T_i at the position $(\bar{x}_{ij}, \bar{y}_{ij})$ on I_{s_j} to produce I'_{s_j} . The specific placement position can be easily obtained by:

$$(\bar{x}_{ij}, \bar{y}_{ij}) = \frac{\sum_{(x,y)} (x, y) \cdot \mathbf{1}_{\{M_{s_j}(x,y)=T_i\}}}{\sum_{(x,y)} \mathbf{1}_{\{M_{s_j}(x,y)=T_i\}}}$$

Naturally, we can overlay not just the markers but also the segmentation outlines or even the segmentation masks associated with each retained prominent instance. We explore these ablations later. In the end, we obtain the prompted image sequence $[I'_1, \dots, I'_m]$, which is then used as the input to MLLMs.

We refer to our method as *Coarse* because of the following: first, we only visually prompt for instance-level correspondences and not point-level correspondences. Second, the instance-level correspondences are extracted using off-the-shelf tracking models. Despite not being perfectly precise, they still help MLLMs build a better 3D model of the environment. Third, we only visualize a handful of prominent corresponding instances.

3 PROMPTING PROPRIETARY MODELS

We first evaluated the utility of our COARSE CORRESPONDENCES on multiple tasks using proprietary models, including understanding 3D space (ScanQA (Azuma et al., 2022) and OpenEQA (Majumdar et al., 2024) in §3.1) as well as temporal events (EgoSchema (Mangalam et al., 2023) in §3.2). Building on the improvements our method brings to 3D understanding and long video understanding, we further demonstrate that our method also delivers significant gains in navigation tasks (VLN-CE ()). Across all these benchmarks, we augment proprietary MLLMs (e.g., GPT-4V and GPT-4O) with COARSE CORRESPONDENCES and evaluate its zero-shot performance. We show that COARSE CORRESPONDENCES significantly improves the base GPT models and can substantially surpass many current state-of-the-art methods that have undergone specialized fine-tuning, even while using much fewer images as input. All experiments were conducted using A100 80G GPUs.

3.1 SPATIAL UNDERSTANDING

Benchmarks. The validation set of ScanQA dataset contains 4675 questions about 71 scenes. Questions in ScanQA require basic recognition, 3D localization, and 3D embodied capabilities (Duan et al., 2022). The validation set contains two ground-truth answers per question for evaluation with models that produce free-form answers. OpenEQA Dataset is an open-vocabulary dataset benchmarking spatial environment understanding and embodied reasoning. We evaluate on OpenEQA’s EM-EQA data split, which contains over 1600 high-quality human-generated questions. The subset tests the episodic memory of an agent moving through a 3D environment over time.

Baselines. For ScanQA, we evaluate COARSE CORRESPONDENCES by augmenting both GPT-4{V,O}, Gemini and Claude models. Besides, we also consider 3D specialized models (Yu et al., 2019; Azuma et al., 2022; Hong et al., 2024) fine-tuned on ScanQA. For OpenEQA, we compare against language-only models to account for language bias (LLaMA2 (Touvron et al., 2023)), commonly used general-purpose multimodal LLMs (GPT-4 (OpenAI et al., 2024), Claude3 (Anthropic, 2024), Gemini-Pro (Team et al., 2024), GPT-4V with 15 and 50 frames.

Metrics. For ScanQA, following prior works, we adopt BLEU (Papineni et al., 2002) scores, METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) as our evaluation metrics. For OpenEQA, we follow their evaluation approach by using GPT-4 to compare the generated answers with the ground-truth answers and assign a score. We report the average score across all questions.

Results. For ScanQA, as shown in Table 1, compared to raw input, COARSE CORRESPONDENCES consistently improves the overall performance of different proprietary models. For instance, on the strongest model, GPT-4o, COARSE CORRESPONDENCES brings improvements of 5.7 BLEU-2, 3.2 METEOR, 6.5 ROUGE-L, and 15 CIDEr points. Compared to methods that are specifically designed for 3D understanding tasks, fine-tuned with specialized 3D SFT data, or even those that use 3D point clouds instead of 2D images as input, we observe that a general-purpose MLLM can still outperform them, especially when enhanced with COARSE CORRESPONDENCES. Moreover, we found that this can be achieved using far fewer images as input.

