

LandmarkLens: Predicting and Presenting Effective Landmarks for Mixed-Reality Urban Exploration

Chu Li
Allen School of Computer Science
University of Washington
Seattle, USA
chuchuli@cs.washington.edu

Yotam Sechayk
The University of Tokyo
Tokyo, Japan
sechayk-yotam@g.ecc.u-tokyo.ac.jp

Jared Hwang
Allen School of Computer Science
University of Washington
Seattle, USA
jaredhwa@cs.washington.edu

Jon E. Froehlich
Allen School of Computer Science
University of Washington
Seattle, USA
jonf@cs.washington.edu

Takeo Igarashi
The University of Tokyo
Tokyo, Japan
takeo@acm.org

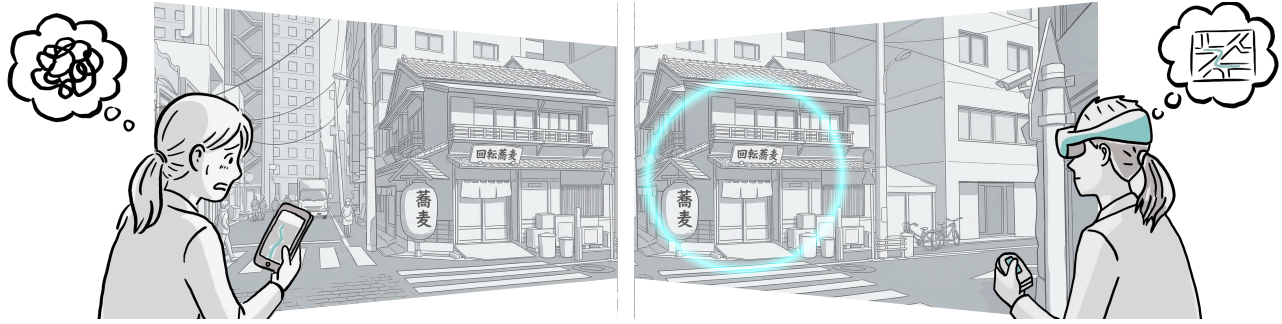


Figure 1: LandmarkLens highlights navigation-relevant landmarks in a mixed-reality street-view environment. Colored circles mark predicted landmarks; directional arrows guide users toward highlights outside their field of view.

Abstract

People with a poor sense of direction (SOD) struggle to build cognitive maps for effective spatial navigation, and existing navigation tools prioritize efficiency over spatial learning. To understand how navigation strategies differ by ability, we conducted a landmark attention study with 20 participants (ten good SOD, ten poor SOD) who navigated across four Tokyo neighborhoods in virtual reality (VR). We found systematic group differences in both gaze behavior and the types of landmarks they verbally identify as effective. Based on these findings, we built LandmarkLens, a mixed-reality (MR) navigation system that uses a vision-language model (VLM) to identify and highlight navigation-relevant landmarks. A follow-up study with eight poor-SOD participants showed improved performance in scene recognition, suggesting that guided landmark attention can support landmark-level spatial knowledge acquisition for people with poor SOD, a first step toward broader spatial learning.

CCS Concepts

• **Human-centered computing** → Mixed / augmented reality; Empirical studies in HCI; Interactive systems and tools; • **Computing methodologies** → Computer vision.

Keywords

Urban Navigation, Sense of Direction, Mixed-Reality

1 Introduction

Navigating unfamiliar urban environments is a common and important daily task, yet for people with poor sense of direction (SOD) [16, 21], it is a source of stress and anxiety [13, 21, 49]. People with poor SOD are less likely to explore unfamiliar environments [13, 21], and report greater dependence on GPS-based navigational tools [13, 17, 49]. While existing navigation tools guide users from point A to B efficiently, their turn-by-turn directions distract attention away from environmental features [9, 17, 56]. Over time, this impairs users’ spatial knowledge acquisition [6, 20, 40], and undermines the ability to navigate independently [13, 21].

Prior work has shown that the ability to navigate unfamiliar environments depends on *cognitive mapping*, the process of acquiring, representing, and using knowledge about the surrounding environment [21]. Central to this process are landmarks: objects that are stationary, distinct, and salient, serving as reference points for building mental representations of the environment [34]. Selecting effective landmarks is correlated with a person’s SOD: people with good SOD¹ select fewer but more effective landmarks [23].



This work is licensed under a Creative Commons Attribution 4.0 International License.

¹Throughout this paper, we use “good/poor sense of direction”, “good/poor navigators”, and “skilled/unskilled navigators” interchangeably to refer to participants who scored ≥ 5.0 or ≤ 3.5 on the SBSOD scale [16], respectively.

However, prior work has characterized these differences primarily through verbal self-reports [23], leaving two open questions: (1) how do the *gaze patterns* of good and poor navigators differ during navigation, and (2) can guiding poor navigators’ attention toward landmarks that skilled navigators rely on improve spatial knowledge acquisition?

To address the first question, we conducted a landmark attention study in virtual reality (VR) with 20 participants, split evenly by SOD as measured by the Santa Barbara Sense-of-Direction (SB-SOD) scale [16]. Using a custom navigation system rendering 360° street-level imagery with integrated eye tracking, participants navigated routes across four Tokyo neighborhoods while verbally describing navigation-relevant features. We collected concurrent gaze and verbal data, enabling us to analyze not only what participants *said* they noticed but also what they *actually looked at* — a combination that prior landmark selection studies have not examined. We found systematic group differences in both gaze behavior and landmark articulation. Good navigators surveyed a significantly broader horizontal extent of the scene through both wider gaze spread and greater head rotation, and their verbal descriptions disproportionately referenced durable wayfinding cues such as signage and permanent store markers. Poor navigators, in contrast, more frequently referenced transient and incidental features such as campaign posters, parked bikes, and decorative details. These findings extend prior verbal-report studies [23] with gaze evidence, confirming that the groups differ not only in *what they articulate* but also in *how they visually scan* the environment.

Building on these findings, we introduce LandmarkLens to address the second question: whether guiding attention toward effective landmarks can improve spatial knowledge acquisition for people with poor SOD. LandmarkLens is a mixed-reality (MR) navigation system that highlights landmarks informed by the attention patterns and articulation strategies of skilled navigators, layered on top of conventional turn-by-turn directions. For instance, the system might highlight a neighborhood name sign that people typically overlook, a landmark visible behind you that connects the new street to the one you just left, or a distant tower that orients the pedestrian within the broader cityscape.

To make this approach scalable beyond our study participants, we translated the empirical findings into a landmark prediction pipeline. We used gaze distributions from good navigators to crop input imagery to relevant regions, and qualitative insights from their verbal descriptions to construct vision-language model (VLM) prompts that simultaneously promote effective reference points and suppress the transient features poor navigators tend to over-encode. The resulting MR interface applies the VLM to street-view imagery to predict and present landmarks along a route. Predicted landmarks are overlaid as interactive highlights. Directional arrows draw users toward highlights outside their field of view, and a dwell-to-dismiss interaction rewards deliberate attention to each one. To our knowledge, LandmarkLens is the first system to leverage skilled navigators’ gaze behavior and landmark recognition as an attention-guidance mechanism for people with poor SOD.

We evaluated LandmarkLens in a controlled VR study with eight participants who scored ≤ 3.5 on the SBSOD scale. Participants navigated two routes, one using LandmarkLens and one using a turn-by-turn navigation baseline, with route and condition order counterbalanced. Following [55], we assessed spatial knowledge through three tasks: scene recognition, route choice, and map sketching. Participants using LandmarkLens showed significantly higher scene recognition accuracy and recalled nearly twice as many landmarks in map sketches compared to baseline, while route choice accuracy and map structural accuracy remained comparable across conditions. These results suggest that landmark attention guidance enhances landmark encoding, offering a promising direction for future MR navigation systems that support spatial learning.

In summary, our contributions are: (1) a landmark attention study with 20 participants revealing that good navigators scan a broader horizontal extent of the scene and verbally reference more durable wayfinding cues, while poor navigators disproportionately encode transient and incidental features; (2) LandmarkLens, an MR navigation system that translates these empirical findings into a VLM-based landmark prediction pipeline and presents navigation-relevant landmarks through gaze-interactive overlays, augmenting turn-by-turn directions to encourage environmental engagement; (3) a preliminary within-subjects evaluation with eight participants showing that landmark attention guidance significantly improves scene recognition accuracy, with qualitative insights that inform future landmark-based MR navigation; (4) an open-source dataset of over 3,000 8K 360° street-view images with gaze and landmark annotations across Tokyo, supporting future research in HCI, environmental psychology, and spatial cognition. All data, code, and study materials are available at <https://github.com/makeabilitylab/landmarklens>.

2 Related Work

We situate our work at the intersection of spatial cognition, individual differences in navigation, and landmark-enhanced MR systems.

2.1 Landmarks, Routes, and Cognitive Maps

Our system design and evaluation build on Siegel and White [42]’s widely adopted trichotomy of spatial knowledge: *landmark knowledge*, recognizing discrete objects or scenes; *route knowledge*, encoding landmark sequences and the actions they trigger (e.g., turn left at the church); and *survey knowledge*, building a *cognitive map* of the environment (e.g., [21, 25, 35, 54, 55]).

Central to all three levels are **landmarks**: stationary, distinct, and salient objects that serve as reference points for constructing mental representations of the environment [34]. Researchers have classified landmarks by saliency — visual, cognitive, or structural [26, 44] — and by visibility: global landmarks (e.g., mountains or tall buildings) provide broad orientation from afar, while local landmarks aid navigation at close-range decision points [45]. **Route knowledge** builds on landmarks through associative memory [18], encoding landmark-place relationships (*when I see the town hall, I am at the town center* [31]), place-action pairings (*turn left at the town hall* [51]), and sequential order (*once I pass the town hall, I will arrive at the restaurant* [46]). The most challenging level is **survey knowledge**—a configurational, map-like representation



Figure 2: Our custom VR street-view navigation system for the landmark attention study. Directional arrows indicate available movement options, turn notifications appear at intersections.

encoding the environment’s layout and the spatial relationships among landmarks [37, 48]. Acquiring it requires integrating separately learned landmarks and routes within a common frame of reference, demanding understanding that extends beyond any single viewpoint [21, 33, 55].

In LandmarkLens, we evaluate all three levels: following Wen et al. [55]’s task design, our evaluation assesses scene recognition (landmark knowledge), route choice (route knowledge), and map sketching (survey knowledge), enabling us to measure how landmark attention guidance affects spatial learning at each level.

2.2 Individual Differences in Cognitive Mapping

Individuals differ widely in cognitive mapping ability [16, 21, 22, 52–54], from expert navigators [30] to those with severe navigational impairments [19]. The Santa Barbara Sense-of-Direction Scale [16], a self-report questionnaire assessing navigational abilities on a 7-point scale, has become the most widely used instrument for measuring these differences [22, 23, 53–55]. For reference, a cross-cultural validation ($N = 550$) found a population mean of 4.2 ($SD = 1.1$) on the SBSOD scale [36].

These differences have practical consequences: poor cognitive mapping leads to spatial anxiety [28] and over-reliance on GPS-based navigation tools [13], which in turn draw attention away from environmental features [9, 10], reduce active route planning [56], and impair spatial learning over the long term [40] — creating a vicious cycle in which poor navigators become increasingly dependent on tools that further attenuate their spatial learning.

Researchers have explored interventions including repeated route exposure [50], practice with feedback [24], and verbalization during learning [5]. Of particular relevance is the role of landmark attention: Ishikawa and Nakamura [23] found that people with good SOD select fewer but more effective landmarks, while people with poor SOD select unreliable features. This raises a natural question: can instructing people to attend to effective landmarks facilitate cognitive mapping? To date, this remains untested [21]. We designed LandmarkLens to investigate this possibility.

2.3 Landmark Enhancements in MR

Landmark-based navigation systems effectively support wayfinding and spatial knowledge acquisition [11, 57], and augmented reality (AR) takes this further by integrating virtual cues into the user’s real-world view [7]. For indoor navigation, researchers have used AR to enhance landmarks for mental map development [58], label semantic landmarks to promote spatial learning [29], support older adults with landmark-based instructions [38], and augment landmark perception for people with low vision [3]. For outdoor urban environments, commercial tools such as Google Maps Live

View superimpose only arrows and text, while recent research has explored landmark-based approaches including 3D virtual landmark maps [59], destination landmark overlays [32], and detection models trained on researcher-curated salient landmarks [39].

A key question across these systems is how landmarks are selected. In much prior work, selection has been determined by researchers based on general saliency principles (e.g., [38]), while more recent approaches have adapted selection based on user familiarity [61], formative studies with target users [3], or computational detection models [39]. However, none of these approaches ground landmark selection in the cognitive and perceptual differences between good and poor navigators. LandmarkLens bridges this gap by identifying landmark attention patterns from people with good SOD and using them to guide people with poor SOD, directly addressing the open question of whether attention to effective landmarks can facilitate spatial knowledge acquisition [21].

3 Landmark Attention Study

To understand how people with good versus poor SOD attend to landmarks during urban navigation, we conducted a study combining eye gaze tracking and verbal descriptions. Prior work has examined landmark selection through verbal reports alone [23]; our study extends this by collecting concurrent gaze and verbal data from 20 participants navigating street-view routes across four Tokyo neighborhoods in VR.

3.1 Participants

We recruited 20 participants through university mailing lists and snowball sampling. Potential participants completed a screening survey that included the SBSOD questionnaire [16]. We selected the first ten respondents who scored ≥ 5.0 (good SOD) and the first ten who scored ≤ 3.5 (poor SOD); these cutoffs are consistent with [24] and fall above and below the population mean reported for the SBSOD scale [36]. The mean SBSOD score was 6.0 ($SD = 0.7$) for the good group and 2.9 ($SD = 0.5$) for the poor group.

Participants had a mean age of 26.9 ($SD = 2.6$) in the good group and 28.2 ($SD = 4.1$) in the poor group. The good group included two women and eight men; the poor group included six women and four men. Participants were compensated with a JPY ¥5,000 gift card. See Appendix Table 3 for full participant information.

3.2 Apparatus

Street-Level Imagery. We collected 360° street-level imagery from four Tokyo neighborhoods, selected based on Almazán et al. [1] categorization of Tokyo’s urban archetypes: Nezu (*Village Tokyo*), Koenji (*Local Tokyo*), Ginza (*Office Tower Tokyo*), and Kinshicho

(*Mercantile Tokyo*). We captured 360° timelapse video at 8K resolution using GoPro Max2. The videos were segmented into equirectangular photo frames at 0.1 second intervals, yielding a total of 3,120 panoramic images.

Route Selection. Each neighborhood contained 11 routes: one common route walked by all 20 participants, and ten unique routes assigned one per participant (Figure 14). A total number of 2,377 panoramas were included. Common routes enable direct gaze comparison across participants, while the unique routes maximize imagery diversity. Routes were balanced for number of images and number of turns within each neighborhood. Each participant navigated eight routes (one common and one unique route per neighborhood). To mitigate ordering effects on gaze data, route order was counterbalanced across participants using Latin square design. See Appendix Table 4 for the full route presentation order.

VR System. We developed a custom street-view navigation application in Unity, deployed on a Samsung Galaxy XR headset (Qualcomm Snapdragon XR2+ Gen 2, 3552 × 3840 per-eye micro-OLED display, 109° horizontal field of view, four eye-tracking cameras). The application renders each equirectangular panorama on the interior of a sphere surrounding the user. Participants navigated using a handheld controller. Transitions between panoramas were rendered as cross-fade animations, with a turn animation applied when changing direction at intersections. Directional arrows indicated available movement options at each panorama, and turn notification overlays (e.g., “Turn right”, “Right turn ahead”) appeared when approaching intersections to guide participants along the route. The system logged eye gaze origin, gaze direction, gaze hit point on the panorama (UV coordinates), head position, head rotation, and input events at the headset’s native sampling rate (72 Hz); audio was recorded using an external clip-on microphone.

3.3 Procedure

Introduction. Following a two-participant pilot that informed our introduction script and task guidance, each session began with an introduction to the study and informed consent. The researcher showed participants screenshots explaining the navigation controls. Participants were then fitted with a Samsung Galaxy XR headset and eye gaze tracking was calibrated. A gaze visualization sphere was briefly shown to confirm tracking accuracy, then turned off.

Practice. Participants completed a short practice route (26 images from an unused Nezu route) to familiarize themselves with the system. They were given the task instructions: “Imagine you are giving directions to someone who is visiting this area for the first time. As you navigate each route, please verbally describe any features along the way that you think would be helpful as navigation clues.”

Routes. Participants completed eight routes. After each route, participants removed the headset, verbally described the route they had just taken including any visual clues and directions they could recall. As pilot participants preferred sketching during recall, we also provided pen and paper.

Retrace. Participants were informed at the outset that one route would be randomly selected for them to navigate again from memory. In practice, this was always the final route, ensuring participants remained attentive to navigation cues throughout all routes.

Debriefing. After all routes were completed, we debriefed participants, answered any questions about the study, and showed them the routes on a map upon request.

3.4 Analysis

We analyzed participants’ gaze behavior along four dimensions: gaze spread, fixation rate, fixation duration, and head rotation. For gaze spread, we computed the standard deviation of gaze hit coordinates on the equirectangular panorama. For fixation detection, we used the I-DT algorithm [41] with a dispersion threshold of 0.02 in UV space and a minimum duration of 150 ms. We fitted linear mixed models for gaze spread and head rotation, and Gamma generalized linear mixed models for fixation rate and duration. For the retrace task, we modeled intersection-level accuracy using a binomial generalized linear mixed model with group as a fixed effect and route and participant as random intercepts. One participant (A17) was identified as an outlier via the $1.5 \times \text{IQR}$ criterion on session duration, gaze spread, and head rotation, and was excluded from all quantitative analyses, yielding a final sample of ten good and nine poor navigators.

We transcribed participants’ verbal landmark descriptions and analyzed them using a hybrid deductive-inductive coding approach [2]. One researcher developed an initial codebook based on landmark categories from prior work [23, 44], then iteratively refined it as new codes emerged. To establish reliability, a second researcher independently coded two transcripts, one from each SOD group. Disagreements were discussed until consensus was reached. The remaining 18 transcripts were then recoded using the revised codebook.

3.5 Results

Our dataset contained approximately 1.65 million gaze samples across 19 participants, of which 804K were recorded on common routes. The good SOD group contributed a mean of 76,626 samples per participant ($SD = 17,578$), and the poor SOD group contributed a mean of 75,169 ($SD = 15,099$). Mean total session duration was 75.4 minutes ($SD = 12.0$) for the good group and 71.5 minutes ($SD = 10.2$) for the poor group.

3.5.1 Gaze Behavior. On common routes, good navigators exhibited significantly greater total gaze spread than poor navigators ($F(1, 17) = 5.55, p < .05, \eta_p^2 = .25; M_{\text{good}} = 0.13, SD = 0.03; M_{\text{poor}} = 0.11, SD = 0.03$). This difference was driven by horizontal spread ($F(1, 17) = 6.16, p < .05, \eta_p^2 = .27$), while vertical spread did not differ between groups ($F(1, 17) = 0.87, p = .363, \eta_p^2 = .05$). In our analysis of fixation rate (i.e., how frequently participants locked onto a location) and mean fixation duration (i.e., how long each fixation lasted), we found no significant group differences (rate: $\chi^2(1) = 0.08, p = .772$; duration: $\chi^2(1) = 0.41, p = .520$). See Figure 4, Figure 5.

3.5.2 Head Rotation. Good navigators turned their heads more than poor navigators. On common routes, good navigators showed significantly greater head yaw variability ($F(1, 17) = 6.59, p < .05, \eta_p^2 = .28; M_{\text{good}} = 36.0^\circ, SD = 9.7^\circ; M_{\text{poor}} = 28.2^\circ, SD = 6.8^\circ$) and head pitch variability ($F(1, 17) = 4.93, p < .05, \eta_p^2 = .23; M_{\text{good}} = 10.7^\circ, SD = 2.2^\circ; M_{\text{poor}} = 9.0^\circ, SD = 2.1^\circ$). See Figure 5.

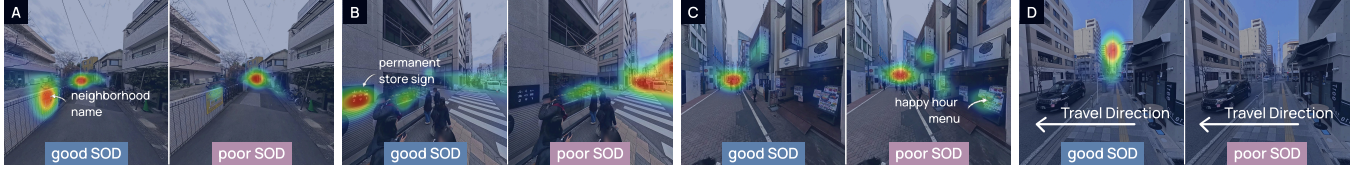


Figure 3: Examples of divergent gaze patterns. Good navigators attended to (A) wayfinding signage and (B) permanent store identifiers. Poor navigators fixated on (C) menus and temporary signage. (D) Good navigators turned to fixate on the Tokyo Skytree as a global landmark; poor navigators did not.

Together, the gaze spread and head rotation findings suggest that good navigators survey a broader horizontal extent of the scene through both eye movements and physical head turns.

3.5.3 Route Retrace Performance. In route retrace, good navigators chose the correct direction at intersections at a higher rate than poor navigators ($M_{\text{good}} = 75.5\%$, $SD = 29.3\%$; $M_{\text{poor}} = 57.5\%$, $SD = 27.8\%$). An analysis of variance based on mixed logistic regression indicated a statistically significant effect of group on intersection accuracy ($\chi^2(1, N = 137) = 4.51$, $p < .05$). Retrace duration did not differ between groups ($M_{\text{good}} = 289.5$ s, $SD = 82.0$; $M_{\text{poor}} = 312.0$ s, $SD = 84.9$; $F(1, 9.2) = 1.01$, $p = .340$, $\eta_p^2 = .10$). The retrace task served to maintain attentiveness throughout the session and to validate the SBSOD-based group assignment: despite each pair retracing a different route, the significant accuracy difference provides converging evidence that the groups captured meaningful differences in navigation ability.

3.5.4 Verbal Data. We coded a total of 1,009 landmark mentions from good navigators and 1,170 from poor navigators across all transcripts (see Appendix Table 5 for full results). Both groups predominantly referenced non-residential buildings, which accounted for 58.6% and 52.0% of mentions respectively. However, the groups diverged in what else they attended to. Good navigators disproportionately mentioned wayfinding signage (5.8% vs. 1.8%), while

poor navigators more frequently referenced trivial details such as decorative elements (0.5% vs. 3.5%) and parts of buildings (0.6% vs. 2.5%). The most striking differences emerged around transient and incidental features. Poor navigators referenced pedestrians (0.1% vs. 1.9%), parked bikes (0.3% vs. 0.9%), and generic signage that does not contain any location information (0.7% vs. 2.6%) far more frequently than good navigators. This is consistent with prior work [23] that people with good SOD select more effective landmarks.

Inspecting frames with the most divergent gaze distributions between groups revealed consistent patterns. Good navigators attended to wayfinding signage (Figure 3A) and permanent store signs such as names engraved in stone (Figure 3B), while poor navigators fixated on menus and temporary store signage (Figure 3C). Most notably, when walking perpendicular to the Tokyo Skytree (Figure 3D), good navigators actively turned their heads to fixate on the tower as a global landmark, while poor navigators did not.

