GUIDEDNAV: SHAPING VISION-LANGUAGE NAVIGATION REPRESENTATIONS THROUGH PRE-ACTION ATTENTION AND SPATIAL GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-and-language navigation (VLN) requires an agent to track progress through an extended instruction, identify landmarks relevant to its current objective, and understand the surrounding spatial structure to act. End-to-end navigation policies typically rely on action supervision to implicitly learn these capabilities. Without explicit guidance on the evidence underlying each decision, however, policies can overfit to correlations predictive of training actions rather than learn the task-relevant grounding and geometry needed for generalization. Our insight is to guide the representations that connect instructions to actions: identifying what remains to be done determines which visual evidence matters, while understanding the geometry around that evidence informs how to move. We introduce **GuidedNav**, a unified framework that explicitly shapes these representations through training-time guidance. Subtask and landmark guidance steer native pre-action attention toward the active instruction segment and relevant visual regions, grounding action prediction in the current objective. Spatial guidance aligns the policy's internal representations with geometric features that capture long-horizon context and local spatial structure. Together, these objectