BEYOND SUCCESS: ASSESSING HUMAN-PLAUSIBLE EXECUTION IN VISION-AND-LANGUAGE NAVIGATION WITH HUMANPATHSCORE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-and-language navigation (VLN) benchmarks measure whether an agent reaches its goal, how efficiently, and how closely its route follows a reference. These essential measures do not fully characterize execution. A successful trajectory may still intersect occupied space, turn abruptly or oscillate, undo progress, or lack visual evidence for an upcoming action. We introduce HumanPathScore (HPS), a reference-trajectory-free, scene-evidence-driven metric that assesses a single trajectory along four observable diagnostic dimensions of human-plausible execution: Physical Walkability, Motion Naturalness, Behavioral Coherence, and View for Action. Together, these constructs operationalize human-plausible execution by measuring whether a human-sized body could traverse the route, movement remains controlled, progress is preserved, and scene evidence supports action. Each construct and its evidence remain visible beside an aggregate that excludes task success. A common trajectory-evidence formulation handles graph-based and continuous trajectories and supports calibrated observations from both mesh-based and 3D Gaussian Splatting reconstructions. We validate HPS in discrete and continuous navigation using blinded human video judgments and matched metric comparisons. HPS shows the strongest overall alignment with human preference among the compared metrics. A separate multi-benchmark analysis shows that HPS yields within-setting method orderings distinct from conventional task and reference metrics. These results support HPS as a diagnostic complement to task-performance and reference-fidelity metrics: existing metrics assess whether an agent succeeds, whereas HPS directly assesses how the execution unfolds.

	Task performance			Human-plausible execution				Evidence used			Applicability		
Metric family	Comp.	Eff.	Fidelity	Physical	Motion	Behavior	View	Human	Mesh	3DGS	Disc.	Cont.	Object
SR/SPL (Anderson et al., 2018a)	✓	✓	–	–	–	–	–	–	–	–	✓	✓	✓
CLS/nDTW/SDTW (Jain et al., 2019)	△	△	✓	–	–	△	–	–	–	–	✓	✓	△
RGS/RGSPL (Qi et al., 2020)	✓	△	–	–	–	–	–	–	–	–	✓	△	✓
OmniNav C/E (Sun et al., 2026)	✓	✓	–	–	–	–	–	–	–	–	–	✓	△
SAGE-CSR (Miao et al., 2025)	△	–	✓	–	–	△	–	–	–	–	△	✓	△
SAGE-ICP (Miao et al., 2025)	–	–	–	△	–	–	–	–	△	–	–	✓	△
SAGE-PS (Miao et al., 2025)	–	–	–	–	△	–	–	–	–	–	△	✓	△
NaviTrace (Windecker et al., 2026)	–	–	✓	△	–	△	△	△	–	–	–	–	–
NTT (Devlin et al., 2021)	–	–	–	△	△	✓	△	✓	–	–	△	✓	△
MMD (Colbert & Saeedi, 2022)	–	–	–	△	△	△	–	✓	–	–	△	✓	△
HumanPathScore (ours)	–	–	–	✓	✓	✓	✓	–	✓	△	✓	✓	✓

Table 1: Scope of representative VLN evaluation families. The four execution columns cover physical walkability, motion naturalness, behavioral coherence, and view for action; Human, Mesh, and 3DGS identify evidence used for scoring. A ✓ marks direct support, △ partial or variant-dependent support, and – no support. Existing metrics measure task outcomes or selected aspects of execution, whereas HPS is designed to cover all four dimensions across discrete, continuous, and object-oriented settings. The table summarizes evaluation scope; empirical human alignment is evaluated separately in Tables 2 and 3.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

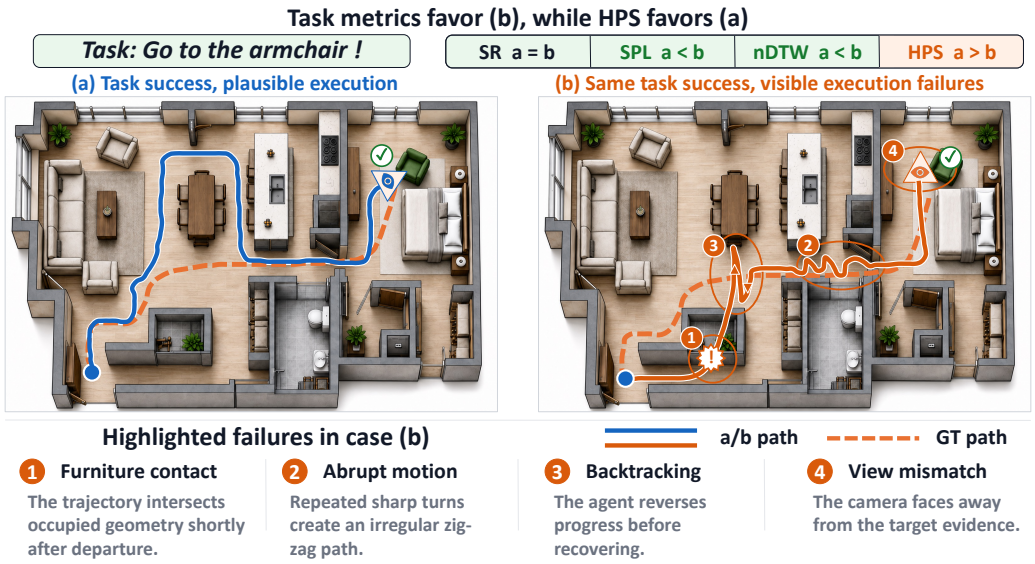

Figure 1: Conceptual illustration of metric complementarity. Both routes meet the endpoint success criterion, so SR remains equal despite (b)'s intermediate body–scene intrusion. Route (b)'s shorter length and closer reference agreement favor SPL and nDTW, respectively. HPS favors (a) by diagnosing the body intrusion, abrupt turns, backtracking, and limited visual support highlighted in (b).