We also observed that the difference between groups was not only what caught their eye, but what they chose to articulate as useful for navigation. When both groups gazed at the same campaign poster in Ginza, their verbal responses revealed different evaluative strategies. A01 (good SOD) dismissed it: “*I don’t need to mention it [campaign poster]. It will get removed later.*” In contrast, A19 (poor SOD) said: “*Hopefully they don’t change whatever that campaign is for next time I’m here.*” This suggests that the gap lies not in perception but in an evaluative filter—good navigators actively suppress transient features from their mental representations, while poor navigators encode them despite awareness of their unreliability.

4 LandmarkLens

LandmarkLens consists of two components: a landmark prediction pipeline and an MR navigation interface. Given a set of 360° images along a route, the prediction pipeline first crops each panorama to the forward-facing region where navigators concentrate their gaze, then passes the cropped images to a VLM that identifies navigation-relevant landmarks based on insights from our landmark attention study (Section 3). The navigation interface renders these landmarks as gaze-interactive overlays layered on top of conventional turn-by-turn directions, guiding users’ attention towards visual features that skilled navigators select as effective clues.

4.1 Landmark Prediction Pipeline

Our landmark prediction pipeline translates findings from our landmark attention study into a scalable approach for identifying navigation-relevant landmarks in street-view imagery. The pipeline operates in four stages: gaze-informed image cropping, VLM-based landmark detection, embedding-based deduplication, and second-pass validation (Figure 6).

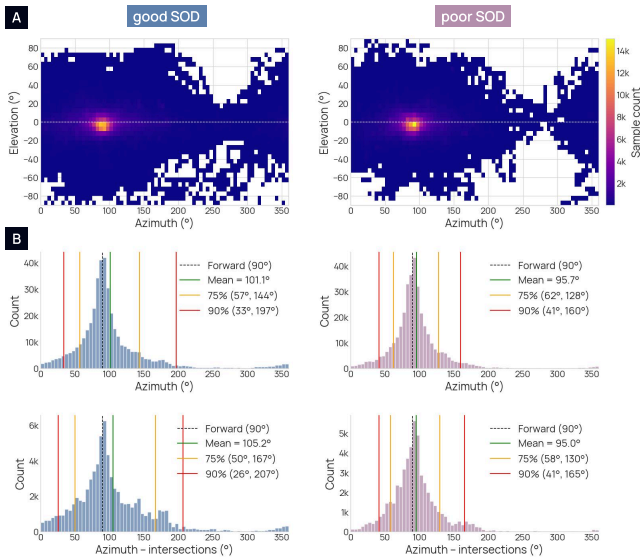


Figure 4: Gaze distribution profiles. (A): 2D heatmaps on equirectangular panoramas. (B): azimuth distributions with percentile boundaries used for VLM cropping extents.

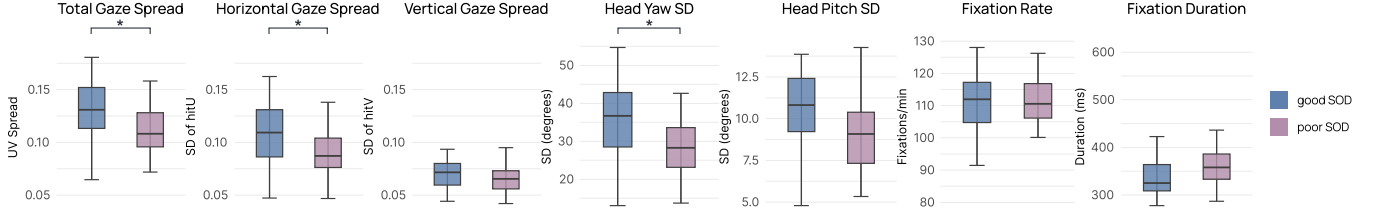


Figure 5: Gaze, fixation, and head movement by group. Good navigators exhibited significantly greater total gaze spread, horizontal gaze spread and head yaw variability. Fixation patterns did not differ significantly between groups.

4.1.1 Gaze-Informed Image Cropping. Raw equirectangular panoramas span a full $360^\circ \times 180^\circ$ field of view, but navigators concentrate their gaze on a forward-facing subset. Passing full panoramas to a detection model dilutes predictions with irrelevant regions: sky, ground and areas behind the walker. We therefore crop each panorama to where good navigators mostly look, using percentile boundaries from their gaze distributions (Figure 4).

Each panorama is first rotated to center the walking direction. At intersection frames, an additional half-turn rotation centers the crop on the bisector between the approaching and turning directions. Based on the 90th-percentile azimuth interval, we crop to 180° at non-intersections (Figure 4). At intersections we widen the crop to 270° so that it covers both the approaching and the turning direction after the half-turn rotation, reflecting broader gaze spread at decision points. Within the crop, a narrower guideline region taken from the 75th-percentile interval marks a preferred detection zone: 90° at non-intersections and 120° at intersections. Vertical extent is fixed across both cases at the 95th-percentile elevation interval, which spans 0.5 in normalized v coordinates.

4.1.2 VLM-Based Landmark Detection. Given the cropped imagery, the pipeline identifies navigation-relevant landmarks using a VLM. **Design Rationale.** We initially explored fine-tuning a DeepGaze model [27] on good-navigator gaze heatmaps to predict visually salient regions. However, the model achieved poor alignment with observed gaze (KL divergence of 1.62 on the Kinshicho common route); it often highlighted road surfaces and blank walls. Moreover, our attention study showed that when the groups gazed at similar regions, the key distinction lies in which features they judged worth encoding for navigation (Section 3.5.4)—a semantic distinction that a pixel-level saliency model cannot represent. This motivated a VLM-based approach, where such distinctions can be expressed through natural language prompting.

Prompt Design. The pipeline uses Google Gemini 3.1 Pro

Preview [12] for landmark detection. The prompt instructs the model to identify landmarks drawn from priority-ranked categories from our qualitative codebook (Table 5), and to prioritize objects whose center falls within the guideline region while still permitting high-priority landmarks in the periphery. Importantly, the prompt emphasizes *relative distinctiveness* within each scene—a vending machine is a stronger anchor on a quiet residential street than on a busy commercial strip. Non-intersection frames are processed in consecutive batches of five with a limit of two landmarks per image, while intersection frames are isolated into single-frame batches with a separate prompt allowing up to three landmarks. Bounding boxes tightly enclose the distinctive visual feature itself — *e.g.*, a café’s neon sign rather than the whole building. Each detection is scored on color prominence, shape prominence, and semantic familiarity (*i.e.*, how easily a person could describe the object), and returned as structured JSON via Gemini’s `response_schema` parameter, including a bounding box, description, category label, and navigational context. See Appendix Section B.1 for details.

4.1.3 Embedding-Based Deduplication. The per-batch prompt deduplication reduces but does not eliminate duplicate detections across adjacent batches. For each detection, we crop the bounding box region from the rotated panorama and extract a feature embedding using the DINOv3 vision encoder [43]. Within each route segment, we compare detections in nearby frames (± 5 frame window) via cosine similarity of their L2-normalized CLS token embeddings. Because the pipeline processes each segment independently, we also compare detections across consecutive segment boundaries—matching the last frame of each segment against the first of the next. Detection pairs exceeding a similarity threshold of 0.70 are linked via union-find, and each group is pruned to retain only its earliest occurrence.

4.1.4 Second-Pass Validation. Finally, each remaining detection undergoes a second-pass validation. The detection’s bounding box

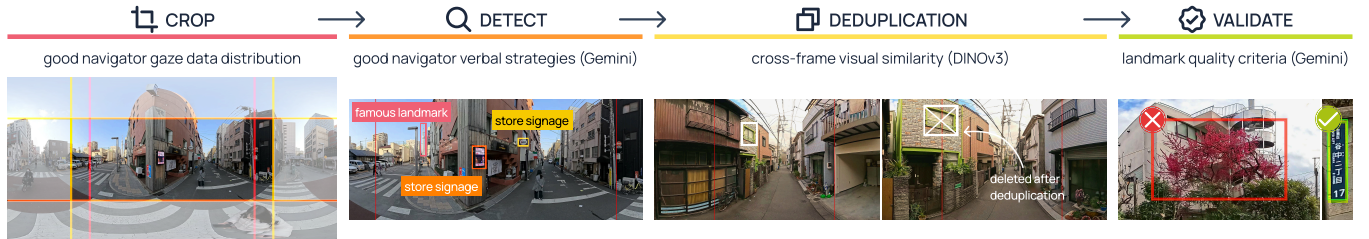


Figure 6: The landmark prediction pipeline: (1) gaze-informed cropping, (2) VLM-based detection using good navigators’ verbal strategies, (3) embedding-based deduplication via DINOv3, and (4) second-pass validation.



Figure 7: Examples of VLM detection results. In (A) and (B) VLM detections closely overlap with gaze hotspots. In (C) and (D), unmatched gaze hotspots correspond to objects such as store interiors, parked cars, and sidewalks—features that attracted visual attention but never appeared in participants’ verbal landmark descriptions.

region is cropped from the source panorama (with a $1.5\times$ buffer) and overlaid with a translucent rectangle marking the original detection area. This annotated crop is sent individually to Gemini along with the first-pass description and category label. A separate validation prompt (Appendix Section B.2) instructs the model to verify that the described object is visible in the cropped region, that the category and bounding box refer to the same real-world object, that the object is permanent rather than transient, and that it is visually distinctive enough to serve as a landmark. The prompt includes an explicit exclusion list of generic municipal infrastructure—street lights, traffic signs, guardrails, utility poles—and temporary features such as banners and flags, reflecting the types of objects our qualitative findings showed poor navigators tend to over-encode. Invalid detections are excluded from the final landmark set.

On average, the pipeline processes each panorama in eight seconds; for example, evaluation Route B (67 panoramas) took 9.2 minutes and cost USD \$1.77 in total.

4.2 Evaluating the Landmark Guidance Pipeline

To assess whether the pipeline identifies objects that good navigators attend to, we compared VLM detections against ground-truth gaze heatmaps aggregated from the ten good-navigator participants

Area	Frames	VLM		Precision	GT		Recall
		Matched	Unmatched		Matched	Unmatched	
Ginza	105	123	19	86.6%	75	64	54.0%
Koenji	177	191	42	82.0%	97	73	57.1%
Nezu	149	85	30	73.9%	32	55	36.8%
Kinshicho	122	123	26	82.6%	45	54	45.5%

Table 1: VLM pipeline evaluation against ground-truth data.

across four common routes. We ran this evaluation on validated but non-duplicated detections. Since gaze data spans multiple consecutive frames for the same object, evaluating against deduplicated detections would artificially penalize recall.

For *precision*—whether VLM detections fall where people actually looked—we compute each detection’s density ratio: the share of total gaze captured by the bounding box divided by the box’s area fraction, where values above 1.0 indicate the detection captures disproportionately more gaze than expected by chance. A detection is classified as matched if its density ratio ≥ 1.0 . For *recall*—whether the VLM found what people looked at—we threshold the gaze heatmap at the 95th percentile of nonzero values and extract connected components as hotspot regions. Each hotspot is matched to VLM detections using $\text{IoU} \geq 0.1$ or $\text{IoMin} \geq 0.15$. A ± 1 frame cross-matching pass accounts for the same landmark attracting gaze on adjacent frames. To avoid inflating both metrics with gaze attributable to forward path-monitoring rather than landmark recognition, we exclude detections and hotspots whose centroid falls within a center exclusion zone (30° wide $\times$ 15% of panorama height) aligned with the travel direction. Table 1 reports results. Figure 7 shows examples of the evaluation output. In cases of unmatched gaze hotspots, they correspond to objects such as store interiors, parked cars, and sidewalks—features that attracted visual attention but never appeared in verbal landmark descriptions.

