UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces

Baining Zhao[1,2*], **Jianjie Fang**[3*], **Zichao Dai**[4*], **Ziyou Wang**[3], **Jirong Zha**[1], **Weichen Zhang**[1,2], **Chen Gao**[4†], **Yue Wang**[4], **Jinqiang Cui**[2], **Xinlei Chen**[1†], **Yong Li**[4]
[1]Shenzhen International Graduate School, Tsinghua University,
[2]Pengcheng Lab, Shenzhen, [3]Northeastern University at Qinghuangdao,
[4]Department of Electronic Engineering, Tsinghua University
chen.xinlei@sz.tsinghua.edu.cn, chgao96@gmail.com, liyong07@tsinghua.edu.cn

Project Page　*Dataset*　*Code*

Abstract

Large multimodal models exhibit remarkable intelligence, yet their embodied cognitive abilities during motion in open-ended urban aerial spaces remain to be explored. We introduce a benchmark to evaluate whether video-large language models (Video-LLMs) can naturally process continuous first-person visual observations like humans, enabling recall, perception, reasoning, and navigation. We have manually control drones to collect 3D embodied motion video data from real-world cities and simulated environments, resulting in 1.5k video clips. Then we design a pipeline to generate 5.2k multiple-choice questions. Evaluations of 17 widely-used Video-LLMs reveal current limitations in urban embodied cognition. Correlation analysis provides insight into the relationships between different tasks, showing that causal reasoning has a strong correlation with recall, perception, and navigation, while the abilities for counterfactual and associative reasoning exhibit lower correlation with other tasks. We also validate the potential for Sim-to-Real transfer in urban embodiment through fine-tuning.

1 Introduction

Humans can process continuous first-person visual observations, enabling them to discern direction, judge distance, and navigate in the three-dimensional space of the real world (Richardson et al., 2010; Burigat et al., 2017; Grauman et al., 2022; Song et al., 2023). This refers to embodied cognition in motion, which highlights that cognitive processes are deeply rooted in the body's interactions with the world (Shapiro, 2019; Newcombe et al., 2023). Naturally, endowing agents with this embodied cognition capability has been a long-term goal in the field of embodied intelligence (Fan et al., 2022; Singh et al., 2023; Zha et al., 2025).

In recent years, large multimodal models (LMMs) (Li, 2023; Wang et al., 2024d; Xu et al., 2025; Zhan et al., 2025) have emerged as a promising approach to achieve this goal. Typically, video-large language models (Video-LLMs) are evaluated on capabilities such as video summarization (Samel et al., 2024; Hua et al., 2024), event question answering (Wang et al., 2024a), and goal localization (Yu et al., 2023). However, the benchmarks used to assess these capabilities are often limited to disembodied third-person video clips, where the agent itself is static (Wu et al., 2024b; Song et al., 2024; Fang et al., 2024). Besides, existing embodied video understanding research (Suglia et al., 2024; Cheng et al., 2024) mainly focuses on robotic arm manipulation (Nair et al., 2022) or indoor/ground-level movement (Marcu et al., 2024; Yang et al., 2024; Zhang et al., 2025). However, the embodied cognitive abilities required for three-dimensional motion in urban open-ended spaces have not been well-defined or assessed. The first-person perspective visual continuous observations generated in this scenario possess the following characteristics:

- Complex Scene and Rich Semantic Information: Urban areas are vast, containing diverse elements like skyscrapers, bridges, and tunnels that provide rich semantic information and pose comprehension and navigation challenges, while dynamic elements like pedestrians and vehicles require real-time adaptation (Yao et al., 2024; Xu et al., 2023).
- Unique Aerial Motion: Aerial navigation (Chen et al., 2024a) involves vertical mobility and a first-person perspective, adding complexity by requiring enhanced embodied cognition for processing diverse motion and observation angles, necessitating advanced spatial awareness and decision-making (Gao et al., 2024; Lee et al., 2024; Xu et al., 2024).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32400–32423

Figure 1: Example of video-language multiple choice question-answering. This figure presents three video examples corresponding to three data sources: real cities in Guangdong Province, China; the simulator EmbodiedCity constructed based on the real city of Beijing, China; and the simulator AerialVLN built on virtual cities. To ensure logical consistency in the movement trajectories within the videos, all video clips consist of continuous perceptual observations generated from ongoing or completed vision-language navigation tasks. For each video clip, we designed different task types to evaluate the embodied cognitive intelligence of Video-LLMs.

From these characteristics, we infer that embodied cognition in urban open spaces poses new challenges, and assessing LMMs' embodied cognitive abilities offers insights for future urban applications (Wang et al., 2024c).

We can use drones to capture motion video in urban spaces as they navigate buildings and dynamic elements. Establishing a benchmark presents challenges: 1) Creating a task set to evaluate embodied capabilities in urban spaces. 2) Obtaining video data: Unlike most high-altitude aerial views (Li et al., 2016; Zhu et al., 2021; Wen et al., 2021), our goal is to record drones maneuvering among urban structures with flexible movement and camera angles. Drones face issues like signal loss, limited range, and crashes due to obstructions and interference, making data collection difficult and costly. 3) Designing logical and purposive motion routes to ensure coherent visual observations.

Accordingly, we introduce a benchmark, **UrbanVideo-Bench**, designed for embodied motion cognition from embodied videos in urban airspace. Firstly, we propose a novel task set comprising 16 tasks characterized by urban spatiotemporal features, as shown in Figure 1 and Table 1. Secondly, we manually operate drones to collect embodied video data from 1) the real cities in Guangdong Province, China, 2) a simulator EmbodiedCity (Gao et al., 2024) built on the real city Beijing, China, and 3) a simulator AerialVLN (Liu et al., 2023) built on virtual cities. Using both real devices and simulators helps to rapidly increase the number of videos (Chen et al., 2017) . The movements in the videos are intentional, directed towards navigating to a specific position within urban space or following a particular route (Wu et al., 2024c). Then, we developed a question-answer generation pipeline with trained human annotators (over 800 hours of effort) and the expertise of LMMs, generating high-quality multiple-choice questions (MCQs). Finally, we quantitatively and qualitatively evaluate widely-used Video-LLMs in zero-shot settings, including both proprietary and open-source models. We additionally attempt supervised fine-tuning (SFT) on two Video-LLMs to validate the effectiveness of our dataset.

Table 1: Task set overview. Embodied cognition in motion is divided into four abilities, each of which is manifested in several tasks. Each task is provided with corresponding handcrafted question prototypes.

Recall	
Trajectory Captioning	Summarize the agent's movement path using visual cues / landmarks.
Sequence Recall	What is the agent's next step after [changing direction to the left over the intersection]?
Object Recall	what is [located next to] the [Central ALL-STAR cafe]?
Scene Recall	Describe scene the agent observes [during the descent to a lower height near the destination].
Start/End Position	Where are the starting point and final destination of the agent's movement?
Perception	
Proximity	How does the distance between the agent and the [rooftop with solar panels in the residential area] change after the [agent descends to street level]?
Duration	Which takes longer, [the agent's movement through the skyscraper alley] or [its descent to the balcony on the 9th floor]?
Landmark Position	Given [navigation goal/route] initially, what is agent's current position relative to [landmark]?
Goal Detection	Given [navigation goal] at starting location, is the target currently visible in the field of view, and if so, what is its position within the view?
Cognitive Map	Summarize historical movement observations into a cognitive map.
Reasoning	
Causal	Why did the agent [perform a descent after ascending alongside the cylindrical building]?
Counterfactual	Instead of [flying over the elevated highway intersection], if the agent chooses to [fly around the cylindrical building], can it complete the task, and how is the alternative route?
Association	Given [navigation goal] at starting location, are there any relevant urban elements or objects in sight when the navigation goal is not visible?
Navigation	
Progress Evaluation	(Route-oriented vision-language navigation) Given [navigation route] at starting location, analyze which step the navigation is currently perform.
High-level Planning	(Goal-oriented vision-language navigation) Given [navigation goal] at starting location, make next plan from the current location.
Action Generation	Given [navigation goal/route] initially, generate the next control action from the current location.

Overall, the innovation of this research is the establishment of **the first benchmark for embodied cognition specifically tailored for motion in urban open-ended aerial spaces**:

- We propose a novel task set comprising 4 categories and 16 tasks to evaluate how Video-LLMs recall, perception, reasoning, and navigation from embodied videos.
- We consequently develop 5.2k multiple-choice questions and 1.5k video clips, derived from real world and simulated environments. The dataset generation pipeline can be extended to other embodied movement videos.
- 17 popular LMMs are evaluated and their shortcomings are analyzed. We also explored the correlation between embodied cognitive abilities and the potential of Sim-to-Real.

2 Related Work

Embodied Capabilities of LMMs. Embodied intelligence refers to the concept that cognitive processes are deeply rooted in the body's interactions with the world (Gupta et al., 2021; Shi et al., 2024; Ding et al., 2024; Zeng et al., 2024). Large Multimodal Models (LMMs) have demonstrated unprecedented visual understanding capabilities and are considered the "brains" for developing embodied agents (Tang et al., 2023; Liang et al., 2024; Huang et al., 2023). Unlike past work (Du et al., 2024; Ramakrishnan et al., 2024) that primarily focused on 2D images, static point cloud or language-based spatial understanding, human comprehension of the world is grounded in continuous visual perception (Zhang et al., 2024; Liao et al., 2024; Majumdar et al., 2023), akin to embodied cognition through video streams. Thus, we need relative video benchmarks from diverse sources to comprehensively evaluate the potential of LMMs in various embodied scenarios.