We also demonstrated the same conclusion on OpenEQA, as indicated in Table 2. By applying COARSE CORRESPONDENCES, we significantly improved the performance of both GPT-4v and GPT-4o, achieving better results with fewer input images. These findings suggest that general-purpose MLLMs are indeed capable of understanding 3D space, and COARSE CORRESPONDENCES can significantly enhance their spatial understanding while simultaneously reducing the number of views needed, which could lower the inference cost and make MLLMs more useful for embodied tasks.

Models	Frame	Accuracy
LLaMA2 (Touvron et al., 2023)	0	28.3
GPT-4 (OpenAI et al., 2024)	0	33.5
Claude3 (Anthropic, 2024)	20	36.3
Gemini-Pro (Team et al., 2024)	15	44.9
GPT-4V (OpenAI, 2023)	15	54.6
GPT-4V (OpenAI, 2023)	50	55.3
Human	Full	86.8
GPT-4V	8	44.8
GPT-4V+CC	8	58.5
GPT-4O	4	49.4
GPT-4O+CC	4	59.1

Table 2: **Comparisons on EM-EQA setting of OpenEQA** . Our method further enhances the embodied ability of MLLMs and exceeds previous methods by a large margin.

Models	Frame	Subset
LongViviT (Papalampidi et al., 2023)	256	56.8
MC-ViT-L (Balažević et al., 2024)	128+	62.6
LLoVi (Zhang et al., 2024)	180	58.3
VideoAgent (Wang et al., 2024)	8,4	60.2
MVU (Ranasinghe et al., 2024)	16	60.3
VideoAgent (Fan et al., 2024)	-	62.8
LangRepo (Kahatapitiya et al., 2024)	-	66.2
GPT-4V	8	64.2
GPT-4V+CC	8	67.4
GPT-4O	8	67.2
GPT-4O+CC	8	73.2

Table 3: **Comparisons on EgoSchema validation set.** COARSE CORRESPONDENCES improves existing MLLMs and surpasses previous finetuned models in a zero-shot manner.

3.2 TEMPORAL UNDERSTANDING

Benchmarks. We evaluated the improvements of our method for long video understanding using the EgoSchema dataset. Each video in EgoSchema is 3 minutes long, with a corresponding question that includes five multiple-choice options. These questions are designed to ensure that answering them

requires viewing a sufficient number of frames from the video. Due to budget constraints, we limited our evaluation to 500 questions from the validation set.

Baselines. The baseline methods we compared against include newly designed and trained model architectures specifically for long video understanding, such as LongViviT (Papalampidi et al., 2023) and MC-ViT-L (Balažević et al., 2024). On the other hand, we also compared methods that rely on text-only foundation models (e.g., GPT-4), i.e., Socratic-based approaches (Zhang et al., 2024; Kahatapitiya et al., 2024), which first use an off-the-shelf image captioning model (Zhao et al., 2023) to convert video frames into captions, and then prompt GPT-4 to answer questions based on those captions. Additionally, we compared agent-based methods (Wang et al., 2024; Fan et al., 2024), which involve using GPT-4 alongside an image captioning model in an agent framework to iteratively perform a series of multi-step reasoning operations to understand long videos. In contrast to these approaches, our method is entirely based on an end-to-end general MLLM architecture, exploring how to further enhance its ability to understand long videos.

Results. COARSE CORRESPONDENCES demonstrates state-of-the-art performance, significantly outperforming existing approaches in a zero-shot manner (Table 3). Compared to the original GPT-4o model, our method improves its performance by 6%. Notably, our method uses far fewer frames than other approaches, yet achieves higher results compared to methods that use many more frames. It is also worth highlighting that even the original GPT-4o, when limited to just 8 frames, already serves as a very strong baseline. This indicates the potential of a sufficiently powerful general-purpose MLLM in long video understanding.

Methods	Success Rate \uparrow	Oracle Success Rate \uparrow	Success weighted by Path Length \uparrow	Trajectory Length \uparrow	Navigation Error \downarrow
GPT-4O	12.00	18.00	10.37	7.31	8.49
GPT-4O+CC	23.00	29.00	21.03	8.12	7.37

Table 4: **Comparison on Navigation task.** COARSE CORRESPONDENCES improves GPT-4o’s performance on R2R dataset for different evaluation metrics. Except for NE, where a lower value indicates better performance, higher values for the other metrics reflect better performance.