4.3 Guiding Landmark Attention in MR

Circle Highlights. As our goal is to direct gaze toward specific objects in the scene, we present landmarks as visual overlays rather than audio cues, an approach supported by prior findings that visual MR guidance reduces navigation errors and cognitive load compared to audio [60]. Each predicted landmark is presented as a



Figure 8: A walkthrough of LandmarkLens during navigation. As the user approaches an intersection, landmark highlights appear on the scene. A directional arrow prompts the user to look toward a highlight they haven’t yet noticed. Once the user gazes at the landmark, the circle pulses and fades, and the user proceeds to the next panorama.

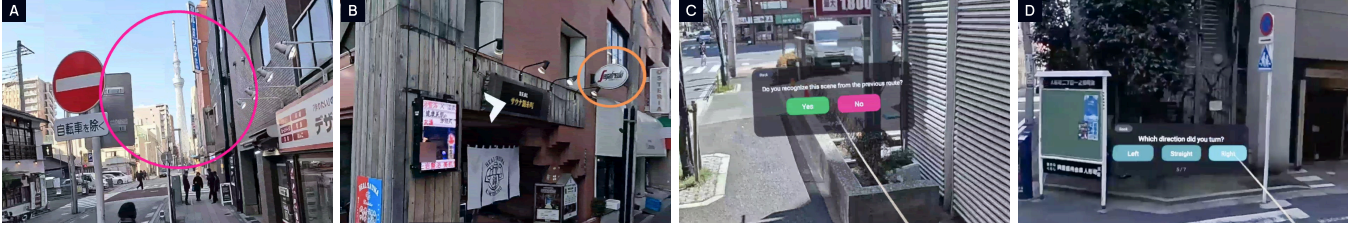


Figure 9: User Evaluation VR interfaces. (A–B) LandmarkLens condition showing circle highlights and directional arrows. (C) Scene-recognition task. (D) Route-choice task.

high-saturation neon circle overlaid on the panorama at the predicted location (Figure 8). We chose circles over bounding boxes for two reasons. First, a circle’s geometric simplicity stands out against the rectilinear urban environment. Second, unlike a tightly fitted bounding box that isolates an object, a circle encompasses both the landmark and its immediate surroundings, encouraging users to encode the object in the context of its environment. Circle color (magenta, orange, or yellow) is assigned based on the visual prominence scores from the VLM output, providing a subtle cue to the relative distinctiveness of each landmark. To balance engagement with distraction, each circle pulses when the user gazes at it and fades out after a two-second fixation. This gaze-to-dismiss interaction lightly gamifies the experience; users “collect” landmarks as they navigate, turning passive observation into an active visual scavenger hunt rewarding deliberate attention.

Directional Arrows. When a predicted landmark is outside the user’s current field of view, a directional arrow appears at the edge of the display pointing toward it, guiding the user to rotate their head toward the landmark. Once the circle enters the field of view, the arrow parks at the display edge and remains visible until the circle is dismissed through the gaze interaction.

Implementation. We implemented LandmarkLens as a VR navigation application in Unity, deployed on a Samsung Galaxy XR headset, extending the street-view navigation system from the landmark attention study with an additional landmark overlay layer.



Figure 10: Evaluation routes and task image locations in Ningyocho. Filled circles: scene-recognition targets; crosses: distractors; open circles: route-choice intersections.

5 Preliminary User Evaluation

To evaluate whether LandmarkLens attention guidance supports spatial knowledge acquisition, we conducted a within-subjects study comparing LandmarkLens against a turn-by-turn baseline.

5.1 Participants

We recruited eight participants (five women, three men; mean age = 27.4, $SD = 2.3$), six from our previous landmark attention study and two through university mailing lists. All participants scored ≤ 3.5 on the SBSOD scale ($M = 2.8$, $SD = 0.4$). Participants were compensated with a JPY ¥5,000 gift card. See Appendix Table 6.

5.2 Study Design

Street-Level Imagery. To avoid familiarity effects for participants who had taken part in the landmark attention study, we collected new imagery from Ningyocho, a neighborhood that shares the same *Mercantile Tokyo* archetype as Kinshicho.

Route Selection. We selected two routes in Ningyocho (Figure 10), balanced on the following criteria: (1) *Length*: Route A was 474.2 m; Route B was 410.9 m. (2) *Structure*: Both routes contained six segments and five turns, with a balanced number of left and right turns. (3) *Decision points*: Both included two additional straight-through intersections. (4) *Visual anchors*: Both routes featured a mixture of commercial and residential buildings, with distinctive visual landmarks at turning intersections. We applied the VLM landmark prediction pipeline (Section 4) to all frames. The pipeline produced 49 highlights on Route A (63 frames) and 58 on Route B (67 frames), yielding a comparable density of 0.8 highlights per frame.

Task Design. Following [55], we assessed participants’ spatial knowledge using three tasks targeting landmark, route, and survey knowledge respectively. No time limit was imposed on any tasks. In the *scene-recognition task*, participants viewed 24 images in VR, presented one at a time in random order, and indicated whether they recognized each scene from the exploration phase. Half of the images were taken from the explored route at non-intersection points; the other half were taken from nearby streets that participants had not explored (Figure 10). All route images were taken from frames where LandmarkLens had placed at least one highlight. In the *route-choice task*, participants viewed seven 360° images taken at intersections along the route and indicated which direction they had traveled during the exploration phase (left, right, or straight; Figure 9). In the *map-sketching task*, participants drew the explored route in as much detail as possible on a blank sheet of paper.

5.3 Apparatus

In both conditions, participants navigated the VR street-view system (Section 3) by tilting the controller joystick, and text notifications at the bottom of the display indicated the required direction at each step (e.g., *Go straight, Turn right*). In the LandmarkLens condition, the system additionally displayed gaze-interactive landmark highlights generated by the VLM pipeline (Section 4). Landmarks were presented as colored circles overlaid on the panorama at predicted locations, with an arrow indicator guiding participants toward circles outside their current field of view. Participants dismissed each circle through a two-second fixation, after which it faded out. Both conditions logged gaze and head-tracking data. The scene-recognition and route-choice tasks were also administered in VR; the map-sketching task was completed on A4 paper.

5.4 Procedure

Each session lasted approximately 60 minutes across four phases: introduction, practice, route exploration with tasks, and post-study debrief. We piloted with two participants to refine the procedure.

Introduction. Each session began with an introduction to the study and informed consent. The moderator described the session structure and explained that after each route participants would be asked whether they recognized certain scenes, which direction they turned at intersections, and to sketch the route from memory.

Practice. Participants completed two short practice routes to familiarize themselves with the system. The first introduced basic navigation controls without landmark overlays; the second introduced the LandmarkLens overlay. Participants were told that the highlighted circles indicated visual elements that might serve as helpful navigation cues, and were encouraged—but not required—to gaze at and dismiss each circle before advancing.

Route Exploration and Tasks. Participants explored two routes, one with LandmarkLens and one with the turn-by-turn baseline, in counterbalanced order. Route-to-condition assignment was also counterbalanced, yielding four counterbalance groups with two participants each. After each route, participants completed the three spatial knowledge tasks in fixed order: map sketching, scene recognition, and route choice. This ordering follows [55] to minimize preceding tasks influencing subsequent ones. Participants then rated seven Likert-scale items assessing confidence, spatial awareness, navigation independence, and perceived mental load.

Post-Study. After completing both routes, participants answered Likert-scale questions on route memory, system usability, and perceived usefulness, followed by a short semi-structured interview.

5.5 Analysis

For task performance on scene recognition and route choice, we used *mixed logistic regression* to model the proportion of correct responses for each task [15].

For map sketching, we scored each sketch on route completeness (segments drawn out of six), landmark count (number of landmarks freely recalled), and shape similarity. For shape similarity, we digitized each sketch as an SVG path, extracted vertex coordinates, and computed *bidimensional regression* [47] between the sketch and the actual route geometry after optimal translation, scaling, and rotation. Correlation values were transformed using Fisher’s r -to- z

transformation prior to statistical testing. For ordinal Likert-scale responses (1–7), we used *mixed ordinal logistic regression* [14]. For responses to open-ended questions, we focused on summarizing high-level themes. One researcher developed themes through open coding [2] of interview transcripts and questionnaire responses.

6 Findings

On average, it took participants 6m 24s to finish the prototype condition ($SD = 2m\ 6s$), and 3m 55s to finish the baseline condition ($SD = 2m\ 42s$). The longer completion time in the prototype condition reflects the gaze-to-dismiss interaction: although not required, most participants chose to dismiss all circles before advancing to the next panorama. Below, we report quantitative results alongside participants’ explanations.

6.1 Task performance

6.1.1 Scene Recognition. Outlier analysis using z -score ($|z| > 2$) and IQR ($1.5 \times IQR$) methods identified one participant (E8) as an outlier on scene recognition; results are reported without this participant ($N = 7$). Participants spent a similar amount of time on the scene recognition task across conditions, with a mean response time of 8.63 s ($SD = 3.94$) in the baseline condition and 9.47 s ($SD = 4.91$) in the prototype condition.

Participants recognized scenes more accurately with the prototype ($M = 79.2\%, SD = 4.2\%$) than with the baseline ($M = 64.9\%, SD = 11.1\%$). An analysis of variance based on mixed logistic regression indicated a statistically significant effect of condition on scene recognition accuracy ($\chi^2(1, N = 14) = 8.36, p < .01$). Participants attributed this improvement to the highlights drawing attention to features they would otherwise overlook. E1 said “*the circles let me remember things I wouldn’t normally notice*,” and E6 remarked that “*the tall buildings are things that I won’t notice without the circles*.” E7 described the experience as “*quite interesting... it highlighted things that I would not usually pay attention to*.”

The excluded participant, E8, offered insight into *why* the highlights did not benefit her scene recognition. She reported that the circles shifted her attention away from her own encoding strategy: “*I’m clearing the circles... but I’m not really paying any attention to where I’m actually going or like where they are in relation to each other*.” She explained that the highlights were most effective “*when it aligned with the kind of stuff I noticed*,” but because they were “*not personalized*,” they disrupted rather than supported her recall.

PID	Route Completeness (%)		Landmark Count		Shape Similarity (r)	
	Baseline	Prototype	Baseline	Prototype	Baseline	Prototype
E1	50%	33%	6	6	0.96	0.94
E2	100%	100%	0	1	0.96	0.98
E3	100%	83%	2	8	0.96	0.74
E4	100%	100%	1	0	0.95	0.97
E5	83%	67%	0	3	0.72	0.68
E6	100%	83%	4	7	0.73	0.92
E7	100%	83%	3	10	1.00	0.72
E8	50%	83%	3	2	0.54	0.99
<i>Mdn</i>	100%	83%	2.5	4.5	0.96	0.93
<i>IQR</i>	50%	17%	3.0	5.5	0.23	0.24

Table 2: Map sketching results by participant and condition.

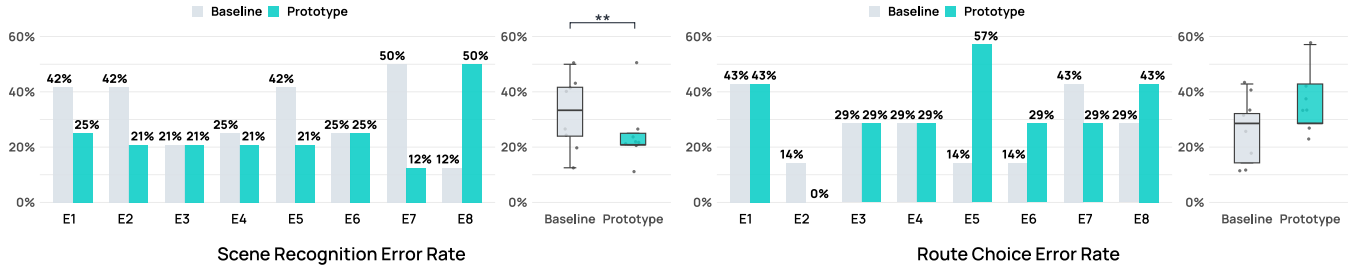


Figure 11: Left: Scene recognition error rate by participant and condition. LandmarkLens significantly improved accuracy over baseline (79.2% vs. 64.9%; $p < .01$). Right: Route-choice error rate per participant by condition. No significant difference ($p = .535$).