Video Benchmarks for LMMs. Traditional video benchmarks cover various tasks like abstract understanding and spatiotemporal analysis (Xu et al., 2017; Wu et al., 2024a; Fu et al., 2024; Gu et al., 2025). They mainly concentrate on understanding the video content (Wu et al., 2024b; Song et al., 2024; Fang et al., 2024), lacking exploration of Video-LLMs' embodied cognitive abilities from an embodied, egocentric perspective. While some research has focused on embodied capabilities in indoor or ground-level scenes (Chandrasegaran et al., 2024; Mangalam et al., 2023; Sima et al., 2024; Marcu et al., 2024; Yang et al., 2024; Chen et al., 2023; Dang et al., 2025; Majumdar et al., 2024), there has been insufficient exploration of embodied

Table 2: The proposed and popular benchmarks for video-large language models. Our benchmark's video sources and scenarios are different from others, focusing on evaluating the embodied cognitive abilities of Video-LLMs related to urban 3D aerial motion.

Benchmark	Video Source	Video Theme	Embodied	Environment	Motion	Video Num.	QA Num.
LongVideoBench (Wu et al., 2024b)	Web-Collected	Life, Movie, News	×	/	/	3.8k	6.7k
MovieChat-1K (Song et al., 2024)	Web-Collected	Movie, TV series	×	/	/	1.0k	14.0k
MMBench-Video (Fang et al., 2024)	YouTube	Life	×	/	/	600	2.0k
HourVideo (Chandrasegaran et al., 2024)	Public Dataset	Human Activity	✓	Indoor/Outdoor	Ground-Level	500	13.0k
EgoSchema (Mangalam et al., 2023)	Public Dataset	Human Activity	✓	Indoor/Outdoor	Ground-Level	5.0k	5.0k
Lingoqa (Marcu et al., 2024)	Self-Recorded	Driving	✓	Outdoor	Ground-Level	28.0k	419.0k
VSI-Bench (Yang et al., 2024)	3 Public Datasets	Indoor Motion	✓	Indoor	Ground-Level	288	5.0k
EgoPlan-Bench (Chen et al., 2023)	2 Public Datasets	Human Activity	✓	Indoor/Outdoor	Ground-Level	419	5.0k
ECBench (Dang et al., 2025)	Public + Self-recorded	Indoor Motion	✓	Indoor	Ground-Level	386	4324
OpenEQA (Majumdar et al., 2024)	Self-recorded	Indoor Motion	✓	Indoor	Ground-Level	180	1.6k
Ours	Self-Recorded	Aerial Agent Motion	✓	City	3D Aerial Space	1.5k	5.2k

abilities in urban open 3D spaces. Part of above-mentioned video benchmarks are shown in Table 2. Comparatively, we independently record embodied motion video data along with corresponding MCQs to evaluate models' cognitive abilities in complex, dynamic urban 3D spaces.

3 Benchmark Design and Construction

We firstly define 16 embodied tasks that assess the embodied cognitive capabilities of Video-LLMs from different four aspects. Then, we describe the dataset generation process. Finally, we provide the statistical characteristics of the dataset.

3.1 Task Set

Considering the characteristics of urban open-ended scenarios, embodied cognition can be divided into four abilities: recall, perception, reasoning, and navigation (Chen and Dolan, 2011; Tang et al., 2021; Chandrasegaran et al., 2024; Chen et al., 2024b). Each ability is evaluated through several specific tasks, which are outlined in Table 1.

Recall tasks evaluates Video-LLM ability to cognitively remember key aspects of urban environments. By integrating the city's semantic elements seen in the video, the task includes Trajectory Captioning, where agents summarize their paths using landmarks. Sequence Recall asks agents to sort out the sequence of actions, while Object Recall focuses on identifying objects near landmarks. Scene Recall involves describing details observed during specific actions, ensuring comprehensive memory and understanding in dynamic urban contexts. The "Start/End Position" indicates whether agents are aware of "where I come from" and "where I am."

Perception include static relative spatial relationships (Landmark Position, Goal Detection), dynamic position changes (Proximity), temporal understanding (Duration), and scene-level comprehension (Cognitive Map) (Chen et al., 2022). The design of these tasks encompasses a range from static to dynamic, spatial to temporal, and micro to macro levels.

Reasoning focuses on analyzing and making sense of its actions and surroundings within the urban environment. This includes understanding the causal relationships behind movements, such as why the agent descends after ascending alongside a building. It also involves counterfactual reasoning, where agents consider alternative routes, like flying around a building instead of over a highway, and assess the viability of these options. Additionally, Video-LLMs are required to recognize associations between visible urban elements and their navigation goals, even when the goals themselves are not in direct view, showcasing urban commonsense and the ability to think critically and adaptively.

Navigation tasks aims to evaluate whether Video-LLMs can plan routes and directly output actions in urban spaces (Wu et al., 2024c; Pu et al.). We assessed two types of vision-language navigation (VLN): route-oriented VLN (Zhou et al., 2024), where the agent is provided with a navigation route (e.g., fly forward to the white building, then turn right, and continue to the lakeside to stop) and can assess the progress of the current route (Progress Evaluation) while ultimately outputting control actions (Action Generation); and goal-oriented VLN (Chaplot et al., 2020), where only a navigation goal (e.g., the lakeside) is given, allowing the agent to autonomously perform high-level navigation planning (High-level Planning) and eventually map it into control actions (Action Generation). Three-dimensional aerial navigation in urban environments is one of the most challenging tasks in embodied intelligence.

The proposed tasks reflect real-world challenges that embodied video systems in urban open-ended spaces may encounter, enhancing the practical relevance of the evaluation.

3.2 Dataset Generation Pipeline

As presented in Figure 2a, we will outline the pipeline developed for creating the dataset, which primarily consists of four steps: video curation, multiple-choice question-answering (MCQ) generation, blind filtering, and human refinement.

3.2.1 Video Curation

The primary consideration for UrbanVideo-Bench is obtaining massive high-quality embodied mobility video data. We collecte drone flight video data in two cities, Shenzhen and Zhaoqing, in Guangdong Province, China. The data collection was conducted using two DJI Mini 4K drones. To further expand the dataset, we chose to acquire data from simulators with two city benchmarks: EmbodiedCity (Gao et al., 2024) and AerialVLN (Liu et al., 2023). They both have the following advantages: a) Realistic environment modeling; b) Support for aerial agents; c) Existing aerial route reference.

We divided our team into two groups. The first group designs the navigation goal/route and provides text instructions for the endpoint, ensuring that the final destination can be reached based on the instructions and common urban knowledge. The second group, consisting of experienced pilots (with over 1000 hours of real drone flight time), performs flights to perform the VLN tasks, collecting first-person perspective data during the flight. Through this approach, the movement of the agent is purposeful, and the collected video data is more conducive to evaluating the capabilities of Video-LLMs. See Appendix B.1 for details.

3.2.2 MCQ Generation

The objective of this phase is to leverage the foundational video understanding and text generation capabilities of a powerful LMM to quickly produce high-quality MCQs for each task. We have designed a Chain-of-thought (CoT) prompting based on characteristics of embodied movement videos, consisting of the following four steps: a) Narration Generation; b) Structured Extraction of Key Information; c) Role-playing; d) Providing Question/Option Templates and Examples. The detail process and prompts are in Appendix B.2. In order to improve the quality of problem generation, we use Gemini-1.5 Flash (Team et al., 2024) with video understanding capability to assist in generation at this stage.

3.2.3 Blind Filtering

The purpose of blind filtering is to eliminate MCQs that can be answered correctly based solely on urban common sense, without any video input. This process aims to evaluate the embodied cognition capabilities of Video-LLMs based on mobile video, rather than their world knowledge or common sense. Blind filtering enhances the quality of the dataset. The specific method involves using n Video-LLMs to guess the answer to any given MCQ case without video input. If all n Video-LLMs guess correctly, the MCQ is removed; otherwise, it is retained.

3.2.4 Human Refinement

The generated MCQs may contain invalid questions, ambiguous options, incorrect answers, and various other issues that require further human refinement. These issues stem from two main sources: a) Even the most advanced Video-LLMs lacks the ability to fully understand embodied movement videos. b) The urban aerial agent scenarios are complex, making video comprehension challenging. We approach the human refinement process from four aspects: a) Invalid/ambiguous questions: For example, when the task specifies for the agent to "navigate to the main entrance," the presence of multiple nearby buildings leads to ambiguity, as the agent is unsure which building's main entrance is intended. The navigation target should be clarified to "navigate to the main entrance of the yellow building on the right." b) Urban element hallucination: This refers to the presence of city elements or objects in the correct option that were never actually present in the video. c) Direction: Directional descriptions are often incorrect or imprecise. d) Choices with insufficient differentiation or errors: One scenario is more than one option is correct. Another is only one is correct, but the ground truth was provided incorrectly. The entire refinement process required over 800 person-hours. See Appendix B.3 for more examples.

