Metric VGGT with GPT-5.6 Sol for Kilometer-Scale Spatial Video Question Answering

Yuheng Zhou, Gabriel Turbinca, Jialei Xu, Zizhuang Wei
Team Huawei_SpaceMind

Abstract

We present a zero-shot system for KilometerVision, a five-choice spatial question-answering task over walking videos of up to ten minutes. GPT-5.6 Sol interprets each question, identifies relevant landmarks, and selects temporal evidence. In parallel, VGGT estimates camera motion and relative depth in overlapping windows, and MetricAnything converts each reconstruction to metric scale. The windows are aligned in $SE(3)$ to form a global walking path. Question-specific solvers use this path to estimate compass direction, travel distance, and route shape. Sol answers route-sequence and revisit questions and provides the answer when route matching is uncertain. Without task-specific training, the system obtains 72.312% top-1 accuracy (713/986) on the official Test set, improving the semantic-plus-compass baseline by 8.316 percentage points.

1 Introduction

KilometerVision is Task 4 of the fourth Perception Test Challenge [1]. It extends the Perception Test benchmark [2] to city-scale walking tours and evaluates metric displacement, accumulated heading, route shape, landmark-conditioned distance, route direction, and revisit time. The official Test set contains 986 five-choice questions over 54 source videos. Evaluated intervals span 6–600 s, with a median duration of 204 s. Reliable answers therefore require both semantic understanding of long videos and a consistent geometric representation of camera motion.

Our method combines semantic video reasoning with metric geometry. GPT-5.6 Sol [3] interprets the question, examines globally sampled frames, and requests denser sampling from relevant temporal segments. VGGT [4] and MetricAnything [5] recover a metric camera path from overlapping windows. The question category determines whether the answer is computed from heading, distance, route shape, or Sol's semantic and temporal evidence. All components are used without KilometerVision training labels or task-specific fine-tuning.

2 Method

2.1 Overall method

Given a video interval V, question q, and five candidate options $A = \{a_1, \ldots, a_5\}$, the system first identifies the question category. Sol processes sampled video frames and returns a semantic answer y_s, a landmark-presence decision, and, when required, a landmark time t_l. The geometric component estimates camera pose and relative depth in overlapping windows, transfers each reconstruction to metric scale, and joins the windows into a planar path $P = \{p_t\}$. A solver selected for the question category converts this evidence into the requested answer.

The final option is selected by

$$\hat{y} = \begin{cases} y_c(P), & \text{compass,} \\ y_d(P), & \text{start/end distance,} \\ y_l(P, t_l), & \text{landmark distance and present,} \\ y_m(P, A), & \text{Map Trace with } c_m = 1, \\ y_s, & \text{otherwise,} \end{cases} \tag{1}$$

Here, $c_m = 1$ indicates that route extraction succeeds and the map-matching scores satisfy both confidence conditions defined in Sec. 2.4. Compass and start/end-distance questions use geometric estimates. Landmark distance additionally uses Sol's landmark-presence decision and t_l. Map Trace uses the geometric result when $c_m = 1$ and otherwise uses y_s. Route-direction-sequence and revisit-time questions are answered by Sol.

2.2 Semantic video reasoning

For each question, Sol receives the corresponding video interval, question, and options. We use `gpt-5.6-sol` at `xhigh` reasoning effort, temperature 0, and a 2,048-token output limit. Global sampling provides 32–64 JPEG frames with a maximum side length of 896 pixels. Sol may request up to two additional temporal segments, each containing at most 48 frames, to examine brief landmark appearances, turns, or revisits. It returns y_s together with the landmark-presence decision and t_l when needed. On the Test set, Sol processes 840 non-compass questions, of which 609 use additional temporal sampling.

2.3 Metric camera path estimation

VGGT predicts intrinsics, extrinsics, relative depth, confidence, and camera positions for each window. Distance questions use 2 FPS, 10 s windows, and a 5 s stride; Map Trace uses 1 FPS, 20 s windows, and a 10 s stride. For window w, MetricAnything supplies monocular metric depth D^M. We estimate the scale of VGGT depth D^V from five evenly spaced frames F_w and their high-confidence central pixels C_i:

$$s_w = \text{median}_{i \in F_w} \left[\text{median}_{p \in C_i} \frac{D_i^M(p)}{D_i^V(p)} \right]. \tag{2}$$

Adjacent windows share three frames. Their rotations are projected to $SO(3)$, and translation is estimated from the median position difference of the shared frames. This $SE(3)$ transformation aligns each new window with the preceding path. Camera positions observed in multiple windows are averaged with larger weights near the window centers. Singular value decomposition estimates the vertical axis, after which the path is projected onto the 2-D walking plane.

2.4 Question-specific geometric solvers

For compass questions, camera-forward vectors are projected onto the ground plane. Incremental signed yaw is summed,

added to the stated initial compass direction, and matched to the nearest circular option. For start/end distance, the estimate is $d = \|p_T - p_0\|_2$. For landmark distance, Sol first decides presence; if present, geometry uses $d = \|p_T - p_{l_l}\|_2$, and if absent the semantic "not in video" option is retained. Thus 276 of 297 distance answers use metric geometry and 21 use Sol's presence decision.

For Map Trace, the red route in each candidate image is skeletonized, connected between its start and end symbols, simplified, and resampled to $N = 128$ ordered points. Let P be the centered video path and M_k the centered path of candidate k. The normalized route-matching error is

$$E_k = \min_{s>0, R\in SO(2)} \frac{\mathrm{RMSE}(sPR, M_k)}{\mathrm{RMS}(M_k)}, \tag{3}$$

where reflection and temporal reversal are not allowed. The map answer is used only when route extraction succeeds and the best two scores satisfy $E_{(2)} - E_{(1)} \geq 0.03$ and $(E_{(2)} - E_{(1)})/E_{(2)} \geq 0.15$. For example, $E_{(1)} = 0.21$ and $E_{(2)} = 0.29$ give margins of 0.08 and 0.276, so the best-matching option is selected. If either condition is not satisfied, the system uses y_s.

3 Experiment

Table 1: Official Test-set composition. All questions have five options.

Question type	Count	Share (%)
Map trace	284	28.8
Start/end distance	150	15.2
Landmark distance	147	14.9
Route direction sequence	147	14.9
Compass orientation	146	14.8
Revisit time	112	11.4
Total	986	100.0

3.1 Dataset

KilometerVision provides no training set, so all pretrained models are evaluated without task-specific fine-tuning. Prompts, sampling settings, window sizes, and score thresholds remain fixed during evaluation. Each question is processed independently, and Test annotations are used only by the official evaluator. The metric is five-choice top-1 accuracy. In addition to the 18 official validation questions, we use pedestrian sequences from ADVIO [6] to check metric distance estimation and route-shape matching.

Table 2: Component results on the official Test set. Gain is measured against the preceding row.

Configuration	Correct	Top-1 (%)	Gain (pp)
Sol + compass geometry	631/986	63.996	–
+ metric distance	663/986	67.241	+3.245
+ route-shape matching	**713/986**	**72.312**	+5.071

Table 3: Component checks on official validation and ADVIO sequences.

Set	Component	Correct	Acc. (%)
Official validation	Compass direction	5/5	100.0
Official validation	Map Trace	4/5	80.0
ADVIO (46)	Metric distance	36/46	78.3
ADVIO (32)	Route matching	31/32	96.9

3.2 Results and analysis

Table 2 shows that metric distance raises accuracy from 63.996% to 67.241%, adding 32 correct answers. Route-shape matching further raises accuracy to 72.312%, adding 50 correct answers. Compared with Sol plus compass geometry, the complete method improves top-1 accuracy by 8.316 percentage points and answers 82 additional questions correctly.

For the 284 Map Trace questions, the method evaluates 1,420 candidate routes. Route extraction succeeds for 277 questions, and 220 satisfy the two score conditions; 100 of these selected options differ from Sol's answer. Geometry determines 146 compass answers, 276 distance answers, and 220 map answers, covering 642 of 986 questions (65.1%). Sol determines the answer for 21 questions in which the landmark is absent, 64 map questions with uncertain geometric matches, 147 route-direction-sequence questions, and 112 revisit-time questions. Semantic reasoning therefore supplies landmark and temporal evidence, while geometry supplies heading, distance, and route shape.

4 Conclusion

We present a zero-shot KilometerVision method that combines GPT-5.6 Sol with a metric camera path estimated by VGGT and MetricAnything. Sol provides semantic and temporal evidence, while question-specific geometric solvers compute compass direction, distance, and route shape. The complete method reaches 72.312% top-1 accuracy on 986 Test questions without task-specific training. These results show that metric geometry complements semantic video reasoning for long walking videos. Remaining limitations include accumulated drift over long loops and uncertainty estimation for route matching.

References

[1] Perception Test Challenge Organizers, "The fourth perception test challenge at ECCV 2026." Official challenge website, 2026. Accessed 2026-08-30.

[2] V. Pătrăucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, J. Heyward, M. Malinowski, Y. Yang, C. Doersch, T. Matejovicova, Y. Sulsky, A. Miech, A. Fréchette, H. Klimczak, R. Koster, J. Zhang, S. Winkler, Y. Aytar, S. Osindero, D. Damen, A. Zisserman, and J. Carreira, "Perception test: A diagnostic benchmark for multimodal video models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[3] OpenAI, "Previewing GPT-5.6 Sol: A next-generation model." OpenAI, 2026. Accessed 2026-08-30.

[4] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual geometry grounded transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

[5] B. Ma, J. Yang, D. Di, X. Zhang, J. Cui, H. Li, Y. Xie, and W. Chen, "MetricAnything: Scaling metric depth pretraining with noisy heterogeneous sources," *arXiv preprint arXiv:2601.22054*, 2026.

[6] S. Cortés, A. Solin, E. Rahtu, and J. Kannala, "ADVIO: An authentic dataset for visual-inertial odometry," in *Proceedings of the European Conference on Computer Vision*, pp. 419–434, 2018.