6.1.2 Route Choice. Participants' route choice accuracy was comparable across conditions ($M_{\text{prototype}} = 67.9\%$, $SD = 17.1\%$; $M_{\text{baseline}} = 73.2\%$, $SD = 11.4\%$). An analysis of variance based on mixed logistic regression indicated no significant effect of condition on route choice accuracy ($\chi^2(1, N = 16) = 0.39$, $p = .535$). Response times for the route choice task were comparable between conditions, with a mean of 9.49 s ($SD = 5.26$) in the baseline condition and 9.79 s ($SD = 5.97$) in the prototype condition.

A recurring theme in the interviews was that the volume of highlights could interfere with route encoding. E5 mentioned that “too many highlights can make me [...] only remember those surroundings, but not the route,” and E8 similarly noted that “because the circles were constant, it was harder for me to remember what landmarks were associated with some kind of directional change.”

6.1.3 Map Sketching. Map sketching performance was assessed along three dimensions: route completeness, landmark count, and shape similarity. Table 2 reports per-participant results.

Participants recalled nearly twice as many landmarks with the prototype ($Mdn = 4.5$, $IQR = 5.5$) as with the baseline ($Mdn = 2.5$, $IQR = 3.0$). An analysis of variance based on mixed Poisson regression indicated a statistically significant effect of condition on landmark count ($\chi^2(1, N = 16) = 5.80$, $p = .016$). Route completeness was comparable across conditions ($Mdn_{\text{baseline}} = 100\%$, $IQR = 50\%$; $Mdn_{\text{prototype}} = 83\%$, $IQR = 17\%$); an analysis of variance based on mixed logistic regression indicated no significant difference ($\chi^2(1, N = 16) = 0.77$, $p = .381$). Structural accuracy did not differ significantly between conditions. The median bidimensional correlation was comparable across baseline ($Mdn_{\text{baseline}} = .96$) and prototype ($Mdn_{\text{prototype}} = .93$). A paired-samples t-test on Fisher z-transformed correlations indicated no significant difference ($t(7) = -0.12$, $p = .911$, $d = -0.04$). Taken together, the map sketching results suggest that the prototype primarily enhanced landmark encoding—participants recalled nearly twice as many landmarks—without compromising the overall structural accuracy

of their mental representations. See Appendix Figure 17 for full participant sketches and analysis.

6.2 Subjective Experience

6.2.1 Post-Route Self-Assessment. Figure 13 shows the distribution of per-participant differences (prototype – baseline) across seven post-route Likert-scale items. Analyses of variance based on mixed ordinal logistic regression [14] indicated no statistically significant effects of condition on any individual item. Descriptively, however, six of seven items trended in favor of the prototype, with participants reporting higher confidence in scene recognition, intersection recall, surroundings awareness, and finding their way back, as well as lower mental effort. Map sketch confidence was the only item without a clear directional pattern.

6.2.2 Prototype Feedback. When asked which route they would remember better if they navigated it again the next day, six of eight participants chose the prototype route. Figure 12 summarizes participants' responses to prototype-specific questions. The landmark highlight circles were easy to notice ($Mdn = 6.5$), and the directional arrows were well-received, with all participants rating them positively for helping locate highlighted features ($Mdn = 6.0$). Participants generally found the system easy to learn ($Mdn = 6.5$) and most agreed it improved their navigation experience ($Mdn = 6.0$). Opinions were more mixed on whether gaze-based dismissal felt natural ($Mdn = 4.0$) and whether the circles were distracting ($Mdn = 3.0$). Willingness to use such a system with AR glasses in real life was generally positive ($Mdn = 5.0$).

Quantitative ratings were largely positive, and interviews revealed enthusiasm for pushing the system toward more “semantically meaningful and personalized” (E8) highlights. E3, who has a background in architecture, was engaged by *buildings with special façades*, illustrating how personal experience shapes what constitutes a useful landmark. E7 suggested adding contextual information such as a building's history, and E1 proposed color-coding

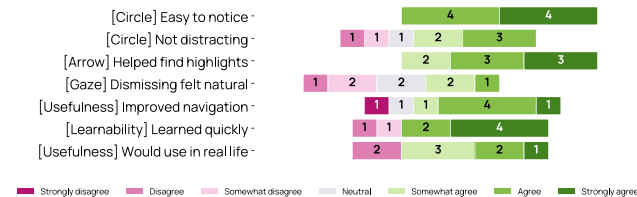


Figure 12: Prototype-specific Likert responses.

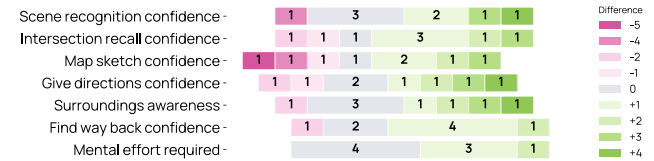


Figure 13: Per-participant Likert-scale differences (prototype-baseline) across post-route self-assessment items. Six of seven items trended in favor of prototype.

circles by building use. For participants with limited Japanese reading proficiency (E4 and E8), highlights on Japanese signage were difficult to retain—“*I can’t remember the signs I can’t read*” (E4).

7 Discussion

Guided Landmark Attention Partially Supports Spatial Knowledge Acquisition. LandmarkLens improved scene recognition but had more modest effects on route choice and map sketching. Following Sorrows and Hirtle’s taxonomy [44], our VLM pipeline primarily detects visual landmarks, which directly support scene-level recognition. However, route and survey knowledge depend more on structural and semantic landmarks, which our pipeline does not yet target. Participant feedback corroborated this: highlights were most useful at turns but overwhelming along straight segments (E3, E5, E8), and landmarks with unreadable text were impossible to remember (E4). Still, because landmark knowledge is the foundation for route and survey knowledge [42], an open question is whether repeated landmark-level gains could, over time, scaffold the acquisition of route and survey knowledge [42]. Investigating this longitudinal effect is a promising future direction.

VLMs as A Viable Approach for Landmark Prediction. VLMs are well-suited because the gap between good and poor navigators is an evaluative filter, not a perceptual one—good navigators see the same features but withhold irrelevant ones from their reports. This filter maps naturally onto VLM instruction-following: we encoded good navigators’ qualitative patterns directly into prompts, something purely vision-based saliency models cannot do. The approach also extends naturally: route structure (GPS, intersection topology) could enable context-sensitive highlighting at decision points, and because selection is prompt-mediated, it could personalize to language proficiency or area familiarity without retraining.

Individual Variability and Personalization. Our pipeline presents the same landmarks to all users, but we did observe within-group differences, where individuals within the good SOD group attended to different landmarks in the same environment (Appendix Figure 15). This aligns with the concept of cognitive landmarks, which hold personal meaning or relevance to an individual [44]. LandmarkLens does not yet account for these individual differences, and personalization based on user preference, language proficiency, and area familiarity is a natural next step.

Ecological Validity of VR-Based Studies. Both our landmark attention study and the user evaluation were conducted in VR using static 360° street-view imagery rather than real-world navigation. Prior work has shown that high-fidelity VR combined with naturalistic movement control can narrow the gap between VR and real-world gaze behavior [8]. However, we acknowledge that our environment lacked dynamic scene elements such as moving vehicles and pedestrians, vestibular cues from physical locomotion, and peripheral visual information beyond the headset’s field of view, all of which may influence gaze behavior and spatial learning during real-world navigation. While many systems intended for AR have been designed and evaluated in VR as a first step (e.g., [4]), future work should evaluate LandmarkLens with real-world navigation.

Future Work. LandmarkLens was evaluated in VR but designed for AR smart glasses—the gaze-interactive overlays (dwell-based dismissal, directional arrows) transfer directly to head-worn displays.

Real-world deployment introduces challenges (lighting variation, dynamic occlusion, inference latency) but also richer spatial encoding through actual locomotion. Our evaluation was also limited in scope: the final study involved eight participants, and all routes were drawn within Tokyo. Evaluation with larger samples and in other regions with different landmark densities or urban layouts would help establish the generalizability of both our pipeline and the attention guidance approach. More broadly, this work points toward a design paradigm for next-generation AR navigation: tools that guide users to their destination may support users in building their capacity to navigate independently, helping break the cycle in which GPS dependence erodes spatial ability [20, 40].

8 Conclusion

This paper presented LandmarkLens, an MR navigation system that guides people with poor sense of direction toward landmarks that skilled navigators rely on. A landmark attention study with 20 participants revealed that good navigators scan a broader horizontal extent of the scene and selectively articulate durable wayfinding cues, while poor navigators disproportionately reference transient features—a gap rooted in evaluative filtering rather than perception. We translated these findings into a VLM-based landmark prediction pipeline and gaze-interactive overlay system. A preliminary evaluation with eight poor-SOD participants showed that LandmarkLens significantly improved scene recognition accuracy and nearly doubled landmark recall in map sketches compared to a turn-by-turn baseline, while route and survey knowledge remained comparable. Participant feedback highlighted that landmarks were most valued at decision points and that highlights should be semantically meaningful and personalized, suggesting that future iterations should adopt context-sensitive highlighting that concentrates cues near intersections and adapts to individual familiarity and preferences.

Acknowledgments

We thank the reviewers for their insightful comments and our study participants for their time. We also thank Mauricio Sousa, Xia Ding, I-Chao Shen, Hongbo Zhao, Zhongyi Zhou, Yuan Li, Jaewook Lee, Quentin Becker, Maria Larsson, and Xinyue Gui for their feedback and support throughout this project. This work was supported by the Institute for AI and Beyond of the University of Tokyo, by the Japan Science and Technology Agency (JST) under the Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) program, Grant Number JPMJAP2401, and by NSF Grant #2411222.