3.3 Dataset Statistics

The proposed dataset includes 1,547 video clips with resolutions of 1280x720 pixels for real drone frames, 960x720 pixels for EmbodiedCity frames, and 520x520 pixels for AerialVLN frames. As shown in Figure 2b&c, the durations of these video clips range from 10 seconds to 10 minutes, and the trajectories of the UAV in the videos encompasses different directions and movement patterns in 3-

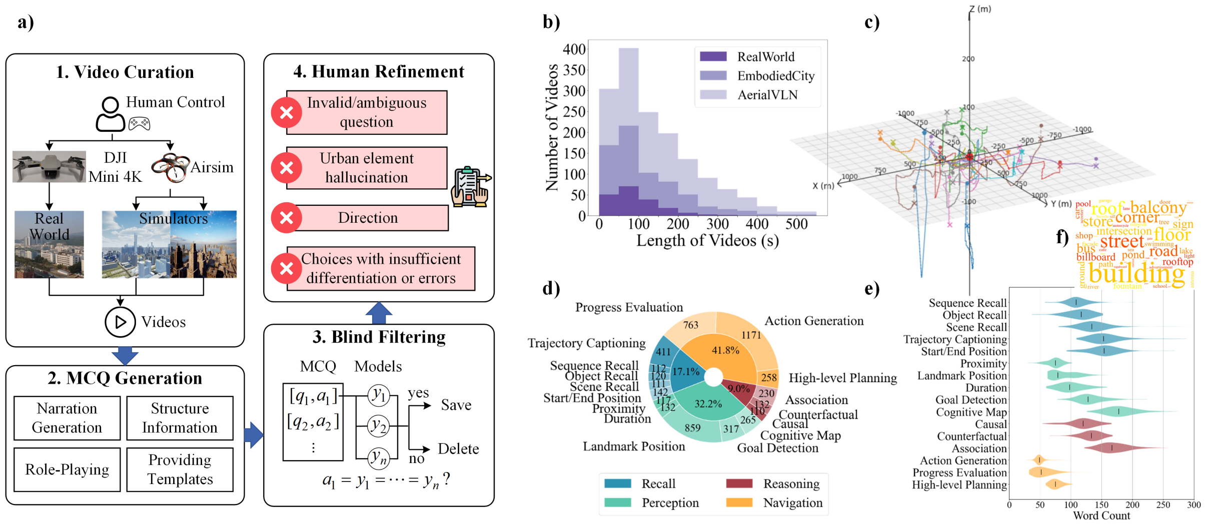

Figure 2: **Dataset Generation Pipeline and Statistics.** a) The pipeline includes four steps: video curation, MCQ generation, blind filtering, and human refinement. b) Histogram of frame count of videos. c) Histogram of path lengths. d) Histogram of word count of questions. e) Violin plot of word count of different categories of questions. f) Word cloud generated from questions and choices.

dimensional spaces, thus suggesting our dataset covers diverse scenarios from brief fine movements, medium-range command execution to long-distance navigation. The dataset also possesses over 5.2K multi-choice questions, including low-level tasks such as Recall and Perception, and also high-level tasks such as Reasoning and Navigation. Seen from Figure 2d, both low-level and high-level tasks take up to approximately 50% of the total questions, enabling a comprehensive evaluation of the various capabilities required for Video-LLMs to perform embodied cognitive tasks during motion in urban open-ended spaces. For each question, the word count ranges from 50 to 250, varying based on the question category and the complexity of the environment and tasks included in the specific video, as is shown in Figure 2e. We finally generate a word cloud in Figure 2f, demonstrating the richness of urban elements included in our MCQs.

4 Experiments

We initially evaluated the performance of 16 popular Video-LLMs on various tasks related to embodied cognition in motion. Subsequently, we conducted detailed analyses focusing on the models, tasks, and video data sources. Finally, we summarized and categorized the reasons for failures across different tasks.

4.1 Experimental Setup

Evaluation Metric: Benefiting from the multiple-choice format of each question, we can directly calculate the accuracy for each task type as well as the overall accuracy.

Baselines: The baseline includes both proprietary and open-source Video-LLMs. For proprietary models, we used state-of-the-art models, including GPT-4o, GPT-4o-mini (OpenAI, 2025), Gemini-1.5 Flash, Gemini-1.5 Pro (Team et al., 2024), Gemini-2.0 Flash (Google, 2025) and Qwen-VL-Max (Cloud, 2025). For the open-source models, we focus on those capable of video/multiple image input, including LLaVA-NeXT-Video-7B (Liu et al., 2024a), Kangaroo (Liu et al., 2024b), Qwen2-VL series (Wang et al., 2024b) and InternVL2 series (Chen et al., 2024c). In terms of input frame numbers, the Gemini and Qwen series allow for convenient adjustment of the input parameter *fps*, while others require input as *frame number*. To ensure a fair comparison while considering local computational resource constraints, the goal is to input as many frames as possible.

Others: All local model inference and fine-tuning is performed on three NVIDIA RTX A6000. Detail calculation of evaluation metric, baseline settings, and prompts can be found in Appendix C.

4.2 Model Comparison

The quantitative result is shown in Table 3. We can draw the following conclusions:

- Both proprietary models and open-source models exhibit relatively poor embodied cognitive abilities when navigating urban open-ended spaces. The best-performing model, Qwen-VL-Max,

Table 3: Accuracy of different Video-LLMs on embodied tasks. The former section shows existing popular Video-LLMs' results. The latter section demonstrates fine-tuning results for two models, highlighting sim-to-real potential.

			Recall					Perception					Reasoning			Navigation		
Method	Rank	Avg.	*Trajectory Captioning*	*Sequence Recall*	*Object Recall*	*Scene Recall*	*Start/End Position*	*Proximity*	*Duration*	*Landmark Position*	*Goal Detection*	*Cognitive Map*	*Causal*	*Counterfactual*	*Association*	*Progress Evaluation*	*High-level Planning*	*Action Generation*
Baseline																		
Random	-	19.7	18.5	17.0	20.8	13.5	21.8	37.8	35.6	19.7	18.0	21.9	18.2	25.0	18.3	21.8	15.9	16.4
Proprietary Models (API)																		
Gemini-1.5-Flash[1 fps]	4	40.5	39.7	51.8	61.7	79.3	61.3	47.1	59.8	37.8	28.7	47.9	60.0	42.4	20.0	43.3	32.6	34.4
Gemini-1.5-Pro[1 fps]	3	42.5	58.6	61.6	65.0	72.1	66.2	66.4	63.6	37.4	33.8	46.0	63.6	46.2	23.0	38.8	43.8	31.9
Gemini-2.0-Flash[1 fps]	5	38.3	47.9	58.9	63.3	75.7	57.0	66.4	47.7	27.9	27.8	45.3	62.7	24.2	17.8	39.2	48.4	30.5
GPT-4o-mini[32f]	6	36.5	33.0	53.6	48.3	59.5	56.3	69.7	51.5	33.3	31.3	42.4	65.5	47.7	22.9	30.8	57.5	25.4
GPT-4o[32f]	2	43.6	47.6	58.9	65.0	67.6	61.3	63.0	47.7	36.8	42.4	52.8	66.4	44.7	45.8	34.2	67.8	33.8
Qwen-VL-Max-latest[32f]	1	45.5	44.9	70.5	64.2	75.7	73.9	78.2	43.9	44.8	44.7	61.1	77.3	49.2	23.9	38.8	70.0	29.6
Open-source Models																		
LLaVA-NeXT-Video-7B-hf[32f]	3	38.6	55.7	39.3	43.3	61.3	40.8	58.8	52.3	49.5	16.7	26.8	44.5	20.5	58.7	36.6	52.3	19.2
Phi-3.5-vision-instruct[32f]	2	38.7	67.0	57.1	57.5	64.9	45.1	48.7	45.5	49.2	17.0	52.1	51.8	34.8	13.9	33.2	59.7	15.6
Kangaroo[64f]	1	39.2	27.0	66.1	60.8	69.4	53.5	75.6	57.6	35.5	37.2	60.0	64.5	42.4	19.1	32.5	41.9	32.4
Qwen2-VL-2B-Instruct[0.5 fps]	5	31.9	29.9	54.5	30.8	57.7	24.6	69.7	47.7	22.0	22.1	64.2	46.4	35.6	13.5	28.8	44.2	27.3
Qwen2-VL-7B-Instruct[0.25 fps]	4	36.2	36.5	50.9	47.5	65.8	47.2	52.1	48.5	25.1	28.4	55.8	55.5	29.5	11.7	33.9	59.3	32.7
InternVL2-2B[32f]	11	27.6	19.2	29.5	37.5	55.9	22.5	57.1	37.9	19.3	24.6	39.2	33.6	45.5	33.5	29.2	37.6	20.9
InternVL2-4B[32f]	10	28.1	19.2	37.5	33.3	62.2	24.6	66.4	42.4	23.2	26.5	32.8	36.4	35.6	24.8	29.5	32.2	22.1
InternVL2-8B[32f]	9	28.1	23.4	23.2	35.0	52.3	22.5	58.0	44.7	23.1	27.4	28.3	33.6	45.5	27.0	31.5	35.7	21.4
InternVL2-26B[32f]	8	28.3	24.3	36.6	35.0	61.3	26.8	51.2	40.2	19.9	28.1	32.4	32.7	44.7	26.5	28.9	37.6	22.8
InternVL2-40B[32f]	7	28.4	22.2	19.6	30.8	54.1	21.1	61.3	50.0	23.2	26.5	34.7	27.3	41.7	25.7	32.4	34.9	22.3
InternVL2-Llama3-76B[32f]	6	28.9	19.5	38.4	37.5	54.1	18.3	65.5	48.5	22.9	28.1	33.6	30.9	43.2	27.4	31.3	34.5	23.2
Fine-Tuning:Training set																		
InternVL2-4B(before)[32f]	4	28.0	17.5	32.9	34.2	61.5	26.9	66.2	41.7	21.0	25.3	37.0	37.3	33.3	25.8	30.8	35.0	21.4
InternVL2-4B(after)[32f]	1	31.4	22.0	34.3	40.8	60.0	24.7	73.0	53.6	21.0	44.1	51.1	31.3	42.9	34.4	36.5	35.0	19.6
InternVL2-8B(before)[32f]	3	29.4	22.4	25.7	35.5	53.8	22.6	59.5	48.8	23.2	30.0	30.4	38.8	40.5	33.3	35.9	34.2	20.8
InternVL2-8B(after)[32f]	2	31.2	21.1	35.7	42.1	61.5	23.7	74.3	51.2	21.4	42.4	52.6	29.9	38.1	34.4	36.3	35.8	19.3
Fine-Tuning:Test set																		
InternVL2-4B(before)[32f]	3	28.3	21.3	45.2	31.8	63.0	20.4	66.7	43.8	27.1	27.9	28.5	34.9	39.6	24.1	27.4	29.7	23.0
InternVL2-4B(after)[32f]	2	31.5	25.5	38.1	34.1	60.9	20.4	66.7	37.5	22.1	38.8	33.1	32.6	50.0	31.4	28.1	39.9	28.9
InternVL2-8B(before)[32f]	4	26.5	24.5	19.0	34.1	50.0	22.4	55.6	37.5	22.8	24.5	26.2	25.6	54.2	22.6	24.3	37.0	22.1
InternVL2-8B(after)[32f]	1	31.7	25.5	35.7	34.1	60.9	18.4	66.7	39.6	23.8	37.4	31.5	34.9	50.0	32.8	27.7	39.1	29.4

achieves an average accuracy of only 45.5%. This underscores the value of UrbanVideo-Bench, highlighting that embodied cognitive abilities in urban three-dimensional spaces have not been adequately addressed.

