
COARSE CORRESPONDENCES BOOST 3D SPACETIME UNDERSTANDING IN MULTIMODAL LANGUAGE MODEL - SUPPLEMENTARY MATERIAL

Anonymous authors

Paper under double-blind review

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., 2024; 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., 2024) 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.

```

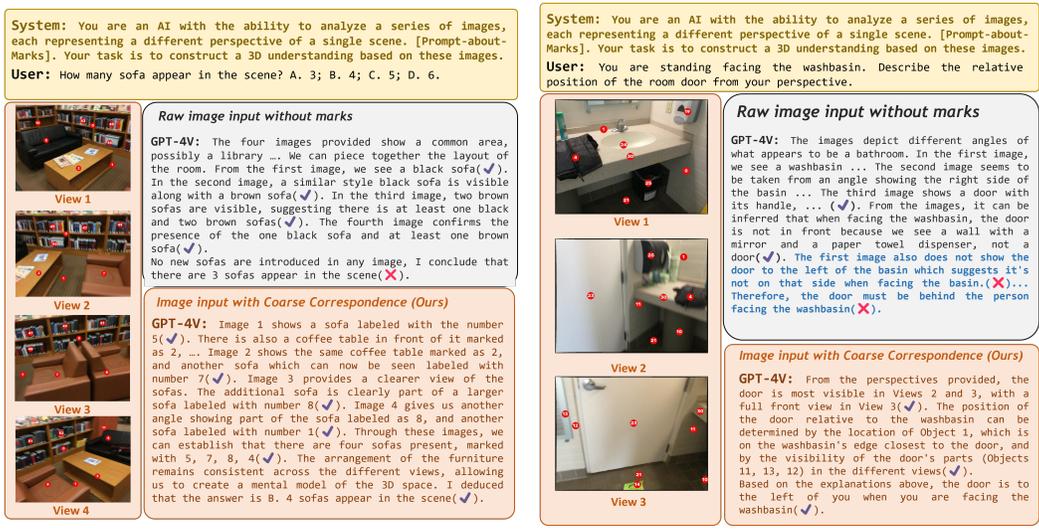
081 # Find center of a mask,
082 # May contains multiple connected components.
083 def find_center(mask):
084     # Go through the middle column, try to find center1
085     exist_y = []
086     x_center = median(left_bound, right_bound)
087     for y in range(upper_bound, lower_bound):
088         if (x_center, y) in mask:
089             exist_y.append(y)
090     if exist_y is not empty:
091         y_center = median(exist_y)
092         center1 = (x_center, y_center)
093     else:
094         center1 = None
095
096     # Go through the middle row, try to find center2 (skip)
097     if avg(center1, center2) in mask:
098         return avg(center1, center2)
099     elif center1 in mask:
100         return center1
101     elif center2 in mask:
102         return center2
103     else:
104         center_naive = ((left_bound + right_bound)//2,
105                        (upper_bound + lower_bound)//2)
106         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),

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161



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

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

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

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

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.



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 3: Hand-crafted coarse correspondence label

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.

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

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.
- 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.
- 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/>.
- 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.
- 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.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos, 2017.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- 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.
- Dingning Liu, Xiaomeng Dong, Renrui Zhang, Xu Luo, Peng Gao, Xiaoshui Huang, Yongshun Gong, and Zhihui Wang. 3daxiesprompts: Unleashing the 3d spatial task capabilities of gpt-4v. *arXiv preprint arXiv:2312.09738*, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

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

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

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

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

285 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
286 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
287 models from natural language supervision. In *International conference on machine learning*, pp.
288 8748–8763. PMLR, 2021.

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

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

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

296 Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent
297 correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:
298 1363–1389, 2023.

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

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

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

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

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

312 Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark
313 prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*,
314 2023a.

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

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

320
321
322
323