References

- [1] Jorge Almazán, Joe McReynolds, Naoki Saito, and Studiolab. 2022. *Emergent Tokyo: Designing the Spontaneous City*.
- [2] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE, Los Angeles London.
- [3] Ruijia Chen, Junru Jiang, Pragati Maheshwary, Brianna R Cochran, and Yuhang Zhao. 2025. VisiMark: Characterizing and Augmenting Landmarks for People with Low Vision in Augmented Reality to Support Indoor Navigation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3706598.3713847
- [4] Yi Fei Cheng, Hang Yin, Yukang Yan, Jan Gugenheimer, and David Lindlbauer. 2022. Towards Understanding Diminished Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3491102.3517452
- [5] Michelene T.H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. Eliciting Self-Explanations Improves Understanding. *Cognitive Science* 18, 3 (1994), 439–477. doi:10.1207/s15516709cog1803_3
- [6] Louisa Dahmani and Véronique D. Bohbot. 2020. Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports* 10, 1 (April 2020), 6310. doi:10.1038/s41598-020-62877-0
- [7] Arindam Dey, Mark Billinghurst, Robert W. Lindeman, and J. Edward Swan. 2018. A Systematic Review of 10 Years of Augmented Reality Usability Studies: 2005 to 2014. *Frontiers in Robotics and AI* 5 (April 2018). doi:10.3389/frobt.2018.00037
- [8] Jan Drewes, Sascha Feder, and Wolfgang Einhäuser. 2021. Gaze During Locomotion in Virtual Reality and the Real World. *Frontiers in Neuroscience* 15 (May 2021). doi:10.3389/fnins.2021.656913
- [9] Aaron L. Gardony, Tad T. Brunyé, Caroline R. Mahoney, and Holly A. Taylor. 2013. How Navigational Aids Impair Spatial Memory: Evidence for Divided Attention. *Spatial Cognition & Computation* 13, 4 (Oct. 2013), 319–350. doi:10.1080/13875868.2013.792821
- [10] Aaron L. Gardony, Tad T. Brunyé, and Holly A. Taylor. 2015. Navigational Aids and Spatial Memory Impairment: The Role of Divided Attention. *Spatial Cognition & Computation* 15, 4 (Oct. 2015), 246–284. doi:10.1080/13875868.2015.1059432
- [11] J Goodman, SA Brewster, and P Gray. 2005. How can we best use landmarks to support older people in navigation? *Behaviour & Information Technology* 24, 1 (Jan. 2005), 3–20. doi:10.1080/01449290512331319021
- [12] Google. 2026. Gemini 3.1 Pro Preview | Gemini API. <https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview>
- [13] Chuanxiuyue He and Mary Hegarty. 2020. How anxiety and growth mindset are linked to navigation ability: Impacts of exploration and GPS use. *Journal of Environmental Psychology* 71 (Oct. 2020), 101475. doi:10.1016/j.jenvp.2020.101475
- [14] Donald Hedeker and Robert D. Gibbons. 1994. A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics* 50, 4 (1994), 933–944. doi:10.2307/2533433
- [15] Donald Hedeker and Robert D. Gibbons. 2006. *Longitudinal data analysis*. John Wiley & Sons.
- [16] Mary Hegarty, Anthony E Richardson, Daniel R Montello, Kristin Lovelace, and Ilavanil Subbiah. 2002. Development of a self-report measure of environmental spatial ability. *Intelligence* 30, 5 (Sept. 2002), 425–447. doi:10.1016/S0160-2896(02)00116-2
- [17] Lukáš Hejtmánek, Ivana Oravcová, Jiří Motýl, Jiří Horáček, and Iveta Fajnerová. 2018. Spatial knowledge impairment after GPS guided navigation: Eye-tracking study in a virtual town. *International Journal of Human-Computer Studies* 116 (Aug. 2018), 15–24. doi:10.1016/j.ijhcs.2018.04.006
- [18] Christopher Hilton and Jan Wiener. 2023. Route sequence knowledge supports the formation of cognitive maps. *Hippocampus* 33, 11 (2023), 1161–1170. doi:10.1002/hipo.23574
- [19] Giuseppe Iaria and Jason J. S. Barton. 2010. Developmental topographical disorientation: a newly discovered cognitive disorder. *Experimental Brain Research* 206, 2 (Oct. 2010), 189–196. doi:10.1007/s00221-010-2256-9
- [20] Toru Ishikawa. 2019. Satellite Navigation and Geospatial Awareness: Long-Term Effects of Using Navigation Tools on Wayfinding and Spatial Orientation. *The Professional Geographer* 71, 2 (April 2019), 197–209. doi:10.1080/00330124.2018.1479970 eprint: <https://doi.org/10.1080/00330124.2018.1479970>
- [21] Toru Ishikawa. 2023. Individual Differences and Skill Training in Cognitive Mapping: How and Why People Differ. *Topics in Cognitive Science* 15, 1 (Jan. 2023), 163–186. doi:10.1111/tops.12605
- [22] Toru Ishikawa and Daniel R. Montello. 2006. Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology* 52, 2 (March 2006), 93–129. doi:10.1016/j.cogpsych.2005.08.003
- [23] Toru Ishikawa and Uiko Nakamura. 2012. Landmark Selection in the Environment: Relationships with Object Characteristics and Sense of Direction. *Spatial Cognition & Computation* 12, 1 (Jan. 2012), 1–22. doi:10.1080/13875868.2011.581773
- [24] Toru Ishikawa and Yiren Zhou. 2020. Improving cognitive mapping by training for people with a poor sense of direction. *Cognitive Research: Principles and Implications* 5, 1 (Aug. 2020), 39. doi:10.1186/s41235-020-00238-1
- [25] Kyungwan Kim and Otmar Bock. 2021. Acquisition of landmark, route, and survey knowledge in a wayfinding task: in stages or in parallel? *Psychological Research* 85, 5 (July 2021), 2098–2106. doi:10.1007/s00426-020-01384-3
- [26] Alexander Klippel and Stephan Winter. 2005. Structural Salience of Landmarks for Route Directions. In *Spatial Information Theory*. Vol. 3693. Springer Berlin Heidelberg, Berlin, Heidelberg, 347–362. doi:10.1007/11556114_22
- [27] Matthias Kümmeler, Harneet Singh Khanuja, and Matthias Bethge. 2025. Modeling Saliency Dataset Bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22077–22088.
- [28] Carol A. Lawton. 1994. Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles* 30, 11–12 (June 1994), 765–779. doi:10.1007/BF01544230
- [29] Bing Liu, Linfang Ding, and Liqui Meng. 2021. Spatial knowledge acquisition with virtual semantic landmarks in mixed reality-based indoor navigation. *Cartography and Geographic Information Science* 48, 4 (July 2021), 305–319. doi:10.1080/15230406.2021.1908171
- [30] Eleanor A. Maguire, Katherine Woollett, and Hugo J. Spiers. 2006. London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus* 16, 12 (2006), 1091–1101. doi:10.1002/hipo.20233
- [31] Hanspeter A. Mallot and Stephan Lancier. 2018. Place recognition from distant landmarks: human performance and maximum likelihood model. *Biological Cybernetics* 112, 4 (Aug. 2018), 291–303. doi:10.1007/s00422-018-0751-4
- [32] Bartosz Mazurkiewicz, Marcelo de Lima Galvão, and Ioannis Giannopoulos. 2023. BeeAR: Augmented Reality Beeline Navigation for Spatial Knowledge Acquisition. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI (Sept. 2023), 199:1–199:17. doi:10.1145/3604246
- [33] Tobias Meilinger, Alain Berthoz, and Jan M. Wiener. 2011. The integration of spatial information across different viewpoints. *Memory & Cognition* 39, 6 (Aug. 2011), 1042–1054. doi:10.3758/s13421-011-0088-x
- [34] Alexandra Millonig and Katja Schechner. 2007. Developing Landmark-Based Pedestrian-Navigation Systems. *Trans. Intell. Transport. Syst.* 8, 1 (March 2007), 43–49. doi:10.1109/TITS.2006.889439
- [35] Daniel R. Montello. 1998. A new framework for understanding the acquisition of spatial knowledge in large-scale environments. *Spatial and temporal reasoning in geographic information systems* (1998), 143–154.
- [36] Daniel R. Montello and Danqing Xiao. 2011. Linguistic and Cultural Universality of the Concept of Sense-of-Direction. In *Spatial Information Theory*, Max Egenhofer, Nicholas Giudice, Reinhard Moratz, and Michael Worboys (Eds.). Springer, Berlin, Heidelberg, 264–282. doi:10.1007/978-3-642-23196-4_15
- [37] John O'Keefe and Lynn Nadel. 1978. *The hippocampus as a cognitive map*. Oxford university press.
- [38] Zhiwen Qiu, Mojtaba Ashour, Xiaohu Zhou, and Saleh Kalantari. 2024. Nav-MarkAR: A landmark-based augmented reality (AR) wayfinding system for enhancing older Adults' spatial learning. *Advanced Engineering Informatics* 62 (Oct. 2024), 102635. doi:10.1016/j.aei.2024.102635
- [39] Jinmeng Rao, Song Gao, Yuhao Kang, and Qingyun Du. 2020. Landmarks as Beacons: Pedestrian Navigation Based on Landmark Detection and Mobile Augmented Reality. In *Proceedings of AutoCarto*.
- [40] Ian T. Ruginski, Sarah H. Creem-Regehr, Jeanine K. Stefanucci, and Elizabeth Cashdan. 2019. GPS use negatively affects environmental learning through spatial transformation abilities. *Journal of Environmental Psychology* 64 (Aug. 2019), 12–20. doi:10.1016/j.jenvp.2019.05.001
- [41] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications (ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. doi:10.1145/355017.355028
- [42] Alexander W. Siegel and Sheldon H. White. 1975. The development of spatial representations of large-scale environments. *Advances in child development and behavior* 10 (1975), 9–55. <https://www.sciencedirect.com/science/article/pii/S0065240708600075>
- [43] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Ouab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jegou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. doi:10.48550/arXiv.2508.10104 arXiv:2508.10104 [cs].
- [44] Molly E. Sorrows and Stephen C. Hirtle. 1999. The Nature of Landmarks for Real and Electronic Spaces. In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*. Vol. 1661. Springer Berlin Heidelberg, Berlin, Heidelberg, 37–50. doi:10.1007/3-540-48384-5_3
- [45] Sibylle D. Steck and Hanspeter A. Mallot. 2000. The Role of Global and Local Landmarks in Virtual Environment Navigation. *Presence: Teleoperators and Virtual Environments* 9, 1 (Feb. 2000), 69–83. doi:10.1162/105474600566628

- [46] Marianne Strickrodt, Mary O'Malley, and Jan M. Wiener. 2015. This Place Looks Familiar—How Navigators Distinguish Places with Ambiguous Landmark Objects When Learning Novel Routes. *Frontiers in Psychology* 6 (Dec. 2015). doi:10.3389/fpsyg.2015.01936
- [47] W. R. Tobler. 1994. Bidimensional Regression. *Geographical Analysis* 26, 3 (1994), 187–212. doi:10.1111/j.1538-4632.1994.tb00320.x
- [48] Edward C. Tolman. 1948. Cognitive maps in rats and men. *Psychological Review* 55, 4 (1948), 189–208. doi:10.1037/h0061626
- [49] Alexis Topete, Chuanxiuyue He, John Protzko, Jonathan Schooler, and Mary Hegarty. 2024. How is GPS used? Understanding navigation system use and its relation to spatial ability. *Cognitive Research: Principles and Implications* 9, 1 (March 2024), 16. doi:10.1186/s41235-024-00545-x
- [50] David H. Uttal, Nathaniel G. Meadow, Elizabeth Tipton, Linda L. Hand, Alison R. Alden, Christopher Warren, and Nora S. Newcombe. 2013. The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin* 139, 2 (March 2013), 352–402. doi:10.1037/a0028446
- [51] David Waller and Yvonne Lippa. 2007. Landmarks as beacons and associative cues: Their role in route learning. *Memory & Cognition* 35, 5 (July 2007), 910–924. doi:10.3758/BF03193465
- [52] Steven M. Weisberg and Nora S. Newcombe. 2016. How do (some) people make a cognitive map? Routes, places, and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42, 5 (May 2016), 768–785. doi:10.1037/xlm0000200
- [53] Steven M. Weisberg, Victor R. Schinazi, Nora S. Newcombe, Thomas F. Shipley, and Russell A. Epstein. 2014. Variations in cognitive maps: Understanding individual differences in navigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40, 3 (May 2014), 669–682. doi:10.1037/a0035261
- [54] Wen Wen, Toru Ishikawa, and Takao Sato. 2013. Individual Differences in the Encoding Processes of Egocentric and Allocentric Survey Knowledge. *Cognitive Science* 37, 1 (2013), 176–192. doi:10.1111/cogs.12005
- [55] Wen Wen, Toru Ishikawa, and Takao Sato. 2014. Instruction of verbal and spatial strategies for the learning about large-scale spaces. *Learning and Individual Differences* 35 (Oct. 2014), 15–21. doi:10.1016/j.lindif.2014.06.005
- [56] Katharine S. Willis, Christoph Hölscher, Gregor Wilbertz, and Chao Li. 2009. A comparison of spatial knowledge acquisition with maps and mobile maps. *Computers, Environment and Urban Systems* 33, 2 (March 2009), 100–110. doi:10.1016/j.compenvurbsys.2009.01.004
- [57] Anna Wunderlich and Klaus Gramann. 2021. Landmark-based navigation instructions improve incidental spatial knowledge acquisition in real-world environments. *Journal of Environmental Psychology* 77 (Oct. 2021), 101677. doi:10.1016/j.jenvp.2021.101677
- [58] Jinyue Zhang, Xiaolu Xia, Ruiqi Liu, and Nan Li. 2021. Enhancing human indoor cognitive map development and wayfinding performance with immersive augmented reality-based navigation systems. *Advanced Engineering Informatics* 50 (Oct. 2021), 101432. doi:10.1016/j.aei.2021.101432
- [59] Yiyi Zhang and Tatsuo Nakajima. 2022. Exploring 3D Landmark-based Map Interface in AR Navigation System for City Exploration. In *Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia (MUM '21)*. Association for Computing Machinery, New York, NY, USA, 220–222. doi:10.1145/3490632.3497858
- [60] Yuhang Zhao, Elizabeth Kupferstein, Hathaitorn Rojnirun, Leah Findlater, and Shiri Azenkot. 2020. The Effectiveness of Visual and Audio Wayfinding Guidance on Smartglasses for People with Low Vision. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376516
- [61] Litao Zhu, Jie Shen, Jingyi Zhou, Zdeněk Stachoň, Shuai Hong, and Xing Wang. 2022. Personalized landmark adaptive visualization method for pedestrian navigation maps: Considering user familiarity. *Transactions in GIS* 26, 2 (2022), 669–690. doi:10.1111/tgis.12877