1 Introduction

Vision-and-language navigation (VLN) asks an embodied agent to translate an instruction into goal-directed behavior in a complex environment. Standard benchmarks make this capability measurable through task completion, path efficiency, reference fidelity, and task-specific grounding. Measures such as SR, SPL, nDTW, and RGS are indispensable: they establish whether an agent solved the assigned task and have driven sustained progress in VLN (Anderson et al., 2018b;a; Jain et al., 2019; Qi et al., 2020). Their target, however, is task performance rather than the quality of the executed process.

Task success, path efficiency, and reference agreement describe whether an agent reaches the goal, how efficiently it moves, and how closely it follows a reference. In human-built spaces, these measures leave body-scale traversability, controlled motion, purposeful progress, and action-supporting visibility unmeasured. HPS complements them with an observable execution target. Figure 1 illustrates this distinction: both routes satisfy the endpoint criterion, while their execution quality differs.

Recent evaluation broadens this view through unified navigation benchmarks and 3DGS-based executability metrics (Sun et al., 2026; Miao et al., 2025), yet these advances remain organized around task-specific outcomes or partial execution properties. Human-plausible execution still lacks a unified, interpretable assessment of a single trajectory from its instruction and calibrated scene evidence, without requiring a ground-truth route or separately collected human-trajectory dataset at scoring time (Table 1).

We define *human-plausible execution* through four observable requirements on an individual trajectory: accommodating a human-sized body, keeping local direction changes controlled, preserving purposeful progress, and exposing useful route or instruction evidence. Among successful routes through human-built environments, these requirements identify improvements in clearance, turn regularity, retracing, and visual opportunities.

To assess these requirements, we introduce HumanPathScore (HPS), a reference-trajectory-free, scene-evidence-driven metric that separates Physical Walkability, Motion Naturalness, Behavioral Coherence, and View for Action. HPS reports each construct and its attributed evidence beside an aggregate score, allowing a low score to be diagnosed rather than hidden in a single prediction. Task success remains outside HPS. The scoring interface uses only the executed trajectory, instruction, and calibrated scene observations, making the same audit available across multiple settings.

We validate HPS with blinded human judgments in discrete and continuous navigation, then compare method rankings across five navigation settings. HPS aligns strongly with human preference and complements task-performance and reference-fidelity metrics.

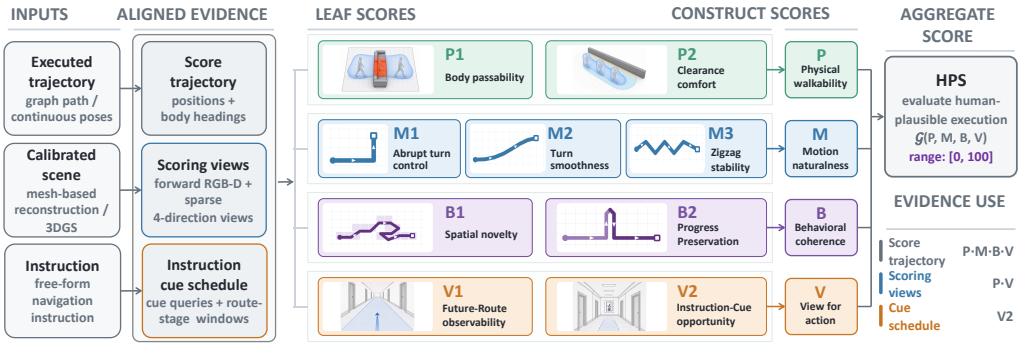

Figure 2: HumanPathScore provides a unified and auditable pipeline for evaluating human-plausible execution. A common input contract aligns each executed trajectory with calibrated scene and instruction evidence across graph-based and continuous navigation, while preserving the provenance of all nine leaf-level diagnoses. The resulting P, M, B, V profile remains directly inspectable as the primary diagnostic output; its equal-exponent geometric mean provides a compact HPS summary, with task success reported separately.

Our main contributions are as follows:

- We formulate human-plausible execution as a distinct single-trajectory evaluation target, characterized by observable diagnostic dimensions beyond task success and human imitation.
- We introduce HumanPathScore (HPS), a reference-trajectory-free and interpretable metric that evaluates VLN agent trajectories from calibrated scene evidence while preserving construct-level diagnostics.
- We validate HPS against blinded human judgments in discrete and continuous navigation and compare method rankings across five settings, demonstrating its complementary value to conventional VLN evaluation.