- Some open-source Video-LLMs outperform part of proprietary models. Specifically, models that have been optimized for video data demonstrate superior performance compared to LMMs that focus on images.
- Smaller parameter models appear to be more unstable. For two models with equivalent average accuracy, the open-source small parameter model tends to have a lower minimum accuracy across all tasks compared to the commercial large parameter model.

4.3 Correlation of Cognitive Abilities

We tend to explore the relationships between these tasks and the underlying cognitive abilities, as shown in Figure 3. Specifically, we compute the pairwise correlations between each column (representing individual tasks) in Table 3, excluding the fine-tuning portion. The implicit assumption in this approach is that if multiple models exhibit similar performance across two tasks, it suggests that the embodied cognitive abilities needed for these tasks are similar. We can derive the following conclusions:

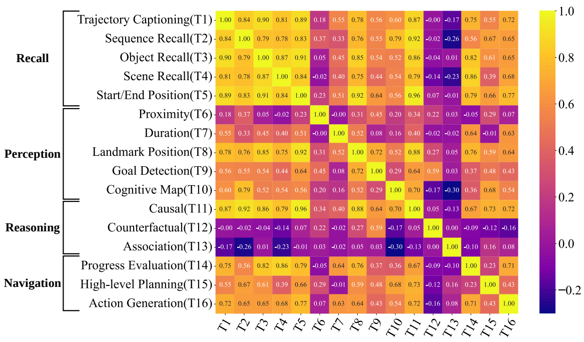

Figure 3: Correlation of Cognitive Abilities. A higher value indicates more similar Video-LLMs' performance on the two tasks, implicitly indicating the relationship between the two tasks.

- The **causal reasoning** task exhibits a high correlation with almost all other tasks. It suggests that the **ability to understand and infer causality is fundamental to a wide range of cognitive processes**. This finding may indicate that causal reasoning is potentially a key factor in the emer-

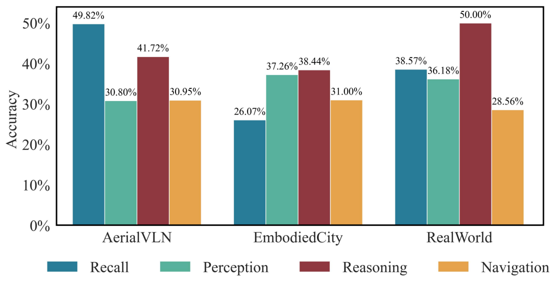

Figure 4: Average performance of 17 Video-LLMs in different video sources.

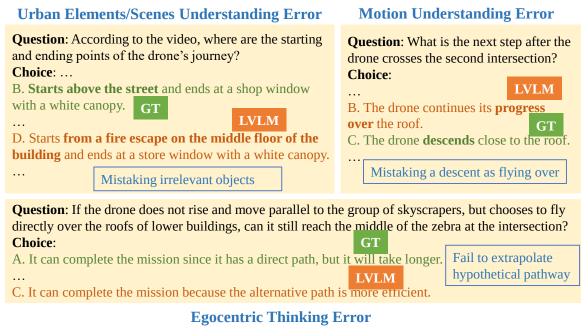

Figure 5: Three common errors in Video-LLMs when performing tasks of urban embodied cognition, along with the corresponding examples.

gence of embodied cognitive in Motion.

- **Recall**-type tasks demonstrate strong inter-correlations among themselves. These tasks all involve memory-related problems, underscoring the ability to recall information is a shared underlying requirement for these tasks, highlighting memory as a central component in cognitive task execution.

- **Navigation** tasks have a high correlation with both **Recall and Perception** tasks. This observation aligns with prior knowledge that effective action and planning depend on robust memory and perceptual capabilities.

- **Counterfactual and Association** reasoning —both high-level reasoning tasks—exhibit low correlations with other task types. These tasks rely on distinct cognitive processes that are not shared with the other tasks in our analysis. This suggests that some embodied cognitive abilities may operate independently rather than as components of a general intelligence framework. Therefore, **when tasks involve these two high-level abilities, targeted training is necessary.**

4.4 Sim-to-Real

We average the performance of all Video-LLMs and listed the results of four categories of cognitive abilities across three different data sources, as shown in Figure 4. From the model performance perspective, there is no significant distributional difference among the various video sources overall. Embodied research has long suffered from a lack of real-world data. Here, we used data from EmbodiedCity and AerialVLN as the training set, and real-world data as the test set. We employed LoRA (Hu et al., 2021) to fine-tune the large models InternVL-4B and InternVL-8B to explore the potential for Sim2Real transfer (see Appendix D for details on fine-tuning.). The results, as shown in Table 3, indicate that both Goal Detection and Association Reasoning on the test set improved by approximately 7% post-fine-tuning. The mean improvements for the two fine-tuned models were 3.2% and 5.2%, respectively.

4.5 Error Analysis

After examining the reasoning processes of Video-LLMs, three common error types were identified, as shown in Figure 5:

- **Error 1: Urban Elements/Scenes Understanding**: Complex urban scenes pose challenges to perception-related tasks for Video-LLMs, resulting in insufficient alignment and urban hallucinations. This means that the models guess based solely on textual content and fail to detect urban objects that are entirely absent in the video during analysis.

- **Error 2: Motion Understanding**: Video-LLMs struggle to distinguish orientation and misinterpret changes in the camera's gimbal angle as vertical movement, indicating limited spatial awareness.

- **Error 3: Egocentric Thinking**: Video-LLMs fail to perform complex embodied reasoning tasks, such as route planning and extrapolation.

Recall and Perception tasks, reliant on visual abilities, suffer from urban elements/scenes and agent motion understanding error. As for the complex and challenging tasks of reasoning and navigation, various errors are prevalent. To analyze current Video-LLMs' shortcomings in detail, we randomly sampled 800 incorrect responses across all Video-LLMs, with 200 cases selected for each task category. Using a combination of Gemini-1.5-Pro and human evaluation, we derive the proportion of each error type in Table 4.

Table 4: Error Analysis Across Tasks

Task	Erro 1 (%)	Error 2 (%)	Error 3 (%)
Recall	38.0	62.0	0.0
Perception	69.5	28.5	2.0
Reasoning	79.5	12.0	8.5
Navigation	37.0	58.0	5.0

Results suggest that Urban Elements/Scenes Understanding Error is the most prominent issue, reflecting challenges in basic multimodal alignment (Peng et al., 2025). Motion Understanding Errors, unique to embodied cognition, are the second most significant. Egocentric Thinking Errors are less frequent, likely due to limited perception abilities, which prevent higher-level errors from being fully exhibited. See Appendix E for more details.

5 Conclusion and Future Work

In this work, we propose UrbanVideo-Bench, a benchmark for embodied motion cognition in urban open spaces, comprising 1.5k video clips and 5.2k multiple-choice questions. We evaluated the performance of 17 currently popular Video-LLMs in terms of recall, perception, reasoning, and planning. The experimental results indicate that the best current Video-LLMs achieve only a 45.5% accuracy rate. Our analysis further reveals that causal reasoning is highly correlated with other tasks such as recall, perception, and planning. Fine-tuning large models with simulation data can enhance their performance on real-world embodied video tasks.

To enhance the embodied cognitive abilities of models in the future, we propose three directions worth further exploration. First, incorporating more diverse and embodied urban datasets, pre-training strategies can address the limitations of current benchmarks that predominantly feature disembodied scenarios. Second, post-training techniques, such as combining supervised fine-tuning with reinforcement learning, are potential to enhance the embodied reasoning capabilities of Video-LLMs. Finally, agent-based approaches are useful to improve performance in specific application domains, leveraging methodologies like SLAM, cognitive mapping, and chain-of-thought reasoning.

Acknowledgments

This paper was supported by National Key Research and Development Program of China 2024YFC3307603, Natural Science Foundation of China under Grant 62371269, 72442026, U23B2030, 62171260, Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund Research Plan (JC20220011), Meituan Academy of Robotics Shenzhen, Talent Program of Guangdong Province (2021QN02Z107), and National Science and Technology Major Project (2024ZD01NL00103).

References

Stefano Burigat, Luca Chittaro, and Riccardo Sioni. 2017. Mobile three-dimensional maps for wayfinding in large and complex buildings: Empirical comparison of first-person versus third-person perspective. *IEEE Transactions on Human-Machine Systems*, 47(6):1029–1039.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Xinlei Chen, Aveek Purohit, Shijia Pan, Carlos Ruiz, Jun Han, Zheng Sun, Frank Mokaya, Patric Tague, and Pei Zhang. 2017. Design experiences in minimalistic flying sensor node platform through sensorfly. *ACM Transactions on Sensor Networks (TOSN)*, 13(4):1–37.

Xuecheng Chen, Haoyang Wang, Yuhan Cheng, Haohao Fu, Yuxuan Liu, Fan Dang, Yunhao Liu, Jinqiang Cui, and Xinlei Chen. 2024a. Ddl: Empowering delivery drones with large-scale urban sensing capability. *IEEE Journal of Selected Topics in Signal Processing*.

Xuecheng Chen, Haoyang Wang, Zuxin Li, Wenbo Ding, Fan Dang, Chengye Wu, and Xinlei Chen. 2022. Deliversense: Efficient delivery drone scheduling for crowdsensing with deep reinforcement learning. In *Adjunct proceedings of the 2022 ACM international joint conference on pervasive and ubiquitous computing and the 2022 ACM international symposium on wearable computers*, pages 403–408.

