Evidence-Log Video Agents with Deterministic Geometry for Kilometre-Scale Spatial Video Question Answering

Yu-An Lu
National Yang Ming Chiao Tung University
yuan.la14@nycu.edu.tw

1 Introduction

KilometerVision [5, 9] is the spatial-reasoning track of the Fourth Perception Test Challenge. It continues the hour-long walking-tour question answering introduced in the 2024 edition [5] and run again in 2025 [6], posing multiple-choice questions about walking-tour videos [17] of up to ten minutes and roughly one kilometre of travel, across six families: map-path matching, camera-heading tracking, textual directions matching, start-to-end distance, image-to-video distance, and revisit detection. Two properties of the task shape our design. A segment is long relative to the temporal resolution at which video models attend and accuracy falls with duration [3], so the evidence for one answer spreads over hundreds of seconds the model must address individually. And several families reward exact quantities such as an integrated turn angle or a metric distance, where multimodal models remain below human accuracy on spatial video question answering [19] and near chance on classic spatial tests [11], while the perception feeding them stays qualitative, as in reading a street sign or naming a city. Our system assigns each kind of work to the component suited to it: the model performs perception and world-knowledge grounding, deterministic image processing performs geometry and measurement, and the interface between them is explicit text. Figure 1 traces one question through it.

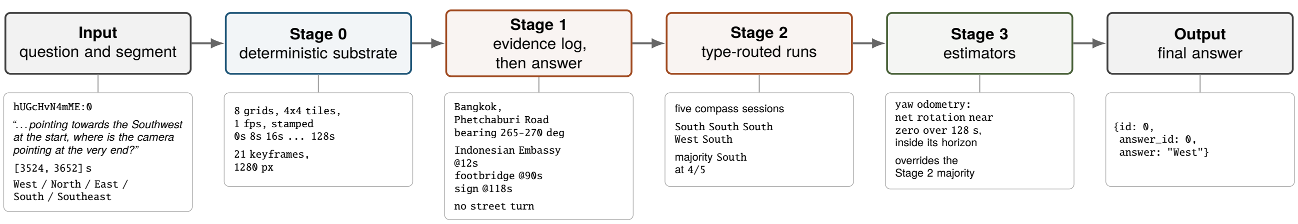

Figure 1: One question carried end to end, each box holding the values its stage produced. Blue marks deterministic components, orange vision-language turns, green measurement. The five sessions favour *South* and the Stage 3 measurement overrides that majority.

2 Method

The pipeline answers one question at a time. A question supplies a video segment delimited by a start and an end second, a question string, and five options, and every stage below works from those materials alone, together with the pretrained models and the fixed constants of Table 1. A question is solved in its own context, with nothing carried over from any other question.

Stage 0: deterministic substrate. Stage 0 turns the material of one question into text-addressable form and uses no model tokens. A rule-based classifier assigns the question to one of the six families the task defines, from its wording alone, selecting the Stage 1 playbook and the Stage 2 sampling budget. Only the footage between that question's own start and end seconds is retrieved. The segment is rendered as 4×4 frame montages [7] at 1 fps up to 360 s and 0.5 fps beyond, 480 px per tile, every tile stamped with its own second so that a time can be named from the pixels [18], plus up to 24 keyframes at 1280 px for reading street names and shopfronts. Where the options are candidate route maps, whose thin polylines are a documented failure mode for vision-language models [10], a deterministic pass recovers the route mask by colour thresholding, the two pins and their start and end identity from the glyph inside each pin head, and the ordered polyline between them, emitted as a sequence of eight-way compass legs with turn angles and attached to the prompt beside the images. A route-mask signature marks duplicate and reversed candidates among that question's own five options.

Stage 1: evidence log and answer. Stage 1 opens a fresh session for the question and closes it once the answer is written. The session first builds an evidence log from the segment in three parts: the real city and streets read from signage with the evidence cited, every heading change with a timestamp, direction, estimated magnitude and uncertainty range, and a timeline of re-identifiable landmarks. It then answers under a playbook selected by family. The compass playbook derives a heading from turn log integration, from the bearing of an identified street, and from sun and shadow azimuth, and reconciles the three. The distance playbooks pin the endpoints to named intersections and read their separation off the identified street layout, the map and textual playbooks match the turn log against the five option descriptions, and the revisit playbook enumerates the

Table 1: Configuration of the system.

Component	Setting
visual context	4×4 grids at 480 px per tile with per-tile timestamps, 1 fps up to 360 s and 0.5 fps beyond, plus up to 24 keyframes at 1280 px spaced at least 5 s apart
map vectorisation	route mask by colour threshold, S and E pin glyph, eight-way leg transcript, normalised route-mask signature
backbone and tool	Gemini 3.7 Flash [4], zero shot, fixed family-specific prompt templates, 8192 max output tokens, `extract_frame(t)` as the only tool, $\leq$ 12 rounds
runs per question	compass 5, map and textual 2 plus an adjudication turn, other families 1, one predefined evidence-scan strategy per session
yaw odometry	Lucas-Kanade on Shi-Tomasi features, 8 fps at 384×216, median horizontal displacement, scale normalised inside each question's own segment
place recognition	MegaLoc at 1 fps, 322×322 input, 8448-d L_2-normalised descriptors
verification	Claude Fable 5, answer-blind, no browsing or external retrieval, one agent per question
setup	no model training or fine-tuning, zero shot on the challenge splits

candidate timestamps. Throughout, the session may call `extract_frame(t)`, returning the full-resolution frame at second t of that segment, for up to twelve rounds, following the tool-using pattern of recent long-video agents [20].