2 Related Work

Task-Oriented Evaluation. VLN spans graph-based route following in R2R and RxR, continuous control in VLN-CE, and grounded targets in REVERIE and ObjectNav (MacMahon et al., 2006; Chen & Mooney, 2011; Anderson et al., 2018b; Chang et al., 2017; Ku et al., 2020; Krantz et al., 2020; Qi et al., 2020; Batra et al., 2020). Interactive, long-horizon, outdoor, and sim-to-real variants further change the agent and evaluation contract (Thomason et al., 2020; Shridhar et al., 2020; Padmakumar et al., 2022; Chen et al., 2019; Anderson et al., 2021; Sun et al., 2026). Across these settings, SR and SPL measure completion and efficiency, whereas CLS, nDTW, and SDTW measure reference-path fidelity (Anderson et al., 2018a; Jain et al., 2019; Ilharco et al., 2019). Object grounding, progress, timing, and instruction-error extensions remain task-specific (Qi et al., 2020; Song et al., 2025; Yokoyama et al., 2021; Taioli et al., 2024; Qiao et al., 2025). HPS evaluates the shared artifact—an executed trajectory—and reports comparisons within each navigation setting.

Process-Aware Evaluation. Recent metrics expose selected properties of execution. SAGE-Bench measures reference-corridor success, collision persistence, and path smoothness in executable 3DGS environments, while NaviTrace evaluates path, endpoint, and semantic consistency from image-space traces (Miao et al., 2025; Windecker et al., 2026). These approaches either require a reference corridor or native collision state, or operate on a narrower trace representation. HPS complements them by using calibrated 3D scene evidence to diagnose a single executed trajectory at construct level.

Human-Centered Evaluation. Social and human-aware navigation evaluates behavior around nearby people, including comfort, legibility, and safety (Kruse et al., 2013; Rios-Martinez et al., 2015; Dragan et al., 2013; Helbing & Molnár, 1995; Francis et al., 2025; Dong et al., 2025; Yue et al., 2024). Human imitation addresses another target: Navigation Turing Tests learn from human traces, while distributional metrics compare human and agent trajectory populations (Devlin et al., 2021; Colbert & Saeedi, 2022). HPS evaluates single-trajectory plausibility from execution evidence, while human judgments serve as validation rather than scoring input.

3 HUMANPATHSCORE

HumanPathScore evaluates human-plausible execution at the trajectory level. It turns calibrated scene and trajectory evidence into interpretable diagnoses and an aggregate score (Fig. 2).

3.1 EVALUATION TARGET AND ALIGNED EVIDENCE

Given an executed trajectory τ, instruction x, and scene interface $\mathcal{E}$, HPS aligns graph paths and continuous poses along traveled distance, while $\mathcal{E}$ provides calibrated RGB-D and camera poses. Graph paths retain piecewise-linear edges and source anchors; VLN-CE uses recorded continuous positions. A standardized observer aligns forward views with local travel direction; sparse directional views are used only to confirm instruction cues.

Mesh-based reconstructions, including official GLB assets, and 3D Gaussian Splatting scenes use this same interface (Kerbl et al., 2023; Li et al., 2025). Depth supports route geometry, color supports instruction cues, and the trajectory provides motion and route-history evidence. HPS scores execution separately from task success, reference paths, human demonstrations, learned human-likeness models, and native simulator collision state; inputs that fail scene-evidence checks remain unscored.

3.2 CONSTRUCT DECOMPOSITION AND SCORING CONVENTION

Human navigation combines perception, movement, and memory (Wolbers & Hegarty, 2010). The constructs address different questions and may co-occur: a route can be smooth yet collide, passable yet retrace, or successful yet lack visual evidence. Their profile also captures trade-offs: rounding a corner may improve motion regularity while reducing clearance. The aggregate supports comparison and reward shaping; the profile explains the changes. Body-scaled interaction grounds P (Warren, 1984); route regularity grounds M (Hicheur et al., 2007); wayfinding grounds B (Wiener et al., 2009); anticipatory visual support grounds V (Hollands et al., 2002).

Diagnostic leaves score plausibility in $[0, 100]$, with higher values better. Physical, motion, and behavioral leaves combine absolute event burden with exposure relative to route length. Stronger or more extensive events lower the score for a fixed route; valid event-free evidence yields 100, and missing evidence remains undefined. Event ownership assigns each deficiency to one sibling leaf. View leaves average positive opportunities over route intervals or cue windows.

HPS uses the same equal-exponent geometric operator throughout the hierarchy. For n factors $s_1, \ldots, s_n$, we define

$$\mathcal{G}(s_1, \ldots, s_n) = 100 \left(\prod_{i=1}^{n} \frac{s_i}{100} \right)^{1/n}. \tag{1}$$

Equal exponents weight constructs equally in log space. The geometric mean limits compensation and preserves scale, summarizing the diagnostic profile for within-setting comparison. Zero factors yield zero HPS; the profile identifies the contributing deficiencies. This operator aggregates leaves within constructs and constructs into $\text{HPS} = \mathcal{G}(P, M, B, V)$.

3.3 PHYSICAL WALKABILITY

People judge whether passage is possible relative to their own body scale. Physical Walkability therefore asks whether the complete route accommodates a person and leaves reasonable surrounding space, rather than only checking waypoint legality.

To apply the same body-space test to mesh-based and 3DGS scenes, HPS fuses calibrated depth into a local truncated signed distance field (TSDF); a Euclidean signed distance field (ESDF) then provides distances to supported occupied geometry (Curless & Levoy, 1996; Oleynikova et al., 2017). Unobserved evidence remains unavailable rather than becoming free space.

Body Passability (P_1). A person cannot follow a route through furniture or walls. HPS sweeps a fixed human-sized region along the reconstructed route; persistent, spatially supported overlap inside this region lowers the score by severity and traveled extent, while local floor continuity prevents ordinary stairs from being treated as intrusion.