3.3 NAVIGATION

Building on the improvements in 3DQA and VideoQA, we hope that our method can also prove effective in embodied tasks such as navigation. Navigation requires an agent to understand 3D space, such as being able to determine the spatial relationship between objects in the instruction and itself, while also performing temporal reasoning to assess the progress toward completing the instruction. We consider conducting experiments on the VLN-CE benchmark (Krantz et al., 2020), which is a continuous simulation environment for low-level action execution in indoor scenes.

Setup. We adopt the val-unseen split from R2R (Krantz et al., 2020) for evaluation. Unlike the previous QA tasks, where all images could be processed at once, in navigation tasks, each image is processed in an online fashion. Specifically, we feed in one image at each iteration of the conversation. Given the significant variation in viewpoints during navigation, we use SAMv2 (Ravi et al., 2024), the state-of-the-art model for long-range tracking, to label each new input image based on episodic history. Then, using the prompted images, we induce the MLLM to output one of four actions at each step: FORWARD (distance), TURN-LEFT (rotation angle), TURN-RIGHT (rotation angle), and STOP. We follow NavGPT (Zhou et al., 2024) to craft input prompts. Considering the high computational cost of navigation tasks, we selected 100 samples from the val-unseen split. Our primary goal is to demonstrate that our method can enhance GPT models’ capabilities in zero-shot navigation tasks, which remains a significant challenge for various types of models.

Metrics. We follow the standard VLN evaluation metrics to evaluate the navigation performance, including success rate (SR), oracle success rate (OS), success weighted by path length (SPL), trajectory length (TL), and navigation error from goal (NE). Note that an episode is considered successful if the agent calls the STOP action within 3m of the goal in the VLN-CE.

Results. As shown in Table 4, our method achieved improvements across all metrics. It can be observed that while GPT-4o performs impressively on many QA tasks, its zero-shot performance

on navigation tasks is relatively low. This may partly be due to the lack of specialized training on action data, making it less accurate in outputs such as determining how many meters to move forward. However, our method reveals another dimension of the problem: MLLMs’ understanding of the 3D spacetime in which they operate can be further enhanced. This is evidenced by the significant improvements in navigation when using COARSE CORRESPONDENCES. We believe that our approach holds great potential for embodied tasks, which can be explored in future research.

4 PROMPTING OPEN MODELS

We further validate the effectiveness of our COARSE CORRESPONDENCES on open-source models. Our primary goal is to demonstrate that our method is not only effective for powerful proprietary models but also provides general improvements to a wide range of MLLMs. We start with the LLaVA model (Liu et al., 2024a) and fine-tune it using a dataset comprising approximately 1.2 million samples of image and video data. Notably, the ScanQA dataset is the *only* dataset related to 3D in this collection. COARSE CORRESPONDENCES is applied only to the ScanQA data, while the other data maintain their original format.

In-domain Evaluation. We first evaluate our model on the ScanQA validation set. As shown in Table 1, our method, compared to fine-tuning on the original ScanQA without COARSE CORRESPONDENCES, significantly enhances the model’s 3D spatial understanding, even surpassing previous VLMs specifically designed for 3D tasks, which involve specialized architectural designs and are fine-tuned on much larger amounts of 3D-related data. This demonstrates that COARSE CORRESPONDENCES can also work effectively for open MLLMs.

Out-domain Evaluation. To further demonstrate the generalizability of our method, we conduct experiments to evaluate the zero-shot performance of our model, fine-tuned on ScanQA, on the SQA3D dataset. As shown in Table 5, on this previously unseen dataset, COARSE CORRESPONDENCES still outperforms the model fine-tuned only on the original ScanQA, proving that our method can generalize to out-of-domain datasets. Even more notably, even without using COARSE CORRESPONDENCES during inference, simply applying it during the training phase already brings improvements. This highlights that our method is not only effective as a prompting technique for inference but also holds potential as a data augmentation method during training, which is worth further exploration in the future.