A Landmark Attention Study Materials & Additional Analysis

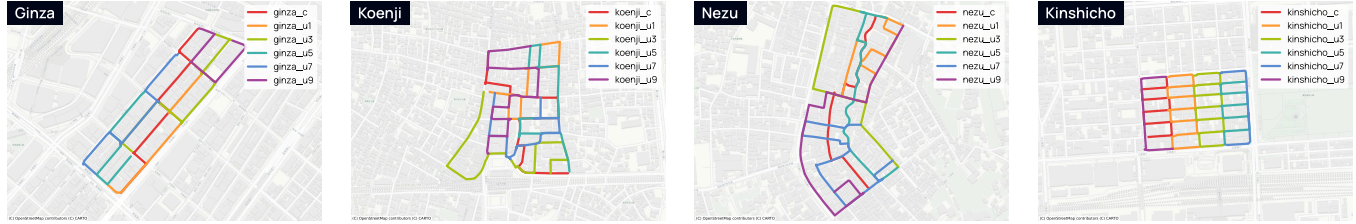


Figure 14: Route selections for landmark attention study. Each neighborhood included one common route (all 20 participants) and ten unique routes (one per participant).

PID	Age	Gender	SBSOD	Group	Languages	JP Read	R1	R2	R3	R4	R5	R6	R7	R8
A01	29	F	5.2	Good	Chinese, English	4	10	6	1	1	1	1	1	1
A02	26	M	5.3	Good	Chinese, English, Japanese	4	1	1	1	1	6	4	1	1
A03	28	M	5.2	Good	Chinese, Japanese, English, German	4	2	6	3	3	1	2	1	1
A04	23	M	6.9	Good	Chinese, English	5	1	1	1	1	1	1	5	5
A05	27	M	6.8	Good	Chinese, Japanese, English	5	2	2	1	1	1	1	1	1
A06	29	M	5.5	Good	Chinese, English	3	2	3	3	3	5	5	2	1
A07	25	M	5.7	Good	Chinese, English, Japanese	5	2	2	2	2	1	1	1	1
A08	30	M	5.7	Good	French, English	2	1	2	1	1	1	1	6	6
A09	29	F	6.6	Good	Chinese, English, Japanese	5	4	5	1	1	1	1	1	1
A10	23	M	6.7	Good	English, Hindi	3	4	1	1	1	5	1	1	1
A11	28	F	3.4	Poor	Chinese, English	3	6	7	3	2	2	2	2	2
A12	27	M	2.9	Poor	Japanese, English	5	1	1	1	1	8	8	1	1
A13	25	F	2.1	Poor	Chinese, Japanese, English	4	5	3	1	1	1	1	1	1
A14	39	F	3.4	Poor	Swedish, English, Japanese, German, French	3	3	2	10	10	3	3	8	8
A15	27	M	3.3	Poor	English, Hindi	5	4	3	2	2	8	7	1	1
A16	30	F	2.8	Poor	Chinese, English, Japanese	4	1	1	1	1	2	1	1	1
A17	28	M	3.4	Poor	English	2	1	1	3	4	1	1	1	1
A18	28	M	2.3	Poor	Chinese, English	2	3	1	1	1	1	1	3	2
A19	25	F	2.3	Poor	English	2	3	2	1	1	1	1	2	1
A20	25	F	2.9	Poor	Chinese, English	4	1	2	1	1	5	4	6	1

Table 3: Participant demographics, spatial ability (SBSOD), and self-reported familiarity with the route walked in each position R1–R8, separated into good and poor navigator groups. Route positions map to specific routes in Table 4; familiarity was rated from 1 (never been here before) to 10 (walk here regularly). Japanese reading levels range from 1 (not at all) to 5 (very well).

PIDs	R1	R2	R3	R4	R5	R6	R7	R8
A01/A11	ginza_c	ginza_u1	koenji_u1	koenji_c	nezu_c	nezu_u1	kinshicho_u1	kinshicho_c
A02/A12	koenji_u2	koenji_c	kinshicho_c	kinshicho_u2	ginza_u2	ginza_c	nezu_c	nezu_u2
A03/A13	nezu_c	nezu_u3	ginza_u3	ginza_c	kinshicho_c	kinshicho_u3	koenji_u3	koenji_c
A04/A14	kinshicho_u4	kinshicho_c	nezu_c	nezu_u4	koenji_u4	koenji_c	ginza_c	ginza_u4
A05/A15	ginza_u5	ginza_c	koenji_c	koenji_u5	nezu_u5	nezu_c	kinshicho_c	kinshicho_u5
A06/A16	koenji_c	koenji_u6	kinshicho_u6	kinshicho_c	ginza_c	ginza_u6	nezu_u6	nezu_c
A07/A17	nezu_u7	nezu_c	ginza_c	ginza_u7	kinshicho_u7	kinshicho_c	koenji_c	koenji_u7
A08/A18	kinshicho_c	kinshicho_u8	nezu_u8	nezu_c	koenji_c	koenji_u8	ginza_u8	ginza_c
A09/A19	ginza_u9	ginza_c	koenji_c	koenji_u9	nezu_u9	nezu_c	kinshicho_c	kinshicho_u9
A10/A20	koenji_c	koenji_u10	kinshicho_u10	kinshicho_c	ginza_c	ginza_u10	nezu_u10	nezu_c

Table 4: Route presentation order (R1–R8) for each of the ten counterbalanced orders. Route labels match the legends in Figure 14. Each order was walked by one good navigator (A01–A10) and one poor navigator (A11–A20). Neighborhoods were presented in blocks of two routes, alternating which of the pair was the common route.



Figure 15: Gaze fixation centers for three frames with the highest within-group dispersion among good navigators. Each dot represents one participant’s primary fixation location. Teal dots represent good SOD participants; pink dots represent poor SOD participants.

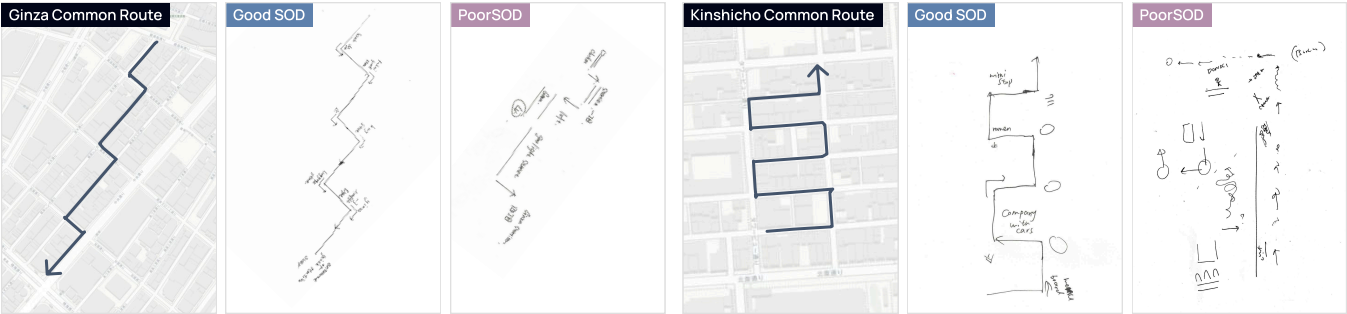


Figure 16: Example sketch maps from landmark attention study for two common routes. Each triplet shows the actual route geometry, a sketch from a good SOD participant, and a sketch from a poor SOD participant. Good SOD participants’ sketches tend to preserve route geometry and relative spatial relationships.

Category	Tag	Example	Good	%	Poor	%
Buildings & Structures	non-residential buildings	restaurants, shops, bars, salon, school, clinic, bank, senior home, community center, gallery, post office	591	58.6	608	52.0
	residential buildings	apartments, single family house	46	4.6	63	5.4
	convenience store	7-11, FamilyMart, Lawson, Mini Stop	17	1.7	29	2.5
	parts of a building	roof, stairs, wall, structurally necessary elements	6	0.6	29	2.5
	structures	Tokyo Tower, Tokyo Skytree	6	0.6	8	0.7
	items inside of a store/house	—	6	0.6	3	0.3
	decorative details	doorbell, potted plants, façade decoration, tiles, structurally unnecessary elements	5	0.5	41	3.5
Urban Infrastructure	road features	crosswalks, T-intersection, four-way intersection, Y split	47	4.7	25	2.1
	street typology	narrow, commercial, residential streets	42	4.2	74	6.3
	transit stations	train stations, bus stops	32	3.2	28	2.4
	parking lots	—	31	3.1	41	3.5
	construction sites	including empty plots of land	14	1.4	25	2.1
	parks	—	6	0.6	3	0.3
	paving	tiled streets, stone paving	3	0.3	9	0.8
	misc	highway, tunnel, pedestrian bridge, train tracks, visible from afar	4	0.4	5	0.4
Signage	signage (wayfinding)	street names, station names	59	5.8	21	1.8
	signage (store)	detached store signs, menus, pricing	21	2.1	23	2.0
	signage (advertisement)	campaign posters, festival posters	8	0.8	12	1.0
	signage (other)	generic signs, security camera, no-smoking	7	0.7	31	2.6
Street Objects	trees/plants	trees, non-potted plants	17	1.7	25	2.1
	vending machines	—	11	1.1	11	0.9
	street furniture	bulletin board, post box, traffic cone	9	0.9	7	0.6
	street art	graffiti	2	0.2	7	0.6
Transient Objects	parked cars	—	15	1.5	10	0.9
	parked bikes	—	3	0.3	10	0.9
	pedestrians	including pets	1	0.1	22	1.9
Total			1009		1170	

Table 5: Distribution of visual elements with example descriptions. Percentages indicate the proportion within each group.

B VLM Pipeline Prompts

B.1 Detection Prompt

You are a visual analyst for a VR navigation tool designed to help people with poor sense of direction.

You will receive a sequence of \$NUM_IMAGES street-level images, taken at intervals while walking along a city street. The images are named image1 through image\$NUM_IMAGES and are provided in walking order (image1 is earliest, image\$NUM_IMAGES is latest).

Each image shows a forward-facing view centered on the walking direction. The images may be cropped from wider panoramas, so the left and right edges represent the boundary of the visible area, not the full scene.

Your task is to identify and locate permanent, visually distinctive objects and landmarks that would help a person orient themselves and remember their route.

OBJECT CATEGORIES (ranked by baseline priority, highest first)

IMPORTANT - Relative distinctiveness:

The ranking reflects general importance, but actual priority depends on context. A lower-ranked object that is the ONLY distinctive feature in a scene is more valuable than a higher-ranked object surrounded by many strong cues.

