KilometerVision: A New Frontier for Large-Scale Spatial Intelligence in VLMs

Aravindh Mahendran[1⋆], Michael King[2⋆], Matthew Koichi Grimes[2⋆], Antoine Yang[2], Tyler Zhu[2,4⋆⋆], Joseph Heyward[2], Tengda Han[2], Shiry Ginosar[5], Chen Sun[3], Dima Damen[2], Simon Osindero[2], Noah Snavely[3], Simon Lynen[6], João Carreira[2], and Viorica Pătrăucean[2(✉)]

[1] Google DeepMind, Berlin, Germany
aravindhm@google.com
[2] Google DeepMind, London, UK
{mjking, mkg, antoineyang, heywardj, tengda, ddamen, osindero, joaoluis, viorica}@google.com
[3] Google DeepMind, USA
{chensun, snavely}@google.com
[4] Princeton University, Princeton, USA
tylerzhu@cs.princeton.edu
[5] Toyota Technological Institute at Chicago, Chicago, USA
shiry@ttic.edu
[6] Google, Zurich, Switzerland
slynen@google.com

Abstract. We push the frontier of large-scale spatial intelligence in Vision-Language Models (VLMs) and introduce the first benchmark that probes geographical layout understanding from real-world videos, spanning up to 1km distances. Inspired by the cognitive science literature, we evaluate models against the hierarchical stages of human spatial awareness: anchoring via landmarks, connecting them through routes, and integrating these into global mental maps. Extensive experiments reveal a fundamental divergence in how current AI models process spatial information. Instead of utilising true path integration or forming geometric survey knowledge, we find that VLMs rely almost entirely on 2D visual recognition and text-matching to bypass complex spatial reasoning.
The benchmark is publicly available at https://perception-test-challenge.github.io/kilometervision.html.

Keywords: spatial intelligence, city-scale, landmarks, routes, maps

1 Introduction

Spatial intelligence is a hallmark of general intelligence [16]. Species across the animal kingdom develop tailored spatial representations, intrinsically linked to

⋆ Core contributors
⋆⋆ Work done during internship at Google DeepMind.

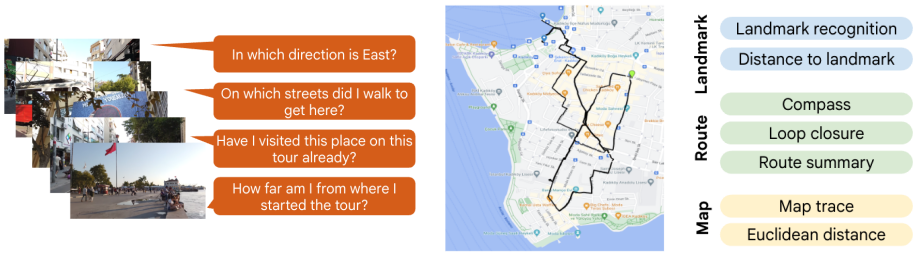


Fig. 1: KilometerVision – the first benchmark probing city-scale spatial understanding from real-world videos spanning 1km distances. *Left:* Real-world hour-long video of a walking tour in Istanbul and relevant questions one would ask when visiting an unfamiliar place. *Centre:* Grounding the video on the map instantly reveals loop closures, distances between landmarks, or sequences of streets visited. *Right:* Using the LANDMARK-ROUTE-MAP paradigm from cognitive science, we define tasks to comprehensively evaluate visual city-scale spatial intelligence in video-language models.

their dominant sensory modalities, enabling them to recognise places, plan paths, and navigate, sometimes across thousands of kilometres [1, 2, 19]. For example, birds rely on geomagnetic fields for orientation, insect-eating bats use echolocation to extract depth, and humans leverage visual and proprioception senses to form mental maps of their environments [44, 45, 58]. Given that current large vision-language models (VLMs) are primarily trained on passive internet-scale data, how do they represent their surroundings?

Multiple works have addressed this question in small-scale environments, studying how VLMs represent objects and their relations [39, 40]. Navigation and planning capabilities have also been studied in simulated interactive environments or factory-like settings [31, 40, 46, 63]. In this work, we push the envelope to city scale and explore if VLMs can infer valid geographic spatial representations from real-world videos of hour-long walking tours.

We present *KilometerVision*, the first benchmark to comprehensively assess VLMs' city-scale spatial intelligence from real-world videos. We take inspiration from the LANDMARKS-ROUTE-MAP paradigm from the large-scale spatial cognitive literature [10, 26, 44] and define tasks in multiple-choice video QA format to comprehensively evaluate the existence of these three types of spatial constructs: *landmarks* (ability to identify and localise salient objects), *routes* (egocentric turn-by-turn representations linking two locations), and *survey knowledge* (allocentric maps integrating multiple routes) [50]; see Fig. 1 (right).

KilometerVision contains 1000 5-way video QAs, defined over 235 YouTube videos. For each question, the model receives a (segment of a) walking tour video together with a question about the video and 5 possible options, out of which only one is correct. The questions are carefully formulated to probe VLMs' visual spatial capabilities rather than their semantic knowledge about various places. We evaluate five state-of-the-art VLMs: Gemini 2.5 Flash [49], PLM-8B [9], Qwen2.5-VL-72B [4], Claude Opus [3], GPT-5 [36]. While our experimental re-

sults show a large performance variation across model families and model sizes, a consistent limitation emerges across all models: rather than utilising true path integration or building geometric mental maps, current VLMs predominantly bypass complex spatial reasoning by relying on 2D visual recognition and semantic text-matching.

2 Related work

Large multimodal models and evaluation: Recent advances in large multimodal models (VLMs) [8, 27, 29, 30, 49, 66] are often achieved by integrating high-dimensional, continuous sensory signals (*e.g.*, visual and audio) into compact representations as inputs to large language models. In order to capture visual information, some approaches [8, 27, 29, 61, 66] employ pre-trained image encoders [37, 65] to extract representations from individual frames, whereas other approaches [32,56] utilize video encoders to extract (short-term) spatiotemporal representations. Beyond distributed representations, it has been demonstrated [30, 54, 57] that interpretable video representations, such as dense captions, often capture sufficient information for many existing video question answering benchmarks, such as NextQA [59] and EgoSchema [33].