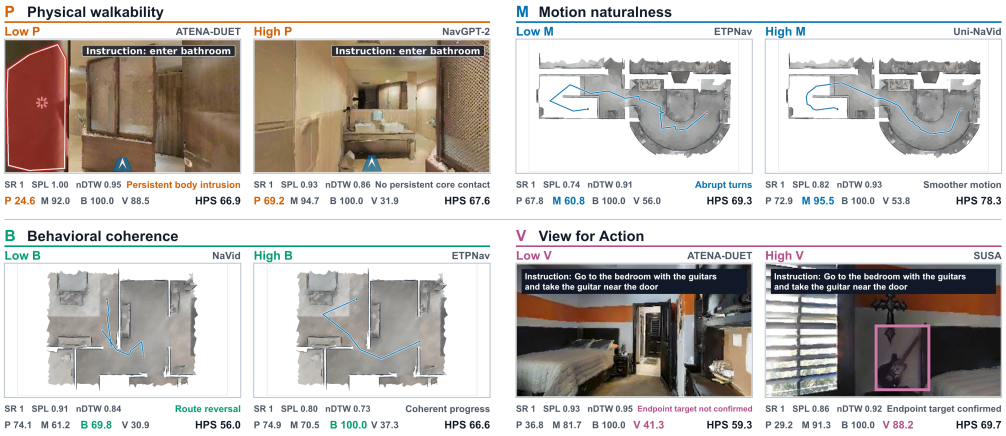

Figure 3: Construct-level HPS evidence. Low–high pairs localize body–scene intrusion, unnatural turns, route cancellation, and instruction-conditioned viewing opportunities. Scores use Eq. 1; view evidence is restricted to its valid trajectory window.

Clearance Comfort (P_2). People usually preserve some surrounding space even when a passage is traversable. HPS measures sustained nearby geometry in a nonoverlapping comfort region outside the body. Body Passability asks whether a person fits, whereas Clearance Comfort asks how confined the passage feels; a narrow route may lower comfort but cannot excuse body intrusion. Their aggregate is $P = \mathcal{G}(P_1, P_2)$.

3.4 Motion Naturalness

Goal-directed human walking forms regular path shapes even when the exact route is not prescribed. Because the same execution may contain sparse graph nodes or dense poses, HPS resamples each route by traveled distance and extracts signed turn events, separating path geometry from output density. Motion Naturalness therefore characterizes navigation-scale route geometry: how sharply direction changes, how a turn develops spatially, and whether alternating corrections persist.

Abrupt Turn Control (M_1). Human turns do not change direction instantaneously. At each source anchor, HPS measures the angle between the incoming and outgoing route segments; Abrupt Turn Control penalizes a large angle concentrated over little local travel.

Turn Smoothness (M_2). Within a genuine turn, direction should also change progressively. Turn Smoothness therefore examines each detected window on the distance-resampled route, favoring angular change spread across neighboring points over the same turn concentrated at one vertex.

Zigzag Stability (M_3). Purposeful walking generally stabilizes after a correction. HPS lowers this leaf for persistent opposite-signed turns while treating a brief correction as ordinary motion. Their aggregate is $M = \mathcal{G}(M_1, M_2, M_3)$.

3.5 Behavioral Coherence

Human wayfinding uses memory and spatial knowledge to organize route-level decisions. Because backtracking can be purposeful (Javadi et al., 2019), Behavioral Coherence targets persistent progress cancellation and redundant route reuse rather than brief corrections. A top-down memory excludes recent travel and assigns cancellation to Progress Preservation before sending the remaining nonrecent returns to Spatial Novelty.

Spatial Novelty (B_1). Human navigation generally expands into new space as it advances. HPS projects the trajectory onto a top-down grid and records visited cells. After excluding recent travel and Progress Preservation intervals, the distance and persistence of returns to visited cells lower the score.

Progress Preservation (B_2). Direct reversal or retracing undoes earlier movement. HPS lowers this leaf according to the persistence of these events, which remain excluded from Spatial Novelty. Their aggregate is $B = \mathcal{G}(B_1, B_2)$.

3.6 VIEW FOR ACTION

View for Action asks whether the upcoming route and a relevant target or landmark are visible when useful under the standardized trajectory observer.

Future-Route Observability (V_1). At each moving position, V_1 measures the contiguous route prefix visible within the fixed preview horizon, ending at the first frustum exit or wall/depth-supported occlusion. Partial preview receives partial credit, weighted by traveled distance; visibility loss therefore contributes in proportion to the affected interval. Stationary and terminal views are excluded. As the standardized observer advances, previously hidden segments are evaluated from later positions.

Instruction-Cue Opportunity (V_2). People use targets and landmarks when they become relevant to a decision. HPS extracts auditable phrases with fixed rules, assigns each a route-stage role, and uses GroundingDINO to confirm the corresponding cue only within that window (Liu et al., 2024). This evidence forms V_2.

An instruction without an auditable cue uses $V = V_1$; otherwise, $V = \mathcal{G}(V_1, V_2)$. Cue confirmations are averaged within roles and combined across roles to form V_2. A zero V_2 propagates to HPS; cue records identify absent confirmations. Render, detector-execution, depth, or alignment failures are recorded as invalid evidence.

HPS returns its aggregate, four constructs, nine leaves, localized evidence, and validity status. Figure 3 shows how these diagnoses expose failures unresolved by task metrics.