Xuecheng Chen, Zijian Xiao, Yuhan Cheng, Chen-Chun Hsia, Haoyang Wang, Jingao Xu, Susu Xu, Fan Dang, Xiao-Ping Zhang, Yunhao Liu, et al. 2024b. Sosscheduler: Toward proactive and adaptive wildfire suppression via multi-uav collaborative scheduling. *IEEE Internet of Things Journal*, 11(14):24858–24871.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2023. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. 2024. Videgothink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623*.

Alibaba Cloud. 2025. Qwen documentation. https://tongyi.aliyun.com/. Accessed: 2025-01-24.

Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. 2025. Ecbench: Can multimodal foundation models understand the egocentric world? a holistic embodied cognition benchmark. *arXiv preprint arXiv:2501.05031*.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. 2024. Understanding world or predicting future? a comprehensive survey of world models. *arXiv preprint arXiv:2411.14499*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, Bangkok, Thailand. Association for Computational Linguistics.

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

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, et al. 2024. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*.

Google. 2025. Gemini api documentation. https://ai.google.dev/gemini-api/docs. Accessed: 2025-01-24.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Qiuyi Gu, Zhaocheng Ye, Jincheng Yu, Jiahao Tang, Tinghao Yi, Yuhan Dong, Jian Wang, Jinqiang Cui, Xinlei Chen, and Yu Wang. 2025. Mr-cographs: Communication-efficient multi-robot open-vocabulary mapping system via 3d scene graphs. *IEEE Robotics and Automation Letters*.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. 2024. Citynav: Language-goal aerial navigation dataset with geographic information. *arXiv preprint arXiv:2406.14240*.

Chunyuan Li. 2023. Large multimodal models: Notes on cvpr 2023 tutorial. *arXiv preprint arXiv:2306.14895*.

Jing Li, Dong Hye Ye, Timothy Chung, Mathias Kolsch, Juan Wachs, and Charles Bouman. 2016. Multi-target detection and tracking from a single camera in unmanned aerial vehicles (uavs). In *2016 IEEE/RSJ* international conference on intelligent robots and systems (IROS)*, pages 4992–4997. IEEE.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. 2024. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. *arXiv preprint arXiv:2409.20365*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024b. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. 2023. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15384–15394.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. 2023. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. 2024. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. 2022. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*.

Nora S Newcombe, Mary Hegarty, and David Uttal. 2023. Building a cognitive science of human variation: Individual differences in spatial navigation.

OpenAI. 2025. Openai api documentation. https://platform.openai.com/docs/overview. Accessed: 2025-01-24.

Ruiying Peng, Kaiyuan Li, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. 2025. Understanding and evaluating hallucinations in 3d visual language models. *arXiv preprint arXiv:2502.15888*.

Xuran Pu, Jianjie Fang, Zhiyuan Deng, Xueqian Wang, Xinlei Chen, et al. A large language model-driven heterogeneous air-ground search swarm. In *ICLR 2025 Workshop on Embodied Intelligence with Large Language Models In Open City Environment*.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. 2024. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*.

Michael J Richardson, Kerry L Marsh, and RC Schmidt. 2010. Challenging the egocentric view of coordinated perceiving, acting, and knowing. *The mind in context*, pages 307–333.

Karan Samel, Apoorva Beedu, Nitish Sontakke, and Irfan Essa. 2024. Exploring efficient foundational multi-modal models for video summarization. *arXiv preprint arXiv:2410.07405*.

Lawrence Shapiro. 2019. *Embodied cognition*. Routledge.

Haochen Shi, Zhiyuan Sun, Xingdi Yuan, Marc-Alexandre Côté, and Bang Liu. 2024. OPEx: A component-wise analysis of LLM-centric agents in embodied instruction following. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 622–636, Bangkok, Thailand. Association for Computational Linguistics.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pages 256–274. Springer.

Kunal Pratap Singh, Jordi Salvador, Luca Weihs, and Aniruddha Kembhavi. 2023. Scene graph contrastive learning for embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10884–10894.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

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

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249.

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

Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. 2024a. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5289–5298.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shuo Wang, David C Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S Arya, Anuj Sharma, Pranamesh Chakraborty, et al. 2024c. The 8th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7261–7272.

Yuhao Wang, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024d. MM-SAP: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9192–9205, Bangkok, Thailand. Association for Computational Linguistics.

Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. 2021. Detection, tracking, and counting meets drones in crowds: A benchmark. In *CVPR*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024a. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024b. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. 2024c. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*.

Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. 2025. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics. *arXiv preprint arXiv:2502.11859*.

Yanggang Xu, Jirong Zha, Jiyuan Ren, Xintao Jiang, Hongfei Zhang, and Xinlei Chen. 2024. Scalable multi-agent reinforcement learning for effective uav scheduling in multi-hop emergency networks. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 2028–2033.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Fanglong Yao, Yuanchang Yue, Youzhi Liu, Xian Sun, and Kun Fu. 2024. Aeroverse: Uav-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models. *arXiv preprint arXiv:2408.15511*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Qingbin Zeng, Qinglong Yang, Shunan Dong, Heming Du, Liang Zheng, Fengli Xu, and Yong Li. 2024. Perceive, reflect, and plan: Designing llm agent for goal-directed city navigation without instructions. *arXiv preprint arXiv:2408.04168*.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*.

Weichen Zhan, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. 2025. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.

Weichen Zhang, Ruiying Peng, Chen Gao, Jianjie Fang, Xin Zeng, Kaiyuan Li, Ziyou Wang, Jinqiang Cui, Xin Wang, Xinlei Chen, et al. 2025. The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? *arXiv preprint arXiv:2504.04540*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10012–10022.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. 2021. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399.

A Dataset Examples

To better illustrate the proposed dataset, we provide MCQ examples and videos from real world, the EmbodiedCity simulator, and the AerialVLN simulator, encompassing all task types. They are presented in Fig. 6, Fig. 7, and Fig. 8.

B Dataset Generation Pipeline

B.1 Details of Video Curation

Selected Simulator Advantages: a) Realistic environment modeling. Both EmbodiedCity and AerialVLN are built on Unreal Engine and include various building types and streets, with more than a hundred categories of micro urban elements, which enriches the semantic information of the obtained embodied mobility video data. b) Support for aerial agents. Both simulators have built-in AirSim plugins, allowing easy control of aerial agents. c) Existing route reference. Research work on vision-language navigation (Zhu et al., 2020) has already been conducted in these simulators, allowing us to obtain some route coordinates and instruction data. Although most routes cannot be directly used (for example, most flying paths in AerialVLN contain many meaningless, repetitive flying actions and lack logicality), they provide some reference for our data collection.

Drone Setup: The drone has only one camera equipped with a gimbal, which can tilt from 0 degrees to 90 degrees downward. The second group consists of experienced pilots (with over 1000 hours of real drone flight time) to ensure the rationality of flight operations. Through this approach, the movement of the agent is purposeful, and the collected video data is more conducive to evaluating the capabilities of Video-LLMs.

B.2 Details of MCQ Generation

In this work, we use LLM to generate the multiple choice questions (MCQ) in the dataset.

a) Narration Generation: Initially, the LMM is prompted to systematically generate a narration of the UAV's trajectory by combining embodied movement videos and destination instructions, based on the videos from Embodied City. Videos from the other two sources are originally combined with trajectory infomation.

b) Structured Extraction of Key Information: The LMM then extracts a list of movements and objects, providing structured text that ensures the subsequent MCQ generation is more aligned with the requirements.

c) Role-playing: In the last MCQ generation prompt, the model is given a specific context and role, enhancing its understanding of the task and adherence to instructions.

d) Providing Question/Option Templates and Examples: For each task, several templates for questions and options are provided, with sections marked for replacement. We also provide detailed task definitions and examples.

In the prompt, we first set up a scenario for the model, one where the model act as a teacher who needs to raise a series of test questions based on the given videos. This role playing trick makes it easier

Figure 6: Videos collected from the real drones and MCQs examples.

for us to explain the details of the question generation task onward. Then, we break the task into several sequential parts, and give the specific requirements for the sub-tasks. The sub-tasks include video understanding, question generation, answer generation, and finally structured input/output, with requirements, templates and examples. As we have 14 categories of questions focusing on different aspect of the video, it's hard to cover them all in one general instruction, so we adopt respectively written instructions for question and answer generation, each having its own task explanation, template and example output. During tests, the LLM sometimes gives random explanatory text before it raises the question, which could be an obstacle in the processing work, so we added specific instructions in the prompt to prevent the model from doing so. The main part of the prompt we use are shown in Figure 9, though most of the explanations, templates and examples for question generating are left out due to space constraints.

In question generation, we find that the LLM still has problem understanding complicated city environments in the videos, so we introduced pre-generated structured object list and movement series to assist the model in understanding the videos. Also, this step is done by prompts shown in Figure 10 and Figure 11.

B.3 Details of Human Refinement

Following the automatic question generation phase, we implemented a systematic human refinement process to ensure the quality and reliability of the generated questions. During the refinement phase, we identified and addressed various issues across the generated questions, with statistics shown in Table 5. This process focused on four major aspects requiring manual intervention, with detailed examples provided in Table 6. First, ambiguous or unclear questions were identified and refined

Figure 7: Videos collected from the EmbodiedCity simulator and MCQs examples.

to ensure precise communication of the intended query. Second, hallucinated urban elements, which were occasionally generated by the LLM but not present in the actual video content, were either corrected or removed to maintain factual accuracy. Third, directional descriptions were standardized by converting absolute directions to relative ones, consistently using the drone as the reference system. This standardization was crucial for maintaining consistency in spatial relationships across the dataset. Finally, multiple-choice options were carefully reviewed and adjusted to ensure appropriate differentiation between choices while maintaining one unambiguously correct answer.