The success of VLMs has inspired researchers to introduce more challenging benchmarks beyond action classification [6] and localization [5, 21]. One such attempt aims to evaluate multimodal concept abstraction and spatiotemporal reasoning, in either synthetic [18,64], egocentric [12,20], or carefully curated real-world scenarios [39]. Another attempt focuses on long-form video understanding, where the videos are usually sourced from movies [41, 48], or procedural demonstrations [20,33,67] and vlogs [14]. Finally, a trend in recent benchmarks aims at offering a comprehensive evaluation on various aspects of multimodal perception and reasoning from diverse domains, such as VideoMME [15] and LVBench [55].

Benchmarking spatial intelligence: Traditionally, spatial understanding was evaluated through low-level tasks like depth estimation [17, 34], pose estimation [23], or 3D reconstruction [13,22,28], tackled by specialised methods (SFM [68], SLAM [13, 22]). Given VLMs' potential to become general perception models, more and more efforts are focusing on evaluating their spatial capabilities using a language interface, often taking inspiration from the cognitive literature to design challenging tasks [31, 39, 40, 63]. The main question that these works try to answer is if the current internet-scale training datasets and strategies can allow spatial intelligence to emerge [31, 40, 46, 51, 52, 63], can VLMs learn a valid world model of physical environments by learning from passively observed data.

Large-scale geographic capabilities: Several works have tried to push the envelope of spatial understanding to geographic scale, due to the unique challenges that such settings expose in terms of memory and long-context understanding. The authors of [51] study the world model learnt by an LLM from a large-scale text taxi rides dataset. Still using only text modality, the authors of [42, 62] design tasks probing geographic knowledge in LLMs. In [43] images are provided as

inputs as well. Other works are attempting to place VLM agents in interactive environments to probe their planning and navigation capabilities [7, 25, 60]. Our proposed benchmark is novel and complements these existing benchmarks, being the first to use real-world videos of walking tours [53] to comprehensively probe VLMs' capabilities to recognise landmarks, routes, and build mental maps of cities from videos alone.

3 Video dataset

3.1 Video source

We rely on high-resolution real-world hour-long YouTube Walking Tour videos filmed in various cities around the world. In these videos, the camera wearer walks around the city, visiting landmarks, sometimes returning to places already visited (*i.e.* loop closures). We start from the nine city-life videos in the Walking Tours dataset [53], and select 226 extra YouTube videos with similar characteristics (high-resolution, hour-long), prioritising diverse coverage of cities in different countries, for a total of 235 videos with about 288 hours of video data.

3.2 Grounding videos on the map

To facilitate extracting annotations for our tasks, we first design an efficient pipeline to ground the videos on the map, *i.e.* extract the (*latitude*, *longitude*) coordinates and *camera pose* of each frame in the video at a given frame rate, *e.g.* 1 FPS. This produces detailed accurate route traces on the map (see Fig. 1, centre), which instantly reveal important aspects that would otherwise require watching a long video, possibly multiple times, to discover, *e.g.* loop closures when walking around in a neighbourhood and returning to the same place on a different path. Given such map traces, we can then use simple heuristics and minimal human annotations to extract at scale ground-truth annotations for various spatial tasks.

We leverage Google's *Visual Positioning System* (VPS) [47], a public API that relies on StreetView imagery to estimate the location and camera pose of any query image. Given a query image, VPS calculates its embedding using an image encoder, then retrieves the (lat, lon) coordinates and camera pose of the nearest neighbour image stored in Google StreetView embedding. For this operation to be efficient, an estimate of the location is also necessary (*e.g.* within 100m of the true location). Given this constraint and to run efficiently on long videos, we rely on human raters to provide the (lat, lon) coordinates corresponding to the start and end of each walking tour video in our dataset. Then, we run the VPS API on consecutive video frames at 1FPS, updating the location estimate with the newly returned location as the video progresses.

In more detail, for each given frame, the VPS output is a tuple $(\text{lat}, \text{lon}, \text{rot}, \text{trans}, \text{confidence})$, where (latitude, longitude) are real scalar values, rot represents the camera rotation as a quaternion in the Earth-Centered, Earth-Fixed

(ECEF) coordinate system rot $\in \mathbb{R}^4$, trans is the camera translation as a matrix $\mathbb{R}^{1\times 3}$, and the confidence score is between 0.0 and 1.0. If this confidence score is larger than a given threshold, we update the estimated location with the newly returned location, otherwise we keep the existing estimate when processing the next frame. To improve the quality of the VPS predictions, we repeat the operation but in reverse order using the last frame's human-annotated position as initial estimate and making VPS calls consuming the video frames in reverse. The two passes are then merged by using a confidence-weighted average between the forward pass predictions and the reverse pass. Finally, a last VPS run is executed, using each frame's weighted average as initial estimate for that frame. Note that, despite the repeated runs, this is a robust pipeline that can be run at scale on thousands of hour-long videos; for our 288 hours of videos, it takes about a day to do the full processing. To circumvent VPS errors in feature-poor areas [24], we apply outlier removal as detailed in the appendix.

To validate this approach of grounding videos on the map, we compared the obtained traces against traces collected by humans using Dynamic Time Warping (DTW) as a metric. To calibrate the DTW metric, *i.e.* get a sense of what is an acceptable error between human-collected ground truth and an automatically extracted VPS path, we collected 2 parallel human annotations for a set of 10 hour-long videos ($\sim$15 hours of video data) using Google Maps as annotation interface and asking human raters to draw on the map the path followed by the camera wearer. Twelve more videos ($\sim$14 hours of data) were used for validation. Overall, about 10% of the data has human annotations. For each video, we calculated the DTW distance between the two human-provided paths, giving us a measure of acceptable variability between two paths representing the same walking tour. We then calculated DTW distances between VPS traces and human-provided traces, and discarded segments of the VPS trace where the DTW metric was above the threshold; more details are included in the appendix. It is worth noting that the cost reduction offered by our pipeline is massive. Human raters need only $\sim$5 minutes per video to mark the location of the start and end of the video, then the video is grounded automatically. Without our pipeline, drawing manually the full path on the map took $\sim$6 hours per 1h-long video and it only provides position information, not pose.