4 EXPERIMENTAL VALIDATION

4.1 EXPERIMENTAL SETUP

Stratified scalar ratings test whether scores track perceived plausibility across execution conditions; independently collected A/B preferences test whether they select the preferred successful execution for the same task. Shared target and landmark cues contextualize visual opportunities. Unless otherwise stated, human validation uses calibrated RGB-D rendered with Habitat-Sim (Savva et al., 2019) from the official Matterport3D (MP3D) mesh-based GLB scenes (Chang et al., 2017). All components and parameters were selected on separate development annotations and frozen before the reported scalar-rating and A/B studies; additional computation details and sensitivity analyses are provided in the supplement.

Scalar ratings. The cohort contains 771 trajectories—402 from R2R and 369 from R2R-CE—stratified by method, success status, path length, and dominant failure type. Stratification and displayed cues were defined independently of HPS scores; annotators were blinded to metric outputs and leaf evidence. All eight annotators independently rate every trajectory from 1 to 7 for Overall plausibility and P, M, B, V, answering Overall first and yielding 30,840 scalar ratings. They see the trajectory video and whitelisted target and landmark noun phrases, with full instructions and method identities withheld. Overall asks how plausible the execution would appear for a person navigating the scene. Spearman correlation uses each trajectory's mean rating.

Rating agreement. Human ratings showed moderate-to-good reliability. For Overall, P, M, B, and V, average-measure absolute-agreement ICCs were 0.80, 0.76, 0.74, 0.68, and 0.58 on R2R and 0.72, 0.70, 0.78, 0.63, and 0.52 on R2R-CE, respectively. The corresponding Overall noise ceilings were 0.89 and 0.85, exceeding the observed HPS–human correlations of 0.858 and 0.763.

A/B preferences. We construct 134 R2R and 123 R2R-CE pairs of successful trajectories from different methods sharing the same scene, instruction, and goal. Each pair receives four independent judgments of which execution better matches human navigation, totaling 1,028 choices. Agreement compares each metric's score ordering with each individual human choice; left–right order is randomized, and metric ties or human no-preference outcomes receive half credit.

Benchmark analysis. We compare SR, SPL, nDTW, and HPS rankings within R2R, RxR, REVERIE, R2R-CE, and RxR-CE (Anderson et al., 2018a; Jain et al., 2019; Ilharco et al., 2019). Methods include recurrent and graph-based agents (Chen et al., 2021; Hong et al., 2021; Chen et al., 2022; Ko et al., 2025; Wang et al., 2024; An et al., 2023; Zhou et al., 2024; Li et al., 2023; Zhang et al., 2026), continuous planners, and video agents (Hong et al., 2022; An et al., 2025; Yu et al., 2025; Ye et al., 2025; Zhang et al., 2024; 2025; Wei et al., 2025; Yao et al., 2025). HPS summaries

(a) Overall human alignment

Category	Metric	R2R		R2R-CE	
		ρ	A/B	ρ	A/B
Path efficiency	SPL	0.335	59.7%	0.575	68.3%
	SoftSPL	0.410	62.7%	0.664	74.8%
Reference fidelity	−Fréchet dist.	0.561	70.1%	0.360	72.4%
	nDTW	0.715	81.3%	0.726	75.6%
	CLS	0.640	75.4%	0.680	74.3%
Process quality	SAGE-PS	0.590	72.4%	0.610	71.5%
	HPS	**0.858**	**88.8%**	**0.763**	**78.9%**

Table 2: Overall human alignment on the complete R2R and R2R-CE cohorts. Spearman correlation uses scalar ratings; A/B reports agreement with separately collected preference judgments, with ties receiving half credit. Relative to nDTW, HPS improves correlation by 0.143 [0.104, 0.181] and 0.037 [0.006, 0.070], respectively; brackets are 95% paired cluster-bootstrap confidence intervals.

(b) Construct-level alignment

Human target	Metric	R2R ρ	R2R-CE ρ
Physical	HPS-P	**0.791**	**0.735**
	Navmesh Clearance	0.664	0.655
	Navmesh Collision	0.409	0.384
Motion	HPS-M	**0.715**	**0.857**
	SAGE-P	0.483	0.538
	SAGE-C	0.143	0.286
	TurnQ Proxy	0.567	0.683
Behavior	HPS-B	**0.837**	0.700
	CLS	0.743	0.602
	nDTW	0.789	**0.808**
	SAGE-CSR	0.460	0.321
View	HPS-V	**0.759**	0.788
	HPS-V_1 only	0.704	**0.812**
	HPS-V_2 only	0.657	0.726

Table 3: Construct-level Spearman correlation on matched trajectories. Navmesh measures are independent proxies, while HPS-V_1 and HPS-V_2 are internal variants. SAGE-P and SAGE-C follow the paper and code specifications, respectively; bold marks the strongest result for each target and setting.

use valid scores; trajectory sources, scoring schemas, backends, and missing reasons are retained. Coverage is reported separately below.

4.2 Overall Human Alignment

HPS shows the strongest overall alignment with human judgment among the compared metrics in both navigation settings. Figure 4 shows its relation to mean Overall plausibility; a random one-third of the rated trajectories is displayed for clarity, while every quantitative comparison uses the complete cohort.

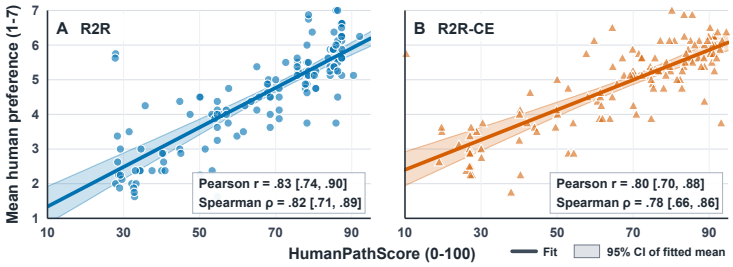

Figure 4: HPS versus mean Overall plausibility on R2R and R2R-CE. A random one-third of trajectories is shown for clarity; its Spearman correlation is within 0.05 of the complete-cohort estimate in both settings. Lines and bands show within-setting fits and 95% cluster-bootstrap confidence intervals.