Method	Acc
LLaVA(Fine-tuned)	36.0
LLaVA+CC(train-only)	37.17
LLaVA+CC	39.13

Table 5: **Comparisons on SQA3D dataset.** COARSE CORRESPONDENCES generalizes well on out-domain dataset.

5 ANALYSIS

5.1 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 (Newcombe, 1989; Tversky & Hard, 2009) 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.

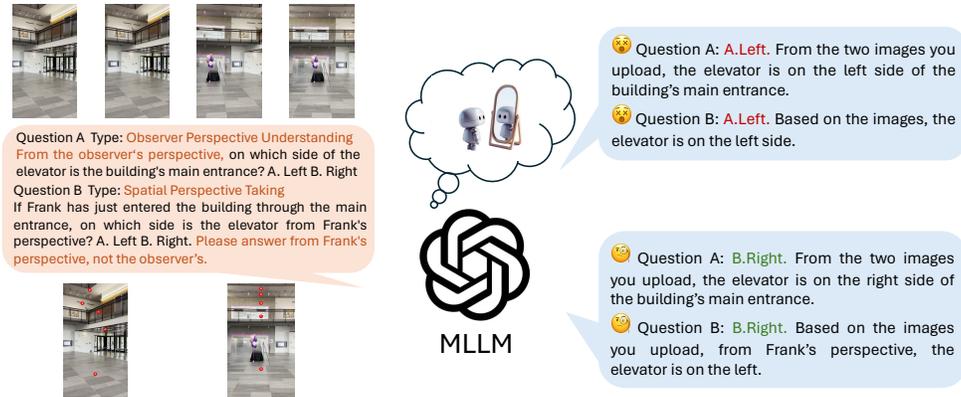


Figure 2: **Illustration of our SOT dataset.** We mention two types of questions: Observer perspective understanding and spatial-perspective taking. COARSE CORRESPONDENCES demonstrates superior effectiveness on the dataset.

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 2, 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. 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 6, 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 2, 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.

More importantly, according to Table 6, 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 results from both left-to-right ($L \rightarrow R$) and right-to-left ($R \rightarrow L$) 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 3. 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

432
433
434
435
436
437
438
439
440
441
442

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 6: **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.

awareness has yet to emerge in MLLMs—at least for now—highlighting a potential direction for future research.

443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464

Ablation on number of marks					
Design Choice	B - 1	B - 2	METEOR	ROUGE-L	CIDER
5	39.7	25.5	17.4	40.8	79.2
8	35.4	18.9	14.6	37.8	74.0
Ablation on mark size					
40px	35.3	17.5	15.5	39.2	76.1
60px	39.7	25.5	17.4	40.8	79.2
80px	33.1	14.3	14.5	37.6	71.3
Ablation on mark type					
markers only	35.9	19.6	15.9	39.5	76.4
+ outline	39.7	25.5	17.4	40.8	79.2
+ mask	33.1	14.3	14.5	37.6	71.3

Table 7: **Ablations on different design choices of COARSE CORRESPONDENCES.** We studied the impact of the number, size, and type of marks on performance. All experiments were conducted on ScanQA using GPT-4V.

465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480

5.2 ABLATION STUDY

Here, we explore the various design decisions in our method.

How does COARSE CORRESPONDENCES differ from other visual prompting methods? Our proposed method calculates and highlights correspondences between images, aiming to elicit 3D and temporal understanding. Other visual prompting methods (namely Set-of-Mark (Yang et al., 2023a), 3DAXiesPrompts (Liu et al., 2023), and Chain-of-thought (Wei et al., 2023)) can also be viewed as alternative prompting methods. Given that the ground-truth answers in existing benchmarks are relatively brief, we selected a scene from ScanQA and manually designed a new question. We qualitatively compare COARSE CORRESPONDENCES against other prompting methods on this new question, as shown in Figure 4.

481
482
483
484
485

The orange part of Figure 4 shows our Coarse Correspondence labels are recognized by GPT-4V. The output answer provides evidence that our coarse correspondence helps GPT-4V develop a mental 3D model of the scene. Set-of-Marks provides no spatial corresponding information and therefore is unhelpful. The Axis labels in 3DAXies can be easily misrecognized by GPT-4V, leading to misleading spatial information. Though Chain-of-Thought helps identify objects, it fails to resolve the “spatial perspective-taking” issue.