The refinement process significantly enhanced the dataset's quality by eliminating potential sources of confusion and ensuring all questions accurately reflected the video content. Human annotators were specifically trained to identify these issues and apply standardized corrections, resulting in a more robust and reliable benchmark for evaluating video-based navigation understanding.

Table 5: Statistics of Issues in Human Refinement Phase

Issue	Count
Invalid/ambiguous questions	761
Urban element hallucination	184
Direction	446
Choices with insufficient differentiation or errors	2363

C Additional Experimental Settings

C.1 Evaluation Metric Calculation

This study employs a multi-stage workflow to evaluate the accuracy of a model in video-based multiple-choice question answering tasks. This workflow integrates regular expressions and model inference for efficiency, complemented by manual validation to ensure robustness. The combined approach achieves precise and scalable accuracy assessment for the target task.The procedure is struc-

Figure 8: Videos collected from the AerialVLN simulator and MCQs examples.

tured as follows:

- **Answer Extraction and Correction.** First, answers are extracted from the model-generated text. Regular expressions are applied to identify standardized outputs (e.g., options starting with "A." or standalone alphabetic characters). For unstructured or ambiguous text, the GPT-4o model is invoked to infer the most plausible option (A–G) based on the output context. Entries unresolved by automated methods are flagged with the special token "l**k". A manual verification step is subsequently performed to correct these flagged entries, ensuring the validity of all answers.

- **Data Alignment and Cleaning.** The corrected answers are aligned with a reference dataset containing ground-truth answers. Data consistency is ensured through format standardization (e.g., removing ".mp4" suffixes from video IDs and normalizing column names). Valid entries are filtered using an inner join operation, merging model predictions with reference data based on video IDs and question categories.

- **Statistical Analysis and Result Export.** Accuracy is computed by category through group-wise comparisons of matched answers, followed by aggregation to derive the overall accuracy. Results are organized into structured tables, detailing per-category and total accuracy values. These outputs facilitate systematic performance evaluation and further analysis.

C.2 Brief Introduction on Baselines

We introduce the participating long-context LMMs as follows:

Gemini-1.5-Flash. Released on February 14, 2024, with an API service, Gemini-1.5-Flash is a lightweight model (parameter count not publicly disclosed). It supports an input token limit of 1 mil-

Prompt for Question Generation

MAIN INSTRUCTIONS:
You are the teacher of the course "Video comprehension and Spatial reasoning". To test the students, you need to make many test questions from a series of egocentric videos taken by a UAV while it flies towards a specific known destiny in a city scene.

Your goal is to raise multi choice questions about the details and spatial or temporal logic from video that satisfy later-on given requirements.

For each video, you must raise a multi choice question, that must be strictly restricted to and strongly related to the video content.

For each question, you need to raise the question itself, and then give 5 choices, labeled as A, B, C, D and E, including ONLY 1 CORRECT answer and 4 wrong answers. Specially, the "Proximity", "Duration" and "Counterfactual" questions should only contain 3 choices, including 1 correct answer and 2 wrong answers. Specially, the "Viewpoint Invariance" questions should only contain 2 choices, including 1 correct answer and 1 wrong answer. For the choices, you need to make sure that the correct answer is not too obvious, and the wrong answers are not too irrelevant. Finally, you need to give the correct answer to the question, which is one of A, B, C, D and E (or one of A, B and C, or one of A and B). You are allowed to put the correct answer at any position among the 5 choices.

All possible questions falls into 14 categories, that covers different aspects of the student's ability to understand the video content, spatial and temporal relationship and causal logic.

The categories, templates and examples of Question, Choices and Answer are given respectively as follows:

1. Captioning
TASK EXPLANATION: This question requires the student to summarize the UAV's movement route in the video by combining the specific objects along the way. Proper answer should be a concise summary of the UAV's movement route, including its starting location, final destiny and specific buildings, objects along the way.

TEMPLATE Question: Summarize the UAV's [movement route] in the video by combining the [specific objects] along the way.

TEMPLATE Choices: The UAV goes from [specific objects] to the [specific objects] by [specific movenments about specific objects].

EXAMPLE OUTPUT:
Question: According to the video, which of the following choices better summarizes the UAV's movement route?
Choices:
A. The UAV goes from [between two tall skyscrapers] to the [the balcony on the 31st floor] by flying straight across the opening ground below.
B. The UAV goes from [the center of the square] to the [left of the main entrance] by descending height to the rooftop.
C. The UAV goes from [the main entrance] to the [right of the main entrance] by turning left to face the main entrance.
D. The UAV goes from [the balcony on the 10th floor] to the [the center of the square] by descending to the ground level.
E. The UAV goes from [between two tall skyscrapers] to the [the balcony on the 30th floor] by flying straight across the opening ground below.
Answer: A

--------------------(The explanation, template and example for other categories) --------------------

NOTICE: The [category] and the [destiny] is known, given by textual input.

EXAMPLE INPUT:
Please raise a [category] multi choice question based on the video taken along the way towards [destiny].
For your reference, the content of the video is shown in a chronological narration, which is given as follows:

........(narration text)........
Also, the objects and positions mentioned in the questions and correct choices should come from the video frames, or you can refer to the list below:
........(object list)........
The frames from the video are as follows:
........(video frames)........

EXAMPLE OUTPUT:
Question: ....?
Choices: A. ....
B. ....
C. ....
D. ....
E. ....
Answer: A

You output must only contain the Question, Choices and Answer in the form shown above, and any other reply, such as "Certainly, here's a .... question about... " is definitely NOT ALLOWED!

The INPUT are as follows:
Please raise a {question_category} multi choice question based on the video taken along the way towards {video_destiny}.
For your reference, the content of the video is shown in a chronological narration, which is given as follows:
{narration_text}
Also, the objects and positions mentioned in the questions and correct choices should come from the video frames, or you can refer to the list below:
{object_list}
The frames from the video are as follows:
{video_frames}

Figure 9: The prompt used in MCQ generation. The prompt contains scenario setup, detailed instructions, respectively written template and example output, and finally input template. Role playing and formatted structure help the LLM to better understand the user's intentions.

lion, an output token limit of 8192, and a maximum video duration of 1 hour. The frame rate is set to 1 fps.

Gemini-1.5-Pro. Released on February 14, 2024, with an API service, Gemini-1.5-Pro is an advanced model (parameter count not publicly disclosed). It supports an input token limit of 2 million, an output token limit of 8192, and a maximum video duration of 2 hours. The frame rate is set to 1 fps.

Gemini-2.0-Flash. Released on December 11, 2024, Gemini-2.0-Flash is the latest lightweight model in the Gemini series, offering improved efficiency and a 12M context length. The input token limit is 1 million, the output token limit is 8192, and the frame rate is 1 fps.

GPT-4o-mini. Released on July 18, 2024, GPT-4o-mini is a compact version of GPT-4o, designed for faster inference with a 128K context length. Parameter count is not publicly disclosed, an input token limit of 128K, an output token limit of 16384, and a frame rate of 32 f.

GPT-4o. Released on May 13, 2024, with an API service, GPT-4o is the latest multimodal LMM from OpenAI, featuring a 128K context length. Parameter count is not publicly disclosed, an input token limit of 128K, an output token limit of 4096, and a frame rate of 32 f.

Qwen-VL-Max-latest. Released in April 2024, Qwen-VL-Max-latest is the most advanced model in the Qwen-VL series, supporting multimodal tasks with a 128K context length. It has an input token limit of 128K, an output token limit of 8192, and a frame rate of 32 f.

LLaVA-NeXT-Video-7B-hf. Released in April 2024, LLaVA-NeXT-Video-7B-hf is a video-focused LMM with 7B parameters. In this experiment, we set the frame rate to 32 f and the output token limit to 512.

Phi-3.5-vision-instruct. Released in August 2024, Phi-3.5-vision-instruct is an upgraded version of Phi-3-Vision-Instruct, with 4.2B parameters and a 128K context length. In this experiment, we set the frame rate to 32 f and the output token limit

Prompt for Object List Generation

MAIN INSTRUCTIONS:
In the "Egocentric UAV Video Scene Recall and Object Extract" task, you are working with a chronological series of frames from an egocentric video taken by a UAV while it flies in a city scene.
Your goal is to write a chronological list of the objects and positions the UAV came across along its route shown in the video frames.

You should give the list in a chronological order, and each item should be a detailed phrase describing the object or position the UAV came across in the video frame, so that there'd be no confusion.

Especially, you should put the starting and ending position of the UAV in the list, on top and bottom respectively.
A narration of movements of the UAV are given to you in textual form, to which you can refer for generation.

The output should be a list of phrases, following the format below:

EXAMPLE OUTPUT: (.json)
{
"The building beside the square", "The window to the right of the main gate", "The red car parked on the side of the road", "The traffic light at the closest intersection"
}

NOTICE: Instead of using " ' " to quote each phrases, you should use " " " to quote each phrases in order to avoid errors in the JSON format. There should be no other characters.
NOTICE: Extra explanation of the output is not allowed. Your output should be a list of phrases ONLY.

The inputs are as follows:
The destiny of the UAV is: {video_destiny}
The movements of the UAV are:
{narrations}
The frames are as follows:
{video_frames}

Figure 10: The prompt used in object list generation. The prompt contains scenario setup, detailed instructions and example output. As a simpler task than question generation, input/output template is not needed.

to 512. The temperature coefficient is set to 0, and the maximum possible output is selected.

Kangaroo. Released on July 17, 2024, Kangaroo is a specialized LMM for multi-image and video understanding, with 8B parameters. Since the frame rate cannot be modified in the code, in this experiment, we used the default frame rate of 64 f and set the output token limit to 256. The temperature coefficient is set to 0, and the maximum possible output is selected.

Qwen2-VL-2B-Instruct. Released on August 30, 2024, Qwen2-VL-2B-Instruct is a lightweight instruct-tuned model with 2B parameters and a 32K context length. In this experiment, we set the resolution to 360 * 420 and adopted the model's default settings. The frame rate is 0.5 fps, which is the maximum limit for our GPU's computational capacity. The output token limit is set to 256.