Table 2 compares HPS with path-efficiency, reference-fidelity, and process metrics on the same trajectories, using both scalar-rating correlation and independently collected A/B preference. SPL combines success with path efficiency; SoftSPL uses continuous progress credit. Discrete Fréchet distance measures the order-aware deviation between the executed and reference trajectories; we negate it so that higher values consistently indicate better agreement.

With task and endpoint success fixed, HPS's higher A/B agreement supports the usefulness of execution-quality assessment among successful alternatives.

The construct results show selective rather than uniform advantages. Physical Walkability and Motion Naturalness lead their independent comparators in both settings, whereas Behavioral Coherence leads only on R2R. Full V leads its internal variants on R2R, while V_1 is slightly higher on R2R-CE,

possibly because continuous observations contain less visible or noisier cue evidence. We use internal variants here because existing VLN metrics do not directly evaluate View for Action. Behavioral and view-related alignment therefore remains setting-dependent.

4.3 Ablation Study

Table 4 asks whether each branch contributes to alignment with overall human ratings. Full HPS and every ablation are evaluated on the same 771 trajectories. Removing a branch re-normalizes the equal-exponent geometric mean over the remaining constructs; for example, $\mathrm{HPS}_{-P} = \mathcal{G}(M, B, V) = (MBV)^{1/3}$. We compute Spearman correlation against the mean human rating and estimate 95% confidence intervals with 10,000 paired cluster-bootstrap resamples.

Removing any branch reduces human alignment in both settings. Motion Naturalness produces the largest decrease, followed by Physical Walkability, indicating that how an agent moves and whether its route accommodates a person are particularly salient to overall judgment. View for Action and Behavioral Coherence produce smaller decreases, showing that they provide additional evidence beyond local geometry and motion. Additional ablations are provided in the supplement.

4.4 Method Ranking and Complementarity

Figure 5 compares task-metric and HPS rankings across five settings. HAMT leads HPS on R2R and RxR despite low conventional ranks; KERM rises from third under all three task metrics to first on REVERIE. In continuous navigation, DualVLN leads both nDTW and HPS on R2R-CE, while Uni-NaVid overtakes it under HPS on RxR-CE. ETP-R1 falls from the top three task-metric ranks to tenth and sixth under HPS, respectively. These orderings show how execution quality complements success, efficiency, and reference fidelity at method level.

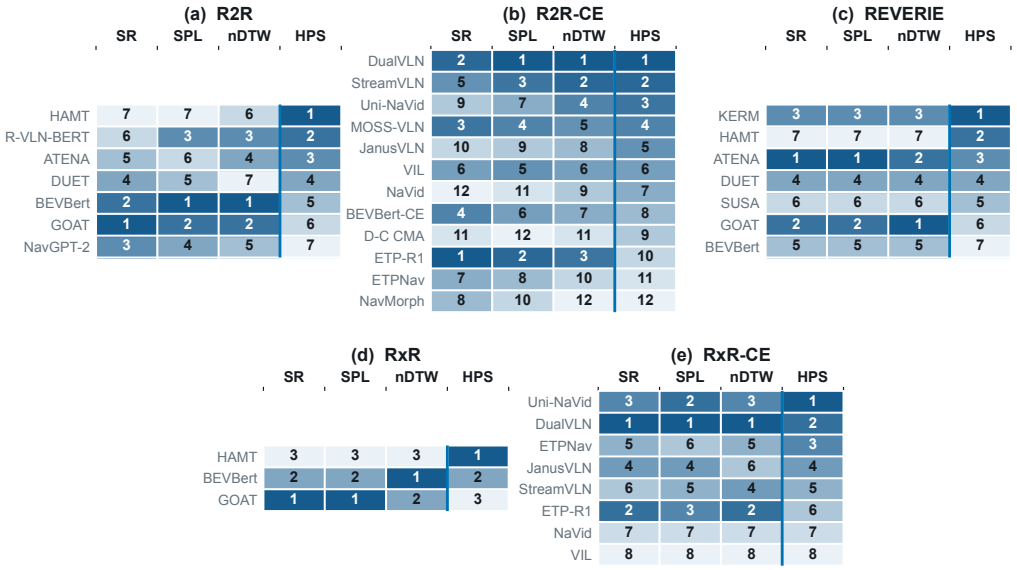

Figure 5: Within-setting method rankings under SR, SPL, nDTW, and HPS across five navigation settings. Rank 1 is best; darker cells indicate better ranks. HPS complements task success, path efficiency, and reference fidelity with execution-quality assessment.

4.5 Reward-Shaping Application

To test use beyond diagnosis, we add a bounded HPS reward to Recurrent VLN-BERT on R2R val-unseen while holding the architecture and evaluation fixed (Table 5).

Relative to nDTW shaping, HPS yields higher SR, SAGE-PS, and HPS point estimates while SPL and nDTW remain similar, suggesting its use as an auxiliary execution-quality signal. Training details are provided in the supplement.