In a recent parallel work [28], the authors relied on ARIA glasses to similarly ground walking tour videos on the map and design a city-scale benchmark for SLAM systems. Compared to this work, which requires a person to wear the ARIA glasses and walk around to collect the videos, our system allows us to ground at scale any *already-filmed* video, with any camera, anywhere in the world.

4 Benchmark tasks

As evidenced in the cognitive literature, continued exposure to a new environment leads to three stages of spatial awareness in humans: LANDMARK-ROUTE-MAP [10]. At the landmark stage, we observe salient objects (*e.g.* a clock tower

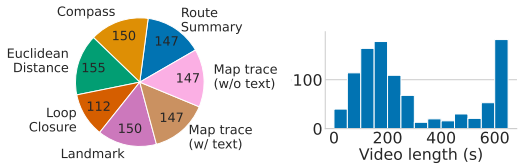


Fig. 2: Left: Distribution of question types in the KilometerVision benchmark. Right: Histogram over video lengths across question types.

or a statue) that we can easily recognise and allow us to anchor ourselves in space. At the route stage, we infer sequences of turns that allow us to move from one landmark to another. Finally, the mental map stage corresponds to an overall layout of the environment that allows us to derive knowledge not readily observed while interacting with the environment, *e.g.* shortcuts. We define evaluation tasks at these three levels of representation to probe if VLMs develop similar constructs. We rely on multiple-choice video QA format because modern VLMs process and generate text natively, providing a standardised and accessible interface for evaluation. Importantly, this format grants us access to the models' intermediate "thinking traces", allowing us to qualitatively investigate the actual strategies they use to solve spatial problems. In contrast, attempting to explicitly extract continuous 3D spatial representations like metric depth maps, 3D point clouds, or dense topological graphs, directly from the latent spaces of general-purpose VLMs, would be highly non-trivial and model specific.

To keep the evaluation efficient, we define the tasks on video segments of up to 10 minutes long. At average walking speed, a 10-minute segment corresponds to about 1km distance traversed; see examples of corresponding map traces in Fig. 5. We include up to 150 QAs for each type of task, again for efficiency reasons. But note that given our pipeline to ground videos on the map, we can efficiently generate a larger number of questions for each task (with the exception of loop closure, as loops occur less frequently in walking tours). Fig. 2 shows the distribution of tasks and video lengths in the benchmark. When compared to prior work, *e.g.* VSI-Bench, our videos are 3-6× longer, covering much larger spatial areas (factory *vs.* neighbourhood).

4.1 Landmarks

Landmark recognition: We define landmark recognition tasks by using a non-standard multiple-choice video QA format. The model is presented with a text question, a video clip, and a picture of a landmark that may or may not have appeared on the path shown in the clip. The questions are of the form:
Does the image show a location in the video, and if so, what is the straight-line distance between the camera location of the image and the camera location of the final video frame?

This is followed by four possible distances in meters (one being the correct one if the landmark is in the given video), plus a fifth *negative* option: *That*

Fig. 3: Examples of landmark frames (Left: Amsterdam, right: Venice).

place was not seen in the video. For simplicity, we select landmark images from the video frames themselves, by filtering for video frames that have very high VPS confidence scores ($> .98$), under the intuition that these are the most distinctive; we also apply non-maximal suppression to keep only the peaks and remove duplicates. Examples of landmark pictures obtained after this filtering are included in Fig. 3, for Walking Tour videos filmed in Amsterdam and Venice, respectively. Fig. 4 shows the distribution of correct answers for landmark distances. We also include questions for which the negative answer is the correct one. In these cases, we extract landmarks from parts of the video that do not overlap with the current segment. Note that this landmark recognition task is an example of needle-in-haystack problem, at which Transformer-like architectures are known to excel [49], and our results confirm this (see Sec. 5.3 and analysis in Fig. 9, right).

Distance to landmark: Note that the second part of the question above related to distance estimation requires map-level knowledge. We use this composite form for efficiency reasons, to run a single model query. In the analysis of the results, we break down the performance into landmark recognition and distance estimation, see Fig. 9, right.

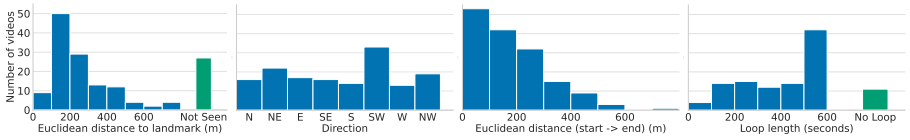


Fig. 4: Left: Distribution of line-of-sight distances (in metres) to landmarks used in our dataset. We also include questions with landmarks that haven't been visited on the tour, for which the correct answer is 'not seen'. Centre-left: Distribution of orientations at the end of the walking tour for the compass task. Centre-right: Distribution of Euclidean distances between the start and the end of each tour. Right: Distribution of loop lengths in seconds; maximum video length is 600 seconds (10 minutes). We also include 11 question-answer pairs where there is no loop.

4.2 Routes

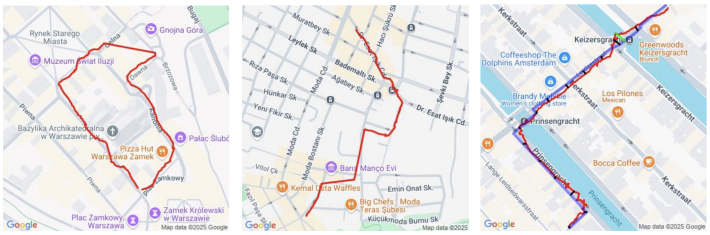
Compass: We formulate compass questions related to global camera orientation by giving the orientation of the camera at the beginning of the clip and asking the model to output the orientation at the end of the clip. These allow us to probe the model's ability to perform simplified path integration, necessary for understanding routes and building mental maps of large-scale environments [38]. To correctly answer this question, the model could keep track of point correspondences across frames and estimate the camera motion. Alternatively, the model could infer the cardinal orientation using visual cues, *e.g.* the sun position. However, such cues are often not available (*e.g.* if the person walks on a narrow street and the sun position cannot be inferred), so the camera motion estimation is a more reliable strategy. We use the following question prompt: *For this question, you need to keep track of the direction that the camera is facing throughout the given walking tour video. Given that the camera is pointing towards the* `<orientation_start>` *at the start, where is the camera pointing at the very end?*