Qwen2-VL-7B-Instruct. Released on August 30, 2024, Qwen2-VL-7B-Instruct is a mid-sized instruct-tuned model with 7B parameters and a 32K context length. In this experiment, we set the resolution to 360 * 420 and adopted the model's default settings. The frame rate is 0.25 fps, which is the maximum limit for our GPU's computational capacity. The output token limit is set to 256.

InternVL2 series (2B, 4B, 8B, 26B, 40B, Llama3-76B). Released on July 4, 2024, the InternVL2 series offers a range of models from 2B to 76B parameters, with context lengths scaling from 32K to 256K, designed for diverse multimodal tasks. In this experiment, we set the resolution to 448 * 448 and adopted the model's default settings. The frame rate is 32 f, and the output token limit is set to 1024.

C.3 Responses of Video-LLMs

Table 7 presents the format correctness rates of six proprietary models through API calls. The probabilities indicate each model's capability to reliably produce outputs adhering to the specified format requirements. These results demonstrate the varying

Prompt for Narration Generation

MAIN INSTRUCTIONS:
In the "Egocentric UAV Video Narration" task, you are working with a chronological series of frames from an egocentric video taken by a UAV while it flies towards a specific known destiny in a city scene.
Your goal is to write a chronological Narration of the route of the UAV that are shown in the video frames.

For the video Narration, you should first show the [starting position] and the [final destiny] of the UAV.
Then you chronologically list the specific [movements] and [positions] according to the video frames.

The [movements] may include but not limited to "goes forward to", "turns left to face", "descend to the height of", etc.
The [positions] may include "the center of the square", "the balcony on the 10th floor", "left of main entrance", etc.
The [destiny] is known, given by textual input.

The video frame may contain movements taken on the way to the destiny, but it is possible that the destiny is not reached in the video frame.
For each or each several frames, you should only describe the movements and positions of the UAV shown in the video frame.

EXAMPLE OUTPUT: (.json)
{
"In the video, the UAV goes from in front of the building with an antenna on the roof to the main entrance."
"Movement 1: The UAV descends to the height of the 1st floor."
"Movement 2: The UAV turns downward to face the main entrance."
"............................"
"Movement X: The UAV goes forward to the main entrance."
}

EXAMPLE EXPLAIN:
The first line summarizes the starting position and the destiny of the UAV.
Each of the following lines contains a single, detailed movement of the UAV, Starting with "Movement X:".
The movements are in chronological order, and each movement should be a detailed phrase describing the movement or position of the UAV in the video, so that there'd be no confusion.

The inputs are as follows:
The destiny of the UAV is: {video_destiny}
The frames are as follows:
{video_frames}

Figure 11: The prompt used in movement extraction. The prompt contains scenario setup, detailed instructions, respective explanation of prompt components, and finally input template.

levels of format adherence among state-of-the-art language models.

C.4 Cost of Proprietary Models

In Table 8, we show the cost of our evaluation of proprietary models. The cost of input is significantly greater than the cost of output, suggesting that the existing proprietary models still face the problem of excessive token amount when receiving video input.

C.5 Prompt

For each Video-LLM, the input includes an embodied movement video, a single MCQ, and a prompt that introduces the background and output format.

For models such as Gemini-1.5-flash, the videos can be uploaded to cloud storage space, the questions are given to the model separately for evaluation, as is shown in the following prompt:

Please assume the role of an aerial agent. The video represents your egocentric observations from the past to the present. Please answer the following questions:
<Question>
<video input>
The template for the answer is:
Option: [] (Only output one option from 'A' to 'E' here, do not output redundant content)
Reason: [] (Explain why you choose this option)

For models such as GPT-4o and Qwen, the

Table 6: Examples of human refinement for generated questions.

Issue	Refinement	Example
Invalid/ambiguous questions	Clarify and complete unclear or ambiguous parts in question stems	Navigation Goal: [8th floor of the building (with a huge dome)], what is the ~~direction~~ (spatial relationship) between the current position and the destination when the drone reaches the current position?
		Navigation Goal: [Rooftop of a nearby blue building]. Given the [~~planned milestone~~ (top of the tallest building ahead)], what should be the drone's next action?
Urban element hallucination	Correct hallucinated elements to accurate elements	Navigation Goal: [A store with a green sign reading "~~Meiyue Styling~~" ("Orange Fresh")] What should the drone approach next?
	Remove hallucinated elements	A. Two buildings on sides, highway behind, parking lot below B. Highway adjacent, park behind, walkway below C. High-rise ahead, highway behind, no sidewalk below D. Roads and trees nearby, many high-rises around,~~parking lot below~~. E. Three buildings ahead , park with lanes in front Answer: D
Direction	Use drone as reference system for directions	~~The drone is positioned on the right side of the highway~~ (The highway is positioned on the right side of the drone).
		~~The target building is positioned on the left side of the drone~~ (drone has the target building in its left field of view).
	Convert absolute directions to relative directions	~~The drone is currently located slightly southwest of the target balcony~~ (The target building is positioned on the lower left side of the drone).
Choices with insufficient differentiation or errors	Correct incorrect options to be accurate	Choices: A. park walkway -> high-rise B. Park -> street -> high-rise level -> building side C. Street -> park -> high-rise level -> 15th floor balcony D. ~~Park -> street -> high-rise level -> roof edge -> 15th floor balcony~~ (Left side of domed building -> right turn-> slight descent -> horizontal move to building) E. High-rise -> descent -> park Answer: D
		Choices: A. Grass ascent is longer B. Tree area crossing is longer C. Equal duration Answer: ~~A~~(B)
	Modify one option to contain factual errors	Choices: A. Ascend B. (Backward after) descend C. Forward (after descent) D. Left turn E. Right turn F. Rotate camera upward G. Rotate camera downward Answer: C

Table 7: Format Correctness Rates of Proprietary Models

Proprietary Models (API)	Rate
Gemini-1.5-Flash	0.992762
Gemini-1.5-Pro	0.983810
Gemini-2.0-Flash	0.969714
GPT-4o-mini	0.912190
GPT-4o	0.961143
Qwen-VL-Max-latest	0.997333

Table 8: Evaluation Cost of Different Models

Model	Input Cost	Output Cost	Total Cost
Gemini-1.5-Flash	$13.41	$0.02	$13.43
Gemini-1.5-Pro	$223.49	$0.29	$223.78
Gemini-2.0-Flash	$17.88	$0.02	$17.90
GPT-4o	$95.49	$0.27	$95.77
GPT-4o-mini	$34.31	$0.04	$34.35
Qwen-VL-Max-latest	$4.31	$0.24	$4.55

videos are given to the model one by one, together with the all the questions based on this video, and the models are then required to answer all the questions one by one. In this way, we avoid uploading the same video repeatedly, so as to reduce time consumption and token number. Such prompt are shown below:

> *Please assume the role of an aerial agent. The video represents your egocentric observations from the past to the present. Please answer the following questions:*
> *<Question 0>, <Question 1>, ...*
> *The template for the answer is:*
> *QA0: Option: []; Reason: [].*
> *QA1: Option: []; Reason: []...*
> *The Option only outputs one option from 'A' to 'E' here, do not output redundant content. Reason explains why you choose this option.*

D Fine-Tuning

In this experiment, we conducted multimodal fine-tuning on the `InternVL2-4B` and `InternVL2-8B` models. During fine-tuning, the visual encoder (`freeze_visual_encoder=True`) and the language model backbone (`freeze_llm=True`) were frozen, and only the language model was lightly adapted using LoRA technology (rank 128, alpha 256, dropout 0.05). The experimental data was based on 70% of the multiple-choice question samples from `All_MCQ.jsonl`, with the associated video data totaling 39.56 GB, stored in the `video_LLM/video` directory. The videos were processed by a pre-trained visual encoder to extract spatiotemporal features, which were then concatenated with text embeddings and fed into the model. The maximum input sequence length was set to 8192 to accommodate the joint modeling of long video segments and complex text.

Training employed the AdamW optimizer (learning rate $1e{-}6$, weight decay 0.05) combined with linear warm-up and cosine annealing scheduling. The batch size per GPU was 4, with a gradient accumulation step of 2, resulting in an effective batch size of 8. The maximum training epoch was set to 1. Video data was efficiently loaded using a distributed data loader, with dynamic mixed precision (`AmpOptimWrapper`) and gradient clipping (maximum norm 1) enabled. Checkpoints were saved every 64 steps to ensure training stability and storage efficiency.

E Error Cases

Table 9 and Table 10 shows the results and reasoning process of an Video-LLM (Language and Vision Learning Model) answering questions on a dataset. The table has four columns:

Question Category: This column describes the type of question, such as Action Generation, High-level Planning, etc.

Question: This column contains the specific questions and the corresponding options.

Ground Truth: This column shows the correct answers to the questions, which are carefully checked in human refinement.

Video-LLM Answer: This column provides the answer given by the Video-LLM, and also the reasoning behind the Video-LLM model's choices, demonstrating the model's reasoning process.

Through these questions and their options, answers, and explanations, this table highlights the shortcomings of the Video-LLM in the embodied abilities.