(a) Branch ablation

Variant	R2R ρ [95% CI]	R2R-CE ρ [95% CI]
Full HPS	**0.858 [0.805, 0.897]**	**0.763 [0.695, 0.817]**
w/o P	0.793 [0.759, 0.826]	0.703 [0.669, 0.736]
w/o M	0.773 [0.740, 0.806]	0.685 [0.652, 0.717]
w/o B	0.833 [0.811, 0.852]	0.741 [0.720, 0.759]
w/o V	0.818 [0.791, 0.844]	0.725 [0.699, 0.750]

Table 4: Leave-one-branch-out alignment with mean Overall ratings on the same trajectories. Each entry reports the absolute Spearman correlation and its 95% paired cluster-bootstrap confidence interval after equal-exponent reaggregation.

(b) Reward shaping

	Task performance				Process quality	
Reward	NE (m)↓	SR↑	SPL↑	nDTW↑	SAGE-PS↑	HPS↑
Goal only	4.2810	59.87	54.10	63.74	0.103	72.85
+nDTW	4.1455	61.09	**55.50**	**65.08**	0.110	73.46
+nDTW+HPS	**4.1222**	**61.89**	55.40	65.03	**0.122**	**74.68**

Table 5: Reward shaping with Recurrent VLN-BERT on R2R val-unseen. NE denotes Navigation Error in meters; bold marks the best point estimate in each column.

4.6 Coverage, Backend Sensitivity, and Cost

On 67,863 full-split trajectories from R2R, R2R-CE, and English-only RxR-CE (27 dataset–method combinations), HPS returns 98.2% finite scores. Coverage ranges from 97.5% to 98.8% across these settings; remaining cases lack sufficient geometric or visual evidence.

On the same R2R-CE trajectories, human alignment remains positive with matched 3DGS observations ($\rho = 0.725$, versus 0.763 for mesh-based GLB), although representation-specific differences remain.

Post-render scoring averages 7.9 s per trajectory on an RTX 4090, with most computation spent on GroundingDINO inference and TSDF construction. Score-validity conventions, backend comparisons, and timing details are provided in the supplement.

5 Limitations and Responsible Use

First, HPS provides a measurable operationalization of human-plausible execution rather than an exhaustive model of human navigation. Its four constructs are grounded in and validated against human judgments, but they do not capture the full diversity of human motion, navigation strategies, or viewing behavior.

Second, HPS evaluates observable evidence left by an execution rather than an agent's latent cognition or attention. This intentional boundary makes the metric reproducible and agent-independent: View for Action measures whether visual evidence is available to support action, without claiming that the agent internally attends to or uses that evidence.

Third, evidence-grounded scoring depends on sufficiently accurate scene reconstruction, camera calibration, and depth observations. Severe reconstruction artifacts, such as floaters in 3DGS scenes, can introduce false occupancy or contact evidence, particularly for Physical Walkability, and reduce confidence in the aggregate HPS. HPS therefore withholds cases that fail its evidence-quality checks rather than interpreting missing or unreliable evidence as an execution failure.

6 Conclusion

HumanPathScore separates task achievement from execution plausibility while preserving Physical Walkability, Motion Naturalness, Behavioral Coherence, and View for Action beside its aggregate. HPS aligns strongly with overall human ratings in discrete and continuous navigation, and its constructs are positively associated with their intended judgments. Across benchmark settings, HPS also yields within-setting method orderings distinct from conventional task and reference metrics. These results support HPS as a complementary and diagnosable evaluation of human-plausible VLN execution.

AI Use Disclosure

In this work, we used generative AI tools to help organize research ideas, write portions of the experimental code, and improve the wording and readability of the manuscript. The authors reviewed all AI-assisted code and text, checked the resulting implementation and claims against project materials and experimental records, and take responsibility for the final content, results, and artifacts.

REFERENCES

Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. BEVBert: Multimodal map pre-training for language-guided navigation. In *ICCV*, 2023.

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. ETPNav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 5130–5145, 2025.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018a.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pp. 3674–3683, 2018b.

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *CoRL*, pp. 671–681, 2021.

Dhruv Batra, Angel X. Chang, Sonia Chernova, et al. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, pp. 667–676, 2017.

David L. Chen and Raymond J. Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI*, pp. 859–865, 2011.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*, pp. 12530–12539, 2019.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, pp. 5834–5847, 2021.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, pp. 16516–16526, 2022.

Ian Colbert and Mehdi Saeedi. Human-like navigation behavior: A statistical evaluation framework. *arXiv preprint arXiv:2203.05965*, 2022.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pp. 303–312, 1996.

Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, and Katja Hofmann. Navigation turing test (ntt): Learning to evaluate human-like navigation. In *ICML*, pp. 2644–2653, 2021.

Yifei Dong, Fengyi Wu, Qi He, et al. Ha-vln 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions. *arXiv preprint arXiv:2503.14229*, 2025.

Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *HRI*, pp. 301–308, 2013.

Anthony Francis, Claudia Perez-D'Arpino, Chengshu Li, et al. Principles and guidelines for evaluating social robot navigation algorithms. *ACM Transactions on Human-Robot Interaction*, pp. 1–65, 2025.

Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, pp. 4282–4286, 1995.

Halim Hicheur, Quang-Cuong Pham, Gustavo Arechavaleta, Jean-Paul Laumond, and Alain Berthoz. The formation of trajectories during goal-oriented locomotion in humans. i. a stereotyped behaviour. *European Journal of Neuroscience*, 26(8):2376–2390, 2007.