The orientation is binned into one of eight possibilities: NORTH, NORTHEAST, EAST, SOUTHEAST, SOUTH, SOUTHWEST, WEST, NORTHWEST. The ground-truth end orientation and the start orientation are extracted automatically from the VPS predictions after outlier removal. Rotation quaternions in the ECEF reference frame are mapped to the local tangent plane to calculate the yaw with respect to the direction pointing towards north at that location. The 4 incorrect options in the multiple-choice task are sampled at random. Clips of up to 10-minutes duration are sampled at random such that the start and end frames have high-confidence VPS predictions. We start by extracting a large number of possible questions, and we filter down to keep a subset of 150 QAs by sampling at random questions with distinct start and end orientations; see the distribution of orientations in Fig. 4, centre-left.

Loop closure detection: This is a critical task in large-scale spatial understanding, commonly used by simultaneous localisation and mapping systems (SLAM) [11,35] to correct navigation drift. SLAM solutions rely on a combination of visual recognition and geometric path integration strategies. We formulate loop closure detection questions in multiple-choice format, bridging in this way the gap between the VLM and SLAM communities.

Given our videos grounded on the map, we extract candidate loops by running a simple heuristic over the VPS predictions, checking if the current VPS point has been visited previously within a 4m radius. Then, we conduct manual inspection of the candidate loops with human raters to remove false positives caused by possible VPS outliers. We use the following question prompt in our benchmark: *Has the location at the end of the video been visited at an earlier point in the video? If so, when?*

The negative options are sampled at random making sure that they are at least 1 minute apart from the correct option. Note that we also sample *negative* questions, *i.e.* questions for which the correct answer is *The location has not been visited before in this video.* Figure 5 shows trace examples of positive and negative



(a) (b) (c)

Fig. 5: (a) Example of a VPS trace with a loop. (b) Example of a VPS trace without a loop. (c) Given the VPS coordinates for a video segment (shown in red), we select high-confidence intermediate VPS points to query the Google Maps Compute Routes API and get a closely-matching trace (shown in blue). Black segments indicate the optimal pairings found by DTW (Dynamic Time Warping) metric. In green, we show the maximum distance (error).

scenarios and Fig. 4 shows the distribution of loop lengths in our dataset. We opted to use this composite formulation that includes the time recall to be able to generate five meaningful answer options instead of the binary Yes / No. We break down the performance into loop closure detection and time recall in Fig. 9, right, for analysis purposes.

Route summary: We define route summary tasks to probe if the model can identify a summary of the path shown in the video in terms of navigation actions (turn left, turn right, go straight). The question prompt is: *Which of these sets of directions correctly represent the path of the person in this video?*

To generate the ground truth answers, we rely on Google Maps Routes API that we query using VPS coordinates for start, end, and intermediate points. To ensure that the generated route closely matches the actual path followed in the video (and indicated by VPS), we sample multiple routes using Routes API using different intermediate points and then we use again DTW distance (mentioned in sec 3.2) to select the best match with the VPS trace; see Fig. 5 (c). If the maximum error between optimal pairs found by DTW is not below a given threshold, we discard the question. The pipeline allows us to generate and check a large number of questions very efficiently, so we can afford to select only the questions that satisfy our error constraints. Finally, we keep as correct answer the text description returned by the Routes API for this best match[7].

To generate the negative options, we consider partially-overlapping segments, forcing the model to correctly identify the start and end point of the tour to discard these. Another negative option that we use is the reversed version of the ground truth path. Fig. 6 (top) shows some possible options, which are also used for the map trace task explained below.

[7] We also experimented with fitting a polyline to the VPS predicted locations and deriving the sequence of turns from it, but this led to noisier redundant navigation instructions.

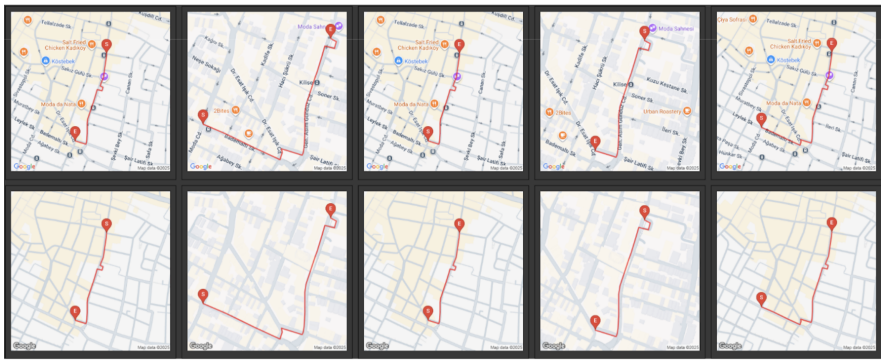


Fig. 6: Top: Map trace images used as options for the Map trace task. **Bottom**: A more challenging map trace task variation where the route traces are in-painted over maps without text labels or landmark icons.

As an ablation, we also experimented with an easier version of the task, where we sample negative options from video segments that do not overlap with the ground truth segment at all; see appendix for more details.

4.3 Survey Knowledge (Maps)

Map trace (with text): At the ultimate MAP representation level, we aim to probe if the mental map inferred by the model for the given walking tour is valid. To this end, we formulate a first task that requires the model to identify the correct map trace of the tour from a set of five possible map traces, see Fig. 6 top. For this setting, we rely on a non-standard video QA configuration, where the question is given as text, and the options are provided as map images (with text labels and landmark icons overlaid on the map) alongside the input video. The question is of the form: *Which of these maps correctly represent the path of the person in this video?*

To generate the correct and the negative options for this task, we use the same pipeline as for the route summary task, but we plot visually on the map the answers from the Google Maps Routes API. Similarly to the route summary task, we set up the same ablation with the easier version of negatives; see appendix.