1. Famous or obvious landmarks - named buildings, monuments, towers, sculptures, temples, shrines, bridges, or any structure a local would recognize by name.
2. Wayfinding signage - street names, maps, directional signs to districts or stations. Do NOT include generic traffic signs.
3. Store signage (permanent) - brand logos and signage attached to buildings. Do NOT include temporary signs or banners.
4. Convenience stores (konbini) - 7-Eleven, FamilyMart, Lawson, etc.
5. Parks and greenspaces - park entrances, gardens.
6. Transit stations and infrastructure - subway entrances, bus stops, station facades.
7. Street furniture - distinctive benches, fountains, clocks, mailboxes.
8. Vending machines - especially clusters or prominent placements.
9. Distinctive non-residential buildings - unusual architecture or strong colors.
10. Other urban infrastructure - overpasses, tunnels, ramps, retaining walls.

OBJECTS TO EXCLUDE

- Transient objects: cars, bikes, pedestrians, animals.
- Generic regulatory signage: stop signs, speed limits, traffic lights.
- Safety and utility signage: fire hydrant signs, AED signs, pole labels.
- Public-order signs on poles: no-smoking, anti-littering, etc.
- Temporary signage: ads, political posters, event banners.
- Temporary commercial signage not attached to buildings.
- Decorative micro-features: small details only visible up close.
- Any object likely to be moved or removed.

DETECTION RULES

Rule 1 - Per-frame cap:

- At most 2 objects per image.
- If \$NUM_IMAGES = 1 (intersection), you may highlight up to 3 objects.

Rule 2 - Directional bias:

- Prefer objects near the center (walking direction).
- Only select edge objects if:
 - * They are major landmarks, OR
 - * Highly distinctive, OR
 - * The only good candidates.

SCORING AND CLASSIFICATION

For each object, provide:

category (1-10): category number

category_label: one of:

"famous landmark", "wayfinding signage", "store signage", "konbini",
 "park/greenspace", "transit station", "street furniture",
 "vending machine", "distinctive building", "urban infrastructure"

color_prominence (0.0-1.0):

- How much the color stands out from surroundings

shape_prominence (0.0-1.0):

- How much the shape stands out from surroundings

semantic_familiarity (0.0-1.0):

- How easy it is to name or describe

All scores should be rounded to 2 decimal places.

BOUNDING BOX FORMAT

Use: [y_min, x_min, y_max, x_max] in range 0-1000.

IMPORTANT - Box the MOST DISTINCTIVE FEATURE, not the entire object.

Examples:

- Building with mural -> box the mural
- Shrine -> box the gate
- Store -> box the logo/sign
- Station -> box the sign

Only box the whole object if the whole object is the distinctive feature.

OUTPUT FORMAT

Return ONLY valid JSON with this structure:

```
{
  "image1": [
    {
      "bbox": [y_min, x_min, y_max, x_max],
      "description": "Describe the object and why it is useful for navigation.",
      "category": 4,
      "category_label": "konbini",
      "color_prominence": 0.75,
      "shape_prominence": 0.40,
      "semantic_familiarity": 0.95,
    }
  ]
}
```

```

    "permanence": 0.80,
    "navigational_context": "Explain its role in navigation."
  }
],
"image2": [],
"image3": []
}

```

Notes:

- Every image key must be present.
- Each image has 0-2 detections (or up to 3 if single image).
- The same real-world object can appear at most twice across all images.
- description: 1-2 sentences (what + why useful)
- navigational_context: 1-2 sentences (where + role in route)

B.2 Validation Prompt

You are a visual analyst validating object detections from a VR navigation tool.

You will receive a single cropped image from a 360-degree panoramic street photo.
A translucent green rectangle is drawn on the image to indicate where the object was detected.
The surrounding area provides additional context.

A previous detection pass produced the following for this region:

- Category: \$CATEGORY
- Description: \$DESCRIPTION

Your task is to use all three signals - the category, the description, and the bounding box location - together to determine whether the detection is valid.
The described object does not need to be strictly inside the green rectangle;
it is fine if the object is nearby or partially outside the box.
The bounding box is an approximate guide, not a precise boundary.

----- VALIDATION CRITERIA

The detection is VALID if ALL of the following are true:

1. The described object is visible in or near the crop.
The object mentioned in the description must be identifiable somewhere in the image.
It does not need to be entirely within the green rectangle - partial overlap or proximity is fine.
2. The category, description, and bounding box point to the same real-world object.
Minor inaccuracies in the description are acceptable as long as the object is recognizable.
Reject only for a fundamental mismatch (e.g., "red torii gate" but the image shows a vending machine).
3. The object is permanent.
It should be a fixed, long-term feature of the environment - not a temporary sign, parked vehicle, pedestrian, seasonal decoration, or movable item.
4. The object is visually distinctive enough to serve as a navigation landmark.
It should be something a person could recognize and describe.
Generic objects (e.g., a plain concrete wall, a generic utility pole) are not valid.

The detection is INVALID if ANY of the following are true:

- The crop is empty, blurry, or too dark to identify any object.

- The described object is not visible anywhere in the image.
- The category, description, and bounding box are fundamentally contradictory.
- The object is transient (vehicle, person, temporary sign, etc.).
- The object is too generic or unremarkable to serve as a navigation aid.
- The object is generic street furniture (e.g., street light poles, traffic lights, utility poles, guardrails, bollards, metal fences, manhole covers, drainage grates, convex traffic mirrors, or similar infrastructure).
- The object is a generic regulatory or public-order sign (e.g., stop signs, yield signs, traffic signs, no-smoking signs, pedestrian signs).
- The object is a temporary banner or sign attached to a pole (e.g., event banners, promotional flags, shopping district pennants).
- The crop captures only a small, unrecognizable fragment of the object.
- The detection highlights a generic building facade with no distinctive feature.

OUTPUT FORMAT

Return ONLY a valid JSON object with no additional text:

```
{
  "is_valid": true or false,
  "reason_is_valid": "One or two sentences explaining your judgment. If valid, briefly confirm what you see. If invalid, explain what is wrong."
}
```

C User Evaluation Materials & Additional Analysis

C.1 Post-Route Questionnaire

Participants rated the following statements on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree):

- (1) I am confident in my scene recognition responses.
- (2) I am confident in my intersection recall responses.
- (3) I am confident that my mental map accurately represents the route.
- (4) I feel I could give directions to someone visiting this area for the first time.
- (5) I felt aware of my surroundings during the route.
- (6) I feel I could find my way back to the starting point without turn-by-turn directions.
- (7) Navigating the route required a significant amount of mental effort.

Participants also rated their prior familiarity with the route on a scale from 1 to 10, where 1 indicated “*completely new to me; I have never walked here before*” and 10 indicated “*I know this route very well; I walk here regularly.*”

C.2 Post-Study Questionnaire

Participants were first asked: “*Which route would you remember better if you had to navigate it again tomorrow?*” and selected either Route A or Route B.

They then rated the following statements on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree):

- (1) The circle highlights were easy to notice.
- (2) The circle highlights were distracting.
- (3) The arrows helped me find highlights I had not yet noticed.
- (4) Dismissing the circles by looking at them felt natural.
- (5) The highlighting system improved my navigation experience.
- (6) I learned how to interact with the highlighting system very quickly.
- (7) I would use such a highlighting system with AR glasses in real life.

C.3 Participant Table

PID	Age	Gender	SBSOD	Languages Fluent In	Japanese Reading	Route 1 Familiarity	Route 2 Familiarity
E1	30	F	2.8	Chinese, English, Japanese	Quite well	Never been there	Never been there
E2	28	M	2.3	Chinese, English	Moderately	Never been there	Never been there
E3	25	F	2.9	Chinese, English	Quite well	Never been there	Never been there
E4	31	M	2.3	Chinese, English	Moderately	Never been there	Never been there
E5	28	F	3.4	English, Chinese	Moderately	Never been there	Never been there
E6	25	F	3.1	Chinese, Japanese, English	Very well	Never been there	Never been there
E7	27	M	2.9	Japanese, English	Very well	Never been there	Never been there
E8	25	F	2.3	English	Slightly	Never been there	Never been there

Table 6: Participant demographics, SBSOD scores, and self-reported familiarity with the two study routes.

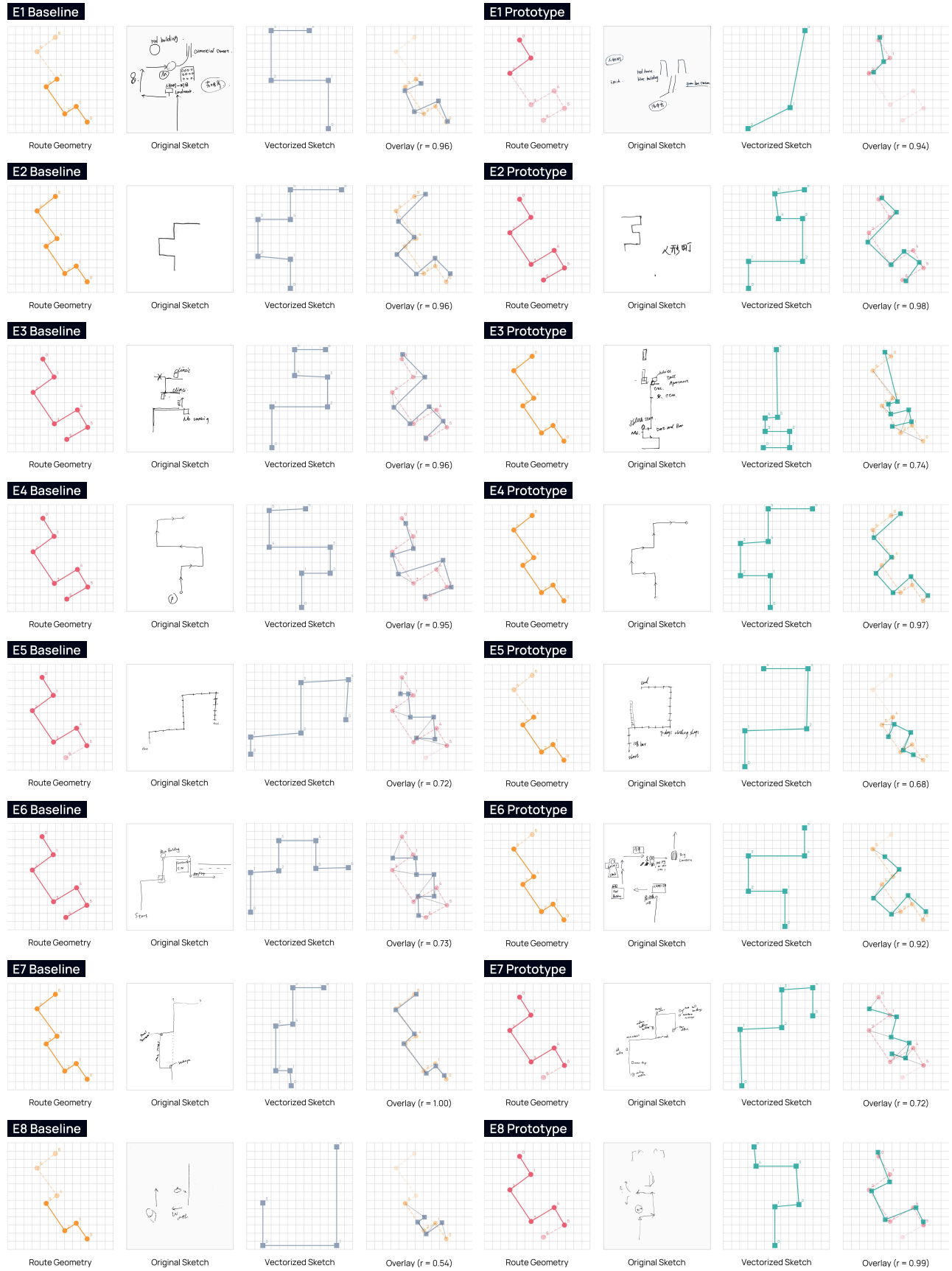


Figure 17: Map sketching task results. For each condition, columns show: the actual route geometry, the participant's original hand-drawn sketch, the vectorized representation of the sketch, and the overlay of the vectorized sketch on the route geometry with the bidimensional correlation coefficient (r). Higher r values indicate greater spatial correspondence between the sketched and actual route geometries.