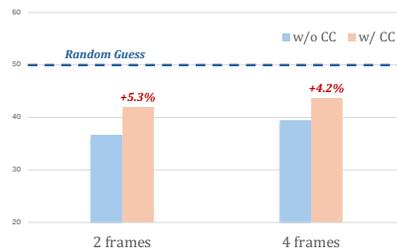


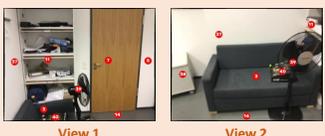
Figure 3: **Comparisons on SOT's spatial perspective-taking questions.** COARSE CORRESPONDENCES improves performance but GPT-4O still performs below random chance.

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

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.

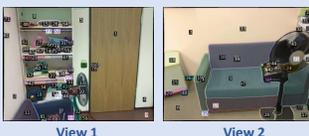
User: You are sitting on the sofa and the electric fan is on your left. Describe the location of the room door from your perspective: A. to the front left of you; B. to the front right of you; C. to the back left of you; D. to the back right of you.

+Coarse Correspondence (Ours)



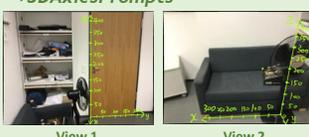
GPT-4V: The first image ... a door marked with number 7, a fan marked with number 39, and items on shelves marked with numbers 11 and 37. In the second image, ... the fan again at number 39, and a sofa marked with number 3(✓). We can infer that the door and the fan are on the same side of the room(✓). Since the fan appears to the left in the first image and is also on the left when sitting on the sofa as per the second image and your description, the door must be to the front of the sofa, as it is not visible in the second image from sofa's perspective(✓). By 'sitting on the sofa' as instructed, and with the electric fan to the left, I can determine the door is A. to the front left of you (✓).

+Set-of-Mark



GPT-4V: The fan is seen on the right side from the sofa's perspective, its leftward when sitting means the orientation is flipped(✓). Therefore, the door that is visible on the right in the first image would be D. behind and to the right when seated on the sofa(✗).

+3DAxesPrompts



GPT-4V: The first image shows ... and a coordinate system with X-axis (horizontally rightwards)(✗) and Z-axis (vertically upwards). The second image shows a sofa on the left with a coordinate system's Y-axis (horizontally extending away from the viewer)(✗) and Z-axis as before. When sitting on the sofa with the fan to your left, the door would be B. diagonally opposite to you on the right side (✗).

+Chain-of-Thought

System: Your response should be as follows.
1)Identification: Identify all labeled objects.
2)Analysis: Analyze the objects and their positions construct a 3D model.
3)Synthesis: Combine the info above, form a comprehensive understanding.
4)Response: Provide a concise answer to the question based on your comprehensive 3D modelling.

GPT-4V:
Identification: Door, shelves, sofa, ... (✓)
Analysis: in image 1 the door is on the right with the shelves adjacent to the door ... In image 2 the sofa is ... (✓).
Synthesis: When sitting on the sofa, the shelves are on the right(✗). The door is next to the shelves.
Response: B.(✗).

Figure 4: **Comparison of different prompting method.** Our proposed **Coarse Correspondences** successfully guided GPT-4V to understand 3D spatial relationships and generate the right answer. Other existing prompting method including image-based **Set-of-Marks**, **3DAxes** and text-based **Chain-of-Thought** failed to answer correctly.

Why use coarse instead of dense correspondences? Instead of filtering and retaining only a handful of coarse correspondences, one ablation we considered is the possibility of using all dense correspondence. Unfortunately, we find that excessively overlaying too many instance marks can degrade performance (Table 7) as they occlude the visual content in the images.

How large should the marks be? We inject the correspondences into MLLMs by overlaying the marks into images. We empirically find an optimal mark size (where 'px' represents the mark's diameter in pixels) in Table 7. Marks that are too small tend to be ignored while those that are too large occlude visual content.

What shape should the marks be? We further studied the appearance of the marks. In addition to red circles with white text, we experimented with adding segmentation outlines and segmentation masks. As shown in Table 7, using segmentation outlines enhances object grounding. However, using segmentation marks occludes visual content and reduces performance.