Map trace (w/o text): To further challenge the video understanding capabilities of the evaluated VLMs, we add a version of the task where the map trace is overlaid on a bare map, without text labels or landmark icons; see Fig. 6 bottom.

Euclidean distance: The cognitive literature defines as the ultimate test for probing the validity of the mental map in large-scale environments the capability of estimating the Euclidean distance between two locations after exploring only non-Euclidean (longer) paths between those locations. In setups that allow interaction, the task is set up as finding a shortcut between two locations after

exploring different connecting paths [10]. In our setup, we define the task as estimating the Euclidean distance between the start and the end of a given walking tour, using the following prompt: *For this question, you need to keep track of where the user is in the real world throughout the walking tour video. Based on that, what is the line-of-sight distance between the start point and the end point of this tour?*

We extract the ground truth start and end positions using the VPS predictions and calculate the L2 distance. The four incorrect options for each question are sampled at random, making sure that they are not within 100m of the ground-truth to avoid ambiguity. Fig. 4, centre-right, shows the distribution of correct answers in the benchmark.

5 Experiments

We evaluate PLM-8B, Qwen2.5-VL-72B, the efficient Gemini model, 2.5 Flash, and the strongest models from the GPT and Claude families, GPT-5 and Claude Opus[8] respectively. We also collected a human baseline and we ran a *blind* Gemini baseline by feeding blank frames to Gemini 2.5 Flash instead of the RGB frames; see Fig. 7. It is important to note that the blind baseline has results at chance-level, confirming that our tasks cannot be answered from text alone.

5.1 Human Baseline

We collected a small human baseline on a subset of the video QAs. We randomly sampled 47-49 questions from each category, for a total of 337 QAs. We recruited 18 crowd-sourced participants (male and female, with advanced English skills) and we collected 10 answers per question, from 10 different participants. Each participant answered questions from between 3 and 7 categories. The overall score for this baseline was 71.3%; see Fig. 7 for accuracy across tasks, with an inter-rater agreement of 0.76 across the dataset. Most of the errors were encountered in the loop closure task, where the participants often mistakenly answered that there was no loop in the video. This is the only task where the human baseline was outperformed by two of the evaluated VLMs. Humans watched the videos at 30 FPS.

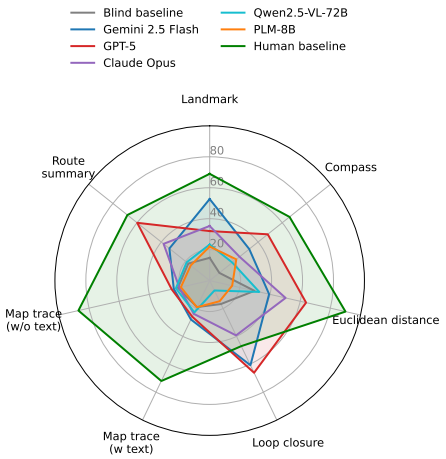


Fig. 7: Performance of different VLMs on KilometerVision, compared to human performance and a blind baseline.

[8] We use the model with 64k thinking tokens. The model without thinking obtains slightly lower performance (35.1% vs 34.9%).

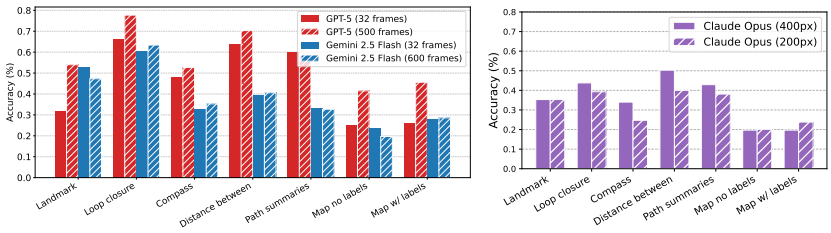


Fig. 8: Impact of frame rate (left) and of spatial resolution (right) on performance for different models.

5.2 Results and Ablations

In our experiments, we found that PLM-8B and Qwen2.5-VL-72B can only process up to 32 frames, and Claude Opus API failed repeatedly on more than 32 frames. For fair comparison, we ran all models by uniformly sampling 32 frames for each question. Fig. 7 summarises the results. Overall, GPT-5 obtains the strongest performance, excelling at loop closure and route summary, but being significantly below human performance at all other tasks, especially tasks involving maps. PLM-8B and Qwen2.5-VL-72B show limited understanding capabilities, being on par with the blind baseline.

Impact of frame rate and spatial resolution: To decouple intrinsic spatial capabilities from the memory limitations of each model, we also evaluated Gemini 2.5 Flash and GPT-5 (the only models that can run reliably with more than 32 frames per video) at their maximum temporal capacity to reflect the current SOTA upper bound and have a fairer comparison with the human baseline. We include results in Fig. 8, left. Gemini 2.5 Flash shows almost no improvement when increasing the number of frames (32→600; ∼38% accuracy), while GPT-5 improves significantly from 45% to 57%. This suggests that only larger-scale models are capable of exploiting additional frames due to higher-level video understanding capabilities. To understand how spatial resolution impacts performance, we ran an ablation with Claude Opus downsampling the frames to 200 pixels on the smaller side (from 400 pixels used in the main experiment). The performance mainly degrades across tasks, with the exception of map related tasks, where the accuracy stays at chance-level irrespective of the spatial resolution.