Stage 2: type-routed sampling. Stage 2 repeats Stage 1 on the same question and takes the majority. Because the return on repeated sampling depends on the question [15], the budget is set by family: five independent sessions for compass, two for map and textual directions, one elsewhere, each session drawing on a different one of five predefined evidence-scan strategies. Where map or textual sessions disagree, an adjudication turn chooses between the two leading options, and because pairwise judging alone is unreliable [16] it receives the deterministic Stage 0 route transcripts alongside them.

Stage 3: geometric and retrieval estimators. Stage 3 computes physical quantities from the segment and supplies them as measurements with stated uncertainty. Net heading change comes from integrating the median horizontal displacement of sparse optical flow [8, 14] at 8 fps on 384×216 frames, a statistic that responds to rotation and stays near zero under forward translation [12]. Its pixel to degree scale is an ephemeral per-instance geometric estimate computed inside that same segment, by snapping the turn magnitudes detected there onto multiples of $45°$. It is discarded once the question is answered and is withheld where a segment holds too few turns to determine it. The result is combined with the starting heading the question itself states, within an integration horizon fixed in advance. Place identity comes from MegaLoc embeddings [1, 13] at 1 fps over the same segment. A revisit question scores its candidate timestamps against the closing frames, and an image-to-video question scores the supplied photograph against the segment frame by frame, with the located frame then pinned like any other endpoint so that its distance to the final position follows. High similarity is admitted as evidence of a match and low similarity is recorded as inconclusive.

Stage 4: verification. Without a verifier, more samples do not convert into more correct selections [2], so Stage 4 re-solves the questions where Stage 2 leaves no majority, using a second reasoning model driven through an agentic command-line harness, one agent per question. That agent sees the frame grids and options for that question alone, without the Stage 2 answer, and returns a choice, a confidence level, and the evidence behind it. The channel is admitted only after scoring on synthetic questions built from known camera trajectories, where an applied yaw fixes the compass answer and a route rendered from the trajectory, with perturbed distractors, fixes the map and textual answers, on the subset where it disagrees with Stage 2.

The final answer then follows a fixed order. The Stage 2 majority stands unless an admitted Stage 3 measurement contradicts it, a map or textual tie goes to the adjudication turn, and any remaining question without a majority goes to Stage 4. The family-specific confidence gates behind each of these steps are set in advance and applied deterministically.

Configuration and data usage. Table 1 lists the settings. Gemini 3.7 Flash [4] and Claude Fable 5 are public pretrained API models and MegaLoc is a public pretrained checkpoint, all used as released. The system is trained on no data of ours and evaluated zero shot.

References

[1] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2852–2858. IEEE, 2025.

[2] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[3] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, et al. Video-MME:

The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025.

[4] Google DeepMind. Gemini 3.7 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-7-Flash-Model-Card.pdf, August 2026.

[5] Joseph Heyward, João Carreira, Dima Damen, Andrew Zisserman, and Viorica Pătrăucean. Perception test 2024: Challenge summary and a novel hour-long VideoQA benchmark. *arXiv preprint arXiv:2411.19941*, 2024.

[6] Joseph Heyward, Nikhil Parthasarathy, Tyler Zhu, Aravindh Mahendran, João Carreira, Dima Damen, Andrew Zisserman, and Viorica Pătrăucean. Perception test 2025: Challenge summary and a unified VQA extension. *arXiv preprint arXiv:2601.06287*, 2026.

[7] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a VLM. *IEEE Access*, 12:193057–193075, 2024.

[8] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th International Joint Conference on Artificial Intelligence*, volume 2, pages 674–679, 1981.

[9] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

[10] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 293–309, 2024.

[11] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations (ICLR)*, 2025.

[12] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry [tutorial]. Part I: The first 30 years and fundamentals. *IEEE Robotics & Automation Magazine*, 18(4):80–92, 2011.

[13] Stefan Schubert, Peer Neubert, Sourav Garg, Michael Milford, and Tobias Fischer. Visual place recognition: A tutorial. *IEEE Robotics & Automation Magazine*, 31(3):139–153, 2024.

[14] Jianbo Shi and Carlo Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600. IEEE, 1994.

[15] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations (ICLR)*, 2025.

[16] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, 2025.

[17] Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M. Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *International Conference on Learning Representations*, volume 2024, pages 31952–31973, 2024.

[18] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13754–13765, 2025.

[19] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643, 2025.

[20] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.