6 CONCLUSION.

We propose a framework called COARSE CORRESPONDENCES prompting. By using off-the-shelf video tracking models to obtain class-agnostic, instance-level correspondences, and conveying this information to MLLMs through visual prompting, we discovered that this simple method, using only 2D images as input—without any specialized architectural design or task-specific SFT—can effectively enhance MLLMs' understanding of 3D space and long videos. This improvement extends to embodied tasks like navigation. Our method not only works on proprietary models but also generalizes to open-source models, and it performs well on both in-domain and out-of-domain datasets. Moreover, it enhances not just inference but also training. Further analysis shows that our method helps MLLMs become more robust to camera motion bias. Additionally, we identified that even GPT models struggle with perspective-taking capability, a fundamental component of human visual intelligence, which presents an important avenue for future exploration to further improve MLLMs.

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

REFERENCES

- AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.
- 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*, pp. 19129–19139, 2022.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.
- Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J. Hénaff. Memory consolidation enables long-context video understanding, 2024.
- 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*, pp. 65–72, 2005.
- 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*, pp. 404–417. Springer, 2006.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024.
- Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.
- 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*, pp. 11287–11297, 2021.
- 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.
- Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019.
- Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. Language repository for long video understanding, 2024.

594 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
595 Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick.
596 Segment anything. *arXiv:2304.02643*, 2023.
597

598 Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph:
599 Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020:
600 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp.
601 104–120. Springer, 2020.

602 Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning
603 events in videos, 2017.
604

605 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-
606 training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*,
607 2023.

608 Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual
609 representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.
610

611 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization
612 branches out*, pp. 74–81, 2004.

613 Dingning Liu, Xiaomeng Dong, Renrui Zhang, Xu Luo, Peng Gao, Xiaoshui Huang, Yongshun Gong,
614 and Zhihui Wang. 3daxiesprompts: Unleashing the 3d spatial task capabilities of gpt-4v. *arXiv
615 preprint arXiv:2312.09738*, 2023.
616

617 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36,
618 2024a.

619 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in
620 neural information processing systems*, 36, 2024b.
621

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

624 Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff,
625 Sneha Silwal, Paul Mccvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li,
626 Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv
627 Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran.
628 Openeqa: Embodied question answering in the era of foundation models. In *Conference on
629 Computer Vision and Pattern Recognition (CVPR)*, 2024.
630

631 Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic
632 benchmark for very long-form video language understanding, 2023.

633 Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie,
634 Danny Driess, Ayzaan Wahid, Zhuo Xu, et al. Pivot: Iterative visual prompting elicits actionable
635 knowledge for vlms. *arXiv preprint arXiv:2402.07872*, 2024.
636

637 Nora Newcombe. The development of spatial perspective taking. *Advances in child development and
638 behavior*, 22:203–247, 1989.

639 OpenAI. Gpt-4v(ision) system card. *OpenAI Blog*, 2023. URL [https://cdn.openai.com/
640 papers/GPTV_System_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf).
641

642 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed:
643 2024-05-22.

644 OpenAI, Josh Achiam, and et al. Gpt-4 technical report, 2024.
645

646 Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean,
647 Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for
contrastively pre-training video-first encoders beyond 16 frames, 2023.

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

651 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
652 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
653 models from natural language supervision. In *International conference on machine learning*, pp.
654 8748–8763. PMLR, 2021.

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

657 Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S. Ryoo. Understanding long
658 videos in one multimodal language model pass, 2024.

659 Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham
660 Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images
661 and videos. *arXiv preprint arXiv:2408.00714*, 2024.

662 Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of*
663 *the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

664 Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red
665 circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023.

666 Simranjit Singh, Georgios Pavlakos, and Dimitrios Stamoulis. Evaluating zero-shot gpt-4v perfor-
667 mance on 3d visual question answering benchmarks. *arXiv preprint arXiv:2405.18831*, 2024.

668 Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent
669 correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:
670 1363–1389, 2023.

671 Gemini Team, Rohan Anil, and et al. Gemini: A family of highly capable multimodal models, 2024.

672 Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia,
673 Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large
674 vision-language models, 2024.