Passive visual observations vs. *embodied perception:* The cognitive literature outlines that humans have difficulties performing large-scale spatial tasks in passive setups without access to proprioception [10]. In an attempt to mimic additional modalities (*e.g.* the vestibular signal), we run ablations with Gemini 2.5 Flash with additional privileged information to see if it can leverage them in a zero-shot regime. First, we investigate if providing camera pose and absolute VPS coordinates help with map trace recognition tasks (Fig 9, left). Then, we investigate if providing maps helps with the Route summary task (Table 1). Finally, we check if providing the ground truth route summary helps with map trace recog-

nition tasks, given that these models are experts at processing text information (Table 2). We can observe that providing camera pose or VPS traces does not help significantly, suggesting that fine-tuning might be needed for the model to learn to leverage them. Providing maps with text labels helps in summarising the routes, but maps without labels actually hurt performance. Similarly, providing the route text summary for map recognition tasks helps, but only when the maps have text labels overlaid. Interestingly, performance increases for these tasks when we remove the video completely and provide only the map with text labels or the text route summaries, suggesting that the model is able to read the map to some extent when it is allowed to allocate all its attention to it, instead of trying to correlate the map information with visual cues in the video.
LLM Thinking ablation: We enable the thinking variant of models where possible, but in practice these models think regardless. With thinking not enabled and the prompt strongly asking the model to answer with just the option number, both Claude Opus and Gemini Flash presented a detailed analysis of the question and options similar to those shown in Fig. 14-16 (Appendix). Thus a strict ablation of this is currently challenging. Nonetheless, we observe the following absolute accuracy changes when attempting to disable thinking in Gemini 2.5 Flash: Euclidean Distance (+15.5%), Loop closure (-3.6%), Route Summary (-4.8%), Compass (-10%). There is no clear trend as the model is often thinking regardless of not being prompted to do so.

5.3 Discussion

In biological intelligence, spatial awareness develops hierarchically: anchoring via visual landmarks, connecting them via egocentric routes (path integration), and ultimately forming an allocentric, geometric map (survey knowledge). Our experiments reveal a fundamental divergence in how current AI models process spatial information.
a) The Landmark stage: visual recognition vs. *spatial grounding.* VLMs excel at visual recognition of landmarks in long videos, as can be observed in Fig. 9 right, where we separate performance for Landmark and Loop closure tasks into visual recognition (*i.e.* determining the presence or absence of a landmark or loop) and selecting the correct distance/time option. However, they acuity lack spatial grounding, given poor performance in estimating the Euclidean distance to the landmark or estimating when the loop started. Model traces (examples included in appendix) reveal an inability to infer 3D depth or scale from 2D pixels; they recognise what a landmark is, but not where it is relative to the observer.
b) The Route stage: semantic dependency over path integration. Biological route knowledge relies on path integration: the continuous tracking of movement and rotation. Our compass, loop closure, and route summary tasks demonstrate that VLMs fail at this physical tracking. When navigating cumulative angle changes, their geometric reasoning collapses. They attempt to recognise turn-left or turn-right actions (*e.g.* for compass), but they fail most of the time at estimating the angles of rotations. This could also be due to the low frame rate that they can

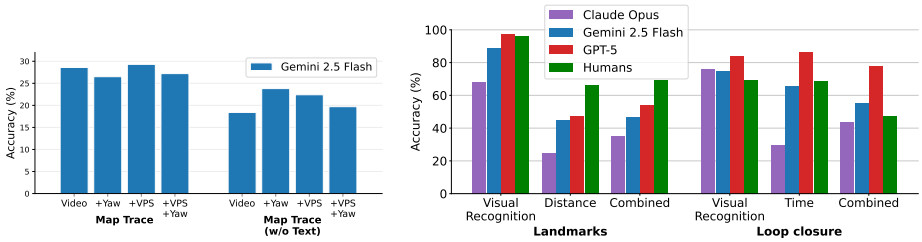


Fig. 9: Left: Gemini 2.5 Flash performance on Map Trace questions with additional camera and/or VPS information. Right: Performance on Landmark and Loop Closure questions separated into Visual Recognition (*i.e.* determining the presence or absence of a landmark or loop) and selecting the correct distance / time option, compared to overall (combined) performance.

	Video	No video
No map	42.2	–
Map trace w/ text	49.0	53.1
Map trace w/o text	37.4	53.1

Table 1: Ablation for the route summary task (top-1 accuracy).

	Video	No video
Map trace w/ text	28.6	–
+ route summary	34.7	46.3
Map trace w/o text	28.6	–
+ route summary	26.5	38.1

Table 2: Ablation for the map trace recognition (top-1 accuracy).

process. Instead, models rely almost exclusively on reading semantic cues (*e.g.* street signs or text labels on the map, see ablations in Tables 1 and 2), effectively substituting reading comprehension for actual geometric path integration.

c) The Map stage: illusion of survey knowledge. Mental maps allow humans to infer allocentric relationships, such as straight-line distances. Our Euclidean distance task and map trace recognition task reveal that VLMs possess only an illusion of this survey knowledge. The models default to text-matching street names (ablations in Tables 1 and 2), confirming that VLMs rely on OCR rather than matching geometric trajectories to map topology. It is worth noting that we tried to create similar visualisations of mental maps as in [63], by defining a grid and asking the model to indicate the location of different landmarks, but preliminary results were not satisfactory.

Loop Closure: bypassing the mental map. Loop closure is the only task where two of the evaluated VLMs achieve super-human performance. However, model traces show that they bypass route or spatial map formation entirely, relying instead on exhaustive, brute-force visual frame matching. While effective, this approach is brittle. Latching onto misleading repeated visual cues can cause false positives (*e.g.* the statue in Fig. 10 for Gemini 2.5 Flash), further highlighting the absence of underlying geometric awareness. More examples of thinking traces are included in the appendix.

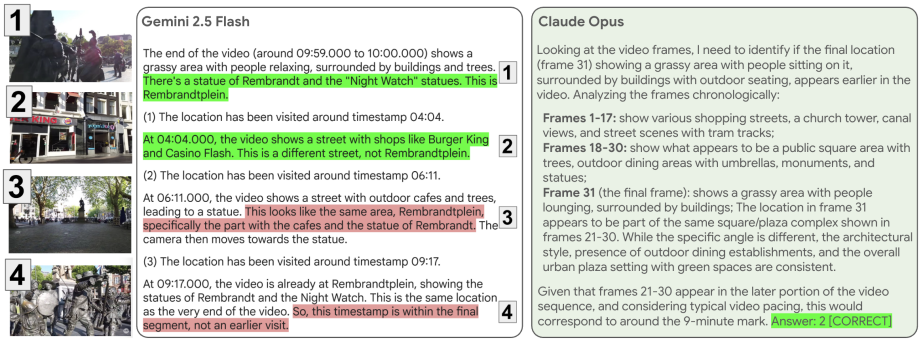


Fig. 10: Example of thinking traces from Gemini 2.5 Flash and Claude Opus on loop closure questions. Both models rely purely on visual recognition to solve the task. Gemini correctly identifies the location at the end of the video 1, and correctly eliminates another option 2. The model incorrectly believes that a different statue that it can see in the distance is the same as the one it sees at the end of the video 3. The model correctly determines that the camera is in the same location at time 9:17 but incorrectly decides that this was not visited earlier 4. **Claude Opus**'s reasoning is generally correct and it correctly identifies option ID 2 as the right answer.