M. A. Hollands, A. E. Patla, and J. N. Vickers. Look where you're going: Gaze behaviour associated with maintaining and changing the direction of locomotion. *Experimental Brain Research*, 143 (2):221–230, 2002.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language BERT for navigation. In *CVPR*, pp. 1643–1653, 2021.

Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *CVPR*, pp. 15418–15428, 2022.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *ACL*, pp. 1862–1872, 2019.

Amir-Homayoun Javadi, Eva Zita Patai, Eugenia Marin-Garcia, et al. Backtracking during navigation is correlated with enhanced anterior cingulate activity and suppression of alpha oscillations and the default-mode network. *Proceedings of the Royal Society B: Biological Sciences*, 286 (1908):20191016, 2019.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, pp. 1–14, 2023.

Heeju Ko, Sung June Kim, Gyeongrok Oh, Jeongyoon Yoon, Honglak Lee, Sujin Jang, Seungryong Kim, and Sangpil Kim. Active test-time vision-language navigation. *arXiv preprint arXiv:2506.06630*, 2025.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, pp. 104–120, 2020.

Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, pp. 1726–1743, 2013.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, pp. 4392–4412, 2020.

Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. KERM: Knowledge enhanced reasoning for vision-and-language navigation. In *CVPR*, pp. 2583–2592, 2023.

Yue Li, Qi Ma, Runyi Yang, et al. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. In *ICCV*, pp. 4961–4972, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55, 2024.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI*, pp. 1475–1482, 2006.

Bingchen Miao, Rong Wei, Zhiqi Ge, et al. Towards physically executable 3d gaussian for embodied navigation. *arXiv preprint arXiv:2510.21307*, 2025.

Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *IROS*, pp. 1366–1373, 2017.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *AAAI*, pp. 2017–2025, 2022.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, pp. 9982–9991, 2020.

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. Navbench: Probing multimodal large language models for embodied navigation. *arXiv preprint arXiv:2506.01031*, 2025.

Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, pp. 137–153, 2015.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, et al. Habitat: A platform for embodied ai research. In *ICCV*, pp. 9339–9347, 2019.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, pp. 10740–10749, 2020.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *CVPR*, pp. 12078–12088, 2025.

Samson Sun, Tianyi Yang, Tengyue Wang, Yikai Xue, Zhengjie Xu, Lingming Zhang, Qichen Zhang, Chao Liang, and Zhipeng Zhang. Beyond isolation: A unified benchmark for general-purpose navigation. In *RSS*, 2026.

Francesco Taioli, Nitzan Bitton-Guetta, Dieuwke Hupkes, Majd Beseiso, Ori Shapira, Panupong Pasupat, Mor Geva, Noam Slonim, and Roee Aharoni. Mind the error! detection and localization of instruction errors in vision-and-language navigation. In *Findings of EMNLP*, pp. 2586–2600, 2024.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *CoRL*, pp. 394–406, 2020.

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *CVPR*, pp. 13139–13150, 2024.

William H. Warren. Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, 10(5):683–703, 1984.

Meng Wei, Chenyang Wan, Xiqian Yu, et al. StreamVLN: Streaming vision-and-language navigation via SlowFast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

Jan M. Wiener, Simon J. Büchner, and Christoph Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition and Computation*, 9(2):152–165, 2009.

Tim Windecker, Manthan Patel, Moritz Reuss, Richard Schwarzkopf, Cesar Cadena, Rudolf Lioutikov, Marco Hutter, and Jonas Frey. Navitrace: Evaluating embodied navigation of vision-language models. In *ICRA*, 2026.

Thomas Wolbers and Mary Hegarty. What determines our navigational abilities? *Trends in Cognitive Sciences*, 14(3):138–146, 2010.

Xuan Yao, Junyu Gao, and Changsheng Xu. NavMorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *ICCV*, pp. 5536–5546, 2025.

Shuhao Ye, Sitong Mao, Yuxiang Cui, Xuan Yu, Shichao Zhai, Wen Chen, Shunbo Zhou, Rong Xiong, and Yue Wang. ETP-R1: Evolving topological planning with reinforcement fine-tuning for vision-language navigation in continuous environments. *arXiv preprint arXiv:2512.20940*, 2025.

Naoki Yokoyama, Sehoon Ha, and Dhruv Batra. Success weighted by completion time: A dynamics-aware evaluation criteria for embodied navigation. In *IROS*, pp. 1562–1569, 2021.

Ting Yu, Yifei Wu, Qiongjie Cui, Qingming Huang, and Jun Yu. MossVLN: Memory-observation synergistic system for continuous vision-language navigation. *IEEE Transactions on Multimedia*, pp. 6690–6704, 2025.

Lu Yue, Dongliang Zhou, Liang Xie, Feitian Zhang, Ye Yan, and Erwei Yin. Safe-vln: Collision avoidance for vision-and-language navigation of autonomous robots operating in continuous environments. *IEEE Robotics and Automation Letters*, pp. 4918–4925, 2024.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. NaVid: Video-based VLM plans the next step for vision-and-language navigation. In *RSS*, 2024.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *RSS*, 2025.

Xuesong Zhang, Yunbo Xu, Jia Li, Ruonan Liu, and Zhenzhen Hu. Agent journey beyond RGB: Hierarchical semantic-spatial representation enrichment for vision-and-language navigation. In *AAAI*, pp. 18791–18799, 2026.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Wang, and Qi Wu. NavGPT-2: Unleashing navigational reasoning capability for large vision-language models. In *ECCV*, pp. 260–278, 2024.