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

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

680 Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image
681 description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern*
682 *recognition*, pp. 4566–4575, 2015.

683 Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video
684 understanding with large language model as agent, 2024.

685 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le,
686 and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

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

689 Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark
690 prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*,
691 2023a.

692 Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu
693 Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Octopus: Embodied vision-language
694 programmer from environmental feedback, 2023b.

702 Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything:
703 Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023c.
704

705 Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for
706 visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and*
707 *pattern recognition*, pp. 6281–6290, 2019.

708 Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas
709 Bertasius. A simple llm framework for long-range video question-answering, 2024.
710

711 Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from
712 large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
713 *Pattern Recognition*, pp. 6586–6597, 2023.

714 Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language
715 navigation with large language models. In *Proceedings of the AAAI Conference on Artificial*
716 *Intelligence*, volume 38, pp. 7641–7649, 2024.
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

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 necessary 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. Over-dependence on these models might erode critical thinking and analytical abilities in the long term.

B RELATED WORK

Multimodal language models Multimodal LLMs(Liu et al., 2024b; Bai et al., 2023) integrate vision encoders (Radford et al., 2021) into large LLMs (Chiang et al., 2023; Touvron et al., 2023), allowing them to directly reason over visual input. Many proprietary models, such as GPT-4 (OpenAI, 2023), Gemini (Team et al., 2024), and Claude (Anthropic, 2024), as well as open-source models like the LLaVA series (Liu et al., 2024b) and BLIP series (Li et al., 2023), have made significant progress in 2D vision-language tasks like image captioning (Chen et al., 2015) and visual question answering (VQA) (Hudson & Manning, 2019; Goyal et al., 2017). Beyond these language-related tasks, many newer attempts applying MLLMs to applications such as autonomous driving (Tian et al., 2024) and robotics (Yang et al., 2023b). Many of these tasks require understanding the 3D space in which they are deployed and reason about how things are changing temporally. We improve the 3D space-time capabilities of such models.

Visual prompting. Effective prompting has been widely proven to improve LLMs across multiple domains. Methods, such as chain-of-thought prompting (Wei et al., 2023), force the model to reason before answering a question. For multimodal LLMs, methods such as Red-circle prompting (Shtedritski et al., 2023) and Set-of-marks (Yang et al., 2023a) can enhance the grounding abilities of CLIP (Radford et al., 2021) and GPT-4V. PIVOT (Nasiriany et al., 2024) employs visual prompting combined with iterative VQA to induce GPT-4V to generate outputs for robotics control. 3DAXIES (Liu et al., 2023) enhances GPT-4V’s ability to use numerical expressions to describe 3D relationships of objects in a single image by annotating a scaled 3D coordinate system on the image. Unlike these works, COARSE CORRESPONDENCES prompts MLLMs to understand the spatial relationships within a complete 3D scene from an image sequence.

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 (Lin et al., 2023) has improved performance on video dense captioning (Krishna et al., 2017) and videoQA (Xiao et al., 2021; Grunde-McLaughlin et al., 2021). To further advance the understanding of temporal relationships in videos, EgoSchema (Mangalam et al., 2023) 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 (Hong et al., 2024) converts multiview images into 3D point clouds and then feeds them into LLMs, demonstrating better results on the ScanQA (Azuma et al., 2022) benchmark for 3D understanding. OpenEQA (Majumdar et al., 2024) 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 (Schonberger & Frahm, 2016) utilize correspondences to better reconstruct 3D scenes. In the past, we relied on handcrafted features like SIFT (Lowe, 2004) or SURF (Bay et al., 2006) to obtain good correspondence. Today, features extracted from deep models (Tang et al., 2023) 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 (Yang et al., 2023c).