6 Conclusion

We introduce *KilometerVision*, the first benchmark that comprehensively evaluates city-scale geographical understanding in large multimodal models (VLMs) using real-world videos. Evaluated against the human LANDMARK-ROUTE-MAPS paradigm, we find that current VLMs operate at the Landmark stage. While they can leverage robust semantic matching and visual recognition to simulate spatial awareness, such as reading street signs to "navigate" or brute-forcing frame matches to detect loop closures, they fundamentally fail at path integration and inferring allocentric survey knowledge. When provided with privileged information in context (GPS, maps, or camera poses), they cannot leverage it, suggesting that such capabilities are not yet mature in state-of-the-art models. We hope that our benchmark and analyses will help the community to improve spatial capabilities in VLMs and build reliable AI assistants and embodied AI.

Acknowledgements

We are very grateful to Andrew Zisserman, Rick Szeliski, and Mehdi S.M. Sajjadi for their guidance and insightful input in shaping the project, and Dilara Gokay for input on Youtube video selection.

References

1. Akesson, S., Boström, J., Liedvogel, M., Muheim, R.: Animal Navigation, pp. 151–178. Oxford University Press (08 2014). https://doi.org/10.1093/acprof:oso/9780199677184.003.0009
2. Andersen, P., Morris, R., Amaral, D., Bliss, T., O'Keefe, J.: The Hippocampus Book. Oxford University Press (12 2006). https://doi.org/10.1093/acprof:oso/9780195100273.001.0001, https://doi.org/10.1093/acprof:oso/9780195100273.001.0001
3. Anthropic: The Claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (2024, date accessed: 2026-06-29)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970 (2015)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
7. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: CVPR. pp. 12530–12539 (2019). https://doi.org/10.1109/CVPR.2019.01282
8. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024), https://arxiv.org/abs/2406.07476
9. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., Martin, M., Wang, H., Rasheed, H., Sun, P., Huang, P.Y., Bolya, D., Ravi, N., Jain, S., Stark, T., Moon, S., Damavandi, B., Lee, V., Westbury, A., Khan, S., Krähenbühl, P., Dollár, P., Torresani, L., Grauman, K., Feichtenhofer, C.: Perceptionlm: Open-access data and models for detailed visual understanding (2025), https://arxiv.org/abs/2504.13180
10. Chrastil, E., Warren, W.: Active and passive spatial learning in human navigation: Acquisition of graph knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition **41**, 1162–1178 (11 2014). https://doi.org/10.1037/xlm0000082
11. Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research **27**(6), 647–665 (2008). https://doi.org/10.1177/0278364908090961, https://doi.org/10.1177/0278364908090961
12. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. IJCV **130**, 33–55 (2022)
13. Floros, G., van der Zander, B., Leibe, B.: Openstreetslam: Global vehicle localization using openstreetmaps. In: 2013 IEEE International Conference on Robotics and Automation. pp. 1054–1059 (2013). https://doi.org/10.1109/ICRA.2013.6630703

14. Fouhey, D.F., Kuo, W.c., Efros, A.A., Malik, J.: From lifestyle vlogs to everyday interactions. In: CVPR. pp. 4991–5000 (2018)
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR. pp. 24108–24118 (2025)
16. Gardner, H.: Frames of Mind: The Theory of Multiple Intelligences. London: Heinemann (1983)
17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
18. Girdhar, R., Ramanan, D.: CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In: ICLR (2020)
19. Gould, J.L., Gould, C.G.: Nature's Compass: The Mystery of Animal Navigation. Princeton University Press (2012)
20. Grauman, K., Westbury, A., Byrne, E., Chavis, Z.Q., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Fuegen, C., Gebreselasie, A., González, C., Hillis, J.M., Huang, X., Huang, Y., Jia, W., Khoo, W.Y.H., Kolár, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K.K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhu, Y., Arbeláez, P., Crandall, D.J., Damen, D., Farinella, G.M., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R.A., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3, 000 hours of egocentric video. In: CVPR (2022)
21. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR. pp. 6047–6056 (2018)
22. Handa, A., Whelan, T., McDonald, J., Davison, A.J.: A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 1524–1531 (2014). https://doi.org/10.1109/ICRA.2014.6907054
23. Hodaň, T., Michel, F., Brachmann, E., Kehl, W., Buch, A.G., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T.K., Matas, J., Rother, C.: Bop: Benchmark for 6d object pose estimation. In: ECCV. p. 19–35. Springer-Verlag (2018)
24. Horvath, V., Toth, C., Barsi, A.: Investigating the accuracy of google's visual positioning system. In: 2025 IEEE/ION Position, Location and Navigation Symposium (PLANS). pp. 1594–1600. IEEE (2025)
25. Huang, J., tse Huang, J., Liu, Z., Liu, X., Wang, W., Zhao, J.: Vlms as geoguessr masters: Exceptional performance, hidden biases, and privacy risks. CoRR **abs/2502.11163** (February 2025), https://doi.org/10.48550/arXiv.2502.11163
26. Kim, K., Bock, O.: Acquisition of landmark, route, and survey knowledge in a wayfinding task: in stages or in parallel? Psychological Research **85**, 2098–2106 (2021)
27. Ko, D., Lee, J.S., Kang, W., Roh, B., Kim, H.J.: Large language models are temporal and causal reasoners for video question answering. In: EMNLP (2023)