Table 9: Error Example Part 1

Question Category	Question	Ground Truth	Video-LLM Answer
Object Recall	According to the video, what is directly above the balcony on the 24th floor? Choices: A. There is a rooftop garden above the balcony on the 24th floor. B. There is a billboard directly above the balcony on the 24th floor. C. There is another balcony directly above the balcony on the 24th floor. D. There is nothing directly above the balcony on the 24th floor. E. There is a drone stationed above the balcony on the 24th floor.	C	Option D; Reason: There is no mention or indication in the video that there is anything directly above the balcony on the 24th floor.
Object Recall	Question: In the video, what is on the left side of the ""Huaji Stone Rolled Rice Rice Roll"" shop? Choices: A. Large parking lot full of cars. B. Another store with a green storefront and a plus sign that says ""Dashenlin"". C. A shop with a black car parked at the door, and the signboard reads ""NOBEE Nobeni"". D. A small, unpaved area with some overgrown vegetation. E. A bus stop with shelters and benches.	C	Option B; Reason: The video shows that on the left side of the "Hugj Stone Rolled Rice Roll" shop, there is another store with a green storefront and a plus sign that says "Dashenlin".
Duration	Which takes longer, the UAV's descent towards the street level or its maneuvering between the buildings? Choices: A. The UAV's descent towards the street level takes longer than its maneuvering between the buildings. B. The UAV's descent towards the street level takes shorter than its maneuvering between the buildings. C. The UAV's descent towards the street level takes the same time as its maneuvering between the buildings.	A	Option B; Reason: The UAV's descent towards the street level is a straightforward movement, while maneuvering between buildings requires more complex navigation and adjustments, which would take longer.
Duration	Which takes longer, the UAV's descent to street level near the bus stop or its flight over the center of the street between skyscrapers? Choices: A. The UAV's descent to the street level takes longer than its flight over the center of the street. B. The UAV's descent to the street level takes shorter than its flight over the center of the street. C. The UAV's descent to the street level takes the same time as its flight over the center of the street.	A	Option B; Reason: The video shows that UAV's flight over the center of the street between skyscrapers takes longer than its descent to the street level near the bus stop.
Trajectory Captioning	According to the video, which of the following sentences better summarizes the movement of the drone? Choices: A. The drone ascends from the facade of the high-rise building, moves forward, turns right on the last side and descends, and reaches the top of the seaside fence. B. The drone descends from the city center to street level and crosses the main thoroughfare to the top of a skyscraper. C. The drone flew from an open street next to a building over several rooftops to a park near the ocean. D. The drone travels from a bridge over the river through numerous skyscrapers to the main city park. E. The drone turns left from above the road and follows the path along the water into the forest area.	A	Option D; Reason: The video shows the drone starting from the facade of a tall building, moving forward, turning right on the last side, and descending to reach the top of the seaside fence. This matches option D.
Start/End Position	Question: According to the video, where are the starting and ending points of the drone's journey? Choice: A. The drone starts at a large intersection on a city street and ends at a store window with a blue canopy. B. The drone starts above the street and ends at a shop window with a white canopy. C. The drone starts from the fire exit in front of the building and ends at the subway entrance. D. The drone starts from a fire escape on the middle floor of the building and ends at a store window with a white canopy. E. The drone starts in the center of the street and ends in the park.	B	Option D; Reason: The video shows the drone starting from a fire escape on the middle floor of the building and ending at a store window with a white canopy.
Proximity	Question: After the drone moves to the outside of the park, how does the distance between the drone and the staircase-like building change? Choices: A. The distance between the drone and the staircase-like building decreases. B. The distance between the drone and the staircase-like building increases. C. The distance between the drone and the staircase-like building remains constant.	A	Option B; Reason: As the drone moves to the outside of the park, it appears to be moving away from the staircase-like building, increasing the distance between them.
Scene Recall	Question: What objects did the drone observe when it hovered over the trees in the middle of the intersection? Choices: A. The drone observes road markings and phone booths as it hovers over the woods. B. The drone observes a bridge, a car, and a lamppost as it hovers over the woods. C. The drone observed skyscrapers, rooftop gardens, and flags while hovering over trees. D. The drone observed a playground, fountain, and street vendors as it hovered over the woods. E. The drone observed a church, a bell tower and a water tower while hovering over the woods.	A	Option C; Reason: When the drone hovers over the trees in the middle of the intersection, it observes skyscrapers, rooftop gardens, and flags.
Counterfactual	Question: If the drone does not rise and move parallel to the group of skyscrapers, but chooses to fly directly over the roofs of lower buildings, can it still reach the trees in the middle of the zebra crossing at the intersection? Choices: A. If the drone chooses to fly directly over the rooftop, it can complete the mission since it has a direct path, but it will take longer. B. If the drone chooses to fly directly over the roof, it will not be able to complete its mission because it will miss the intersection. C. If the drone chooses to fly directly over the roof, it can complete the mission because the alternative path is more efficient.	A	Option C; Reason: If the drone chooses to fly directly over the roofs of lower buildings, it can still reach the trees in the middle of the zebra crossing at the intersection because this alternative path is more efficient and direct.
Counterfactual	Question: If the drone does not fly along the winding road over the lake, but flies directly over the open fields, can it still reach the narrow bridge across the lake in the park? Choices: A. If the drone chooses to fly directly over an open field, it can accomplish the mission because the field provides a clear path, whereas the alternative would require less time. B. If the drone chooses to fly directly over the open field, it will not be able to complete its mission because it will get lost in the middle. C. If the drone chooses to fly directly over the open field, it can complete the mission because the alternative path is more efficient.	B	Option C; Reason: Flying directly over an open field can help the drone avoid trees, making the task easier and potentially faster.
Sequence Recall	Question: What is the next step after the drone crosses the second intersection? Choice: A. The drone ascends to gain a higher vantage point. B. The drone continues its progress over the roof. C. The drone descends close to the roof. D. The drone turns sharply to the right. E. The drone hovers in place.	C	Option: B; Reason: After crossing the second intersection, the drone continues its journey over the roof, as seen in the video.

Table 10: Error Example Part 2

Question Category	Question	Ground Truth	Video-LLM Answer
Causal	Question: Why do drones keep rising after leaving the beach? Choice: A. The drone ascends to capture a wider view of the waterfront. B. The drone rises, avoiding trees and lamp posts. C. The drone rises to avoid hitting the roof of the building. D. The drone ascends to the upper level of the building. E. The drone ascends to the rooftop helipad.	A	Option: C; Reason: The drone rises to align with the upper level of the building, as shown in the video frames.
Landmark Position	Question: The drone is navigating to (the parking space on the right after entering the community entrance). When the drone reaches the current location, what is the direction to the destination? Choice: A. The drone is currently located above the parking space on the right and has just entered the entrance of the community. B. The drone is currently located far away from the community, above a wide avenue. C. The drone is currently located above the community center and not close enough to the parking space. D. The drone is currently located above the left side of the community, near another high-rise building. E. The drone is currently located above the community center and not close enough to the parking space.	A	Option: E; Reason: The drone is currently located over the community center and needs to move towards the parking space on the right.
Goal Detection	Question: The drone is navigating to the balcony on the 13th floor. Is the destination within sight at the current location? If yes, where is the destination in the view? Choice: A. The balcony on the 13th floor is currently located below the center of the drone's field of view. B. The balcony on the 13th floor is currently located in the upper left corner of the drone's field of view. C. The balcony on the 13th floor is currently located on the right side of the drone's field of view. D. The balcony on the 13th floor is currently located in the lower right corner of the drone's field of view.	A	Option: D; Reason: The video does not show the balcony on the 13th floor within the drone's field of view at the current location.
Goal Detection	Question: The drone is navigating to the balcony on the 13th floor. Is the destination within sight at the current location? If yes, where is the destination in the view? Choice: A. The balcony on the 13th floor is currently located below the center of the drone's field of view. B. The balcony on the 13th floor is currently located in the upper left corner of the drone's field of view. C. The balcony on the 13th floor is currently located on the right side of the drone's field of view. D. The balcony on the 13th floor is currently located in the lower right corner of the drone's field of view.	A	Option: D; Reason: The video does not show the balcony on the 13th floor within the drone's field of view at the current location.
Action Generation	The drone is navigating to the 31st floor balcony. What is the next move for the drone, given the target location for the drone's next move? **Choices:** A. Rise. B. Descend. C. Fly forward. D. Turn left. E. Turn right. F. Rotate the camera upward. G. Rotate the camera downward.	B	Option: A; Reason: The drone needs to rise to reach the 31st floor balcony, so the next move should be to ascend. This aligns with option A.
Action Generation	The drone is navigating to the 13th floor balcony. What is the next move for the drone, given the target location for the drone's next move? **Choices:** A. Rise. B. Descend. C. Fly forward. D. Turn left after descending. E. Fly forward. F. Turn left. G. Turn right. H. Rotate the camera upward. I. Rotate the camera downward.	B	Option: F; Reason: The drone needs to rotate the camera upward to align with the balcony on the 13th floor.
High-level Planning	The drone is navigating to a nearby bus stop. Thinking back on past actions and objects seen in the video, what should the drone approach next in order to get from its current location to its destination? **Choices:** A. The center of the square. B. The top of the tallest building in the distance. C. Crossing the street. D. The end of the parking lot. E. Street edge.	E	Option: C; Reason: Crossing the street is the next step for the drone to get to the bus stop.
High-level Planning	The drone is navigating to the largest intersection nearby. Thinking back on past actions and objects seen in the video, what should the drone approach next in order to get from its current location to its destination? **Choices:** A. The top of the traffic light at the intersection. B. Center of the roundabout. C. The highest point of the residential building across the street. D. Street parking. E. The tallest building visible on the horizon.	B	Option: A; Reason: The top of the traffic light at the intersection. To navigate to the intersection, approaching the top of the traffic light would provide a clear vantage point for identifying the intersection.
Progress Evaluation	The drone moves according to a series of movement instructions. What are the drones doing now? **Choices:** A. Fly towards the sun. B. Go to the clear between the buildings. C. Go to the river. D. Go to the brown building. E. Go to the sun.	B	Option: C; Reason: The drone is currently positioned to the right of the trees as it moves through the cityscape.
Progress Evaluation	The drone moves according to a series of movement instructions. What are the drones doing now? **Choices:** A. Fly towards the sun. B. Turn right to the residential building. C. Increase the height of the drone. D. Change the camera gimbal from downward to normal. E. Fly towards the road.	B	Option: D; Reason: The drone is currently capturing a downward view of the city, and changing the camera gimbal to normal will provide a more standard perspective.