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 component, which makes the centering procedure worth explaining.

```

827 # Find center of a mask,
828 # May contains multiple connected components.
829 def find_center(mask):
830     # Go through the middle column, try to find center1
831     exist_y = []
832     x_center = median(left_bound, right_bound)
833     for y in range(upper_bound, lower_bound):
834         if (x_center, y) in mask:
835             exist_y.append(y)
836     if exist_y is not empty:
837         y_center = median(exist_y)
838         center1 = (x_center, y_center)
839     else:
840         center1 = None
841
842     # Go through the middle row, try to find center2 (skip)
843     if avg(center1, center2) in mask:
844         return avg(center1, center2)
845     elif center1 in mask:
846         return center1
847     elif center2 in mask:
848         return center2
849     else:
850         center_naive = ((left_bound + right_bound)//2,
851                        (upper_bound + lower_bound)//2)
852         return center_naive

```

Figure 5: 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

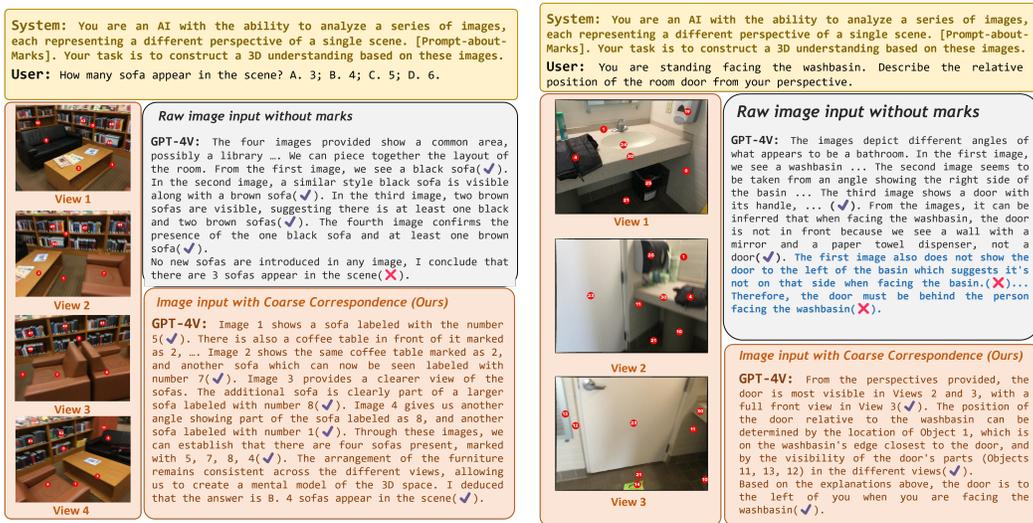
As shown in the pseudo code in Figure 5, firstly we calculate the medium x-index of the masked pixels and loop through this column, trying to find the first center point. Similarly, we calculate the medium y-index of the masked pixels and loop through this row, trying to find another center point. Normally 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 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.

The results of these case studies are shown in Fig. 6. 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

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917



(a) **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.

(b) **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.

Figure 6: Two complicated tasks, i.e. Duplicate Objects Counting and Relative Location Modeling are chosen to demonstrate our method. Zoom in for better view.

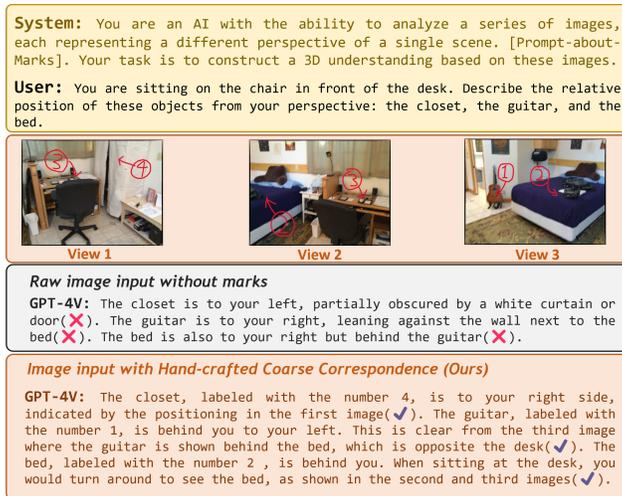


Figure 7: Hand-crafted coarse correspondence label

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 7. This further proves that our proposed method are style-agnostic as long as the marks is able to deliver the spatial correspondence knowledge.

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

E 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 (Kirillov et al., 2023), 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 (Feichtenhofer et al., 2019) 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 (Rajasegaran et al., 2023). Here, we demonstrate that it can more generally help MLLMs understand the 4D spatiotemporal context represented in videos.