28. Krishnan, A., Liu, S., Sarlin, P.E., Gentilhomme, O., Caruso, D., Monge, M., Newcombe, R., Engel, J., Pollefeys, M.: Benchmarking egocentric visual-inertial slam at city scale. In: ICCV (2025)
29. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-OneVision: Easy visual task transfer. TMLR (2025)
30. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
31. Li, L., Bigverdi, M., Gu, J., Ma, Z., Yang, Y., Li, Z., Choi, Y., Krishna, R.: Unfolding spatial cognition: Evaluating multimodal models on visual simulations (2025), https://arxiv.org/abs/2506.04633
32. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-ChatGPT: Towards detailed video understanding via large vision and language models. In: ACL (2024)
33. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. arXiv preprint arXiv:2308.09126 (2023)
34. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
35. Newman, P., Ho, K.: Slam-loop closing with visually salient features. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation. pp. 635–642 (2005). https://doi.org/10.1109/ROBOT.2005.1570189
36. OpenAI: GPT-5 is here. https://openai.com/gpt-5 (2025, date accessed: 2026-06-29)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
38. Palgi, S., Ray, S., Maimon, S., Waserman, Y., Ben-Ari, L., Eliav, T., Tuval, A., Cohen, C., Ali, A., Keyyu, J., Mouritsen, H., Las, L., Ulanovsky, N.: Head-direction cells as a neural compass in bats navigating outdoors on a remote oceanic island. Science **390**(6770) (Oct 2025). https://doi.org/10.1126/science.adw6202
39. Pătrăucean, V., Smaira, L., Gupta, A., Continente, A.R., Markeeva, L., Banarse, D., Koppula, S., Heyward, J., Malinowski, M., Yang, Y., Doersch, C., Matejovicova, T., Sulsky, Y., Miech, A., Frechette, A., Klimczak, H., Koster, R., Zhang, J., Winkler, S., Aytar, Y., Osindero, S., Damen, D., Zisserman, A., Carreira, J.: Perception test: A diagnostic benchmark for multimodal video models. In: NeurIPS (2023), https://openreview.net/forum?id=HYEGXFnPoq
40. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? In: ICLR (2025), https://openreview.net/forum?id=WK6K1FMEQ1
41. Rawal, R., Saifullah, K., Farré, M., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813 (2024)
42. Roberts, J., Lüddecke, T., Das, S., Han, K., Albanie, S.: Gpt4geo: How a language model sees the world's geography. NeurIPS Workshop on Foundation Models for Decision Making (May 2023)
43. Roberts, J., Lüddecke, T., Sheikh, R., Han, K., Albanie, S.: Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In: CVPR2024 Workshops. pp. 554–563 (06 2024). https://doi.org/10.1109/CVPRW63382.2024.00060

44. Siegel, A.W., White, S.H.: The development of spatial representations of large-scale environments. In: Reese, H.W. (ed.) Advances in Child Development and Behavior, vol. 10, pp. 9–55. JAI (1975). https://doi.org/https://doi.org/10.1016/S0065-2407(08)60007-5, https://www.sciencedirect.com/science/article/pii/S0065240708600075
45. Simmons, J.A.: The resolution of target range by echolocating bats. The Journal of the Acoustical Society of America **54**(1), 157–173 (1973)
46. Song, X., Chen, W., Liu, Y., Chen, W., Li, G., Lin, L.: Towards long-horizon vision-language navigation: Platform, benchmark and method. In: CVPR (2025)
47. SpiegelShai, E., Peer, S., Shvartzman, B., Demri, A., Avni, O.: Visual positioning system. U.S. Patent Application US20200401617A1 (2019)
48. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR. pp. 4631–4640 (2016)
49. Team, G.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
50. Tolman, E.C.: Cognitive maps in rats and men. Psychological review **55 4**, 189–208 (1948), https://api.semanticscholar.org/CorpusID:42496633
51. Vafa, K., Chen, J.Y., Rambachan, A., Kleinberg, J., Mullainathan, S.: Evaluating the world model implicit in a generative model. In: NeurIPS (2024), https://openreview.net/forum?id=aVK4JFpegy
52. Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., Kambhampati, S.: Planbench: an extensible benchmark for evaluating large language models on planning and reasoning about change. In: NeurIPS. Curran Associates Inc. (2023)
53. Venkataramanan, S., Rizve, M.N., Carreira, J., Asano, Y.M., Avrithis, Y.: Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In: ICLR (2024)
54. Wang, S., Zhao, Q., Do, M.Q., Agarwal, N., Lee, K., Sun, C.: Vamos: Versatile action models for video understanding. In: ECCV (2024)
55. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: ICCV. pp. 22958–22967 (2025)
56. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: InternVideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
57. Wang, Y., Yang, Y., Ren, M.: LifelongMemory: Leveraging llms for answering queries in long-form egocentric videos. arXiv preprint arXiv:2312.05269 (2024)
58. Wiener, J.M., Buchner, S.J., Holscher, C.: Taxonomy of human wayfinding tasks: A knowledge-based approach. Spatial Cognition & Computation **9**, 152–165 (05 2009). https://doi.org/10.1080/13875860902906496
59. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR. pp. 9777–9786 (2021)
60. Xu, Y., Pan, Y., Liu, Z., Wang, H.: FLAME: Learning to navigate with multimodal llm in urban environments. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9005–9013 (2025)
61. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xGen-MM (BLIP-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)

62. Yang, A., Fu, C., Jia, Q., Dong, W., Ma, M., Chen, H., Yang, F., Wu, H.: Evaluating and enhancing spatial cognition abilities of large language models. International Journal of Geographical Information Science (04 2025). https://doi.org/10.1080/13658816.2025.2490701
63. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. In: CVPR (2025)
64. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. In: ICLR (2020), https://openreview.net/forum?id=HkxYzANYDB
65. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
66. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: EMNLP (2023)
67. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: CVPR. pp. 3537–3545 (2019)
68. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion. Acta Numerica **26**, 305–364 (2017). https://doi.org/10.1017/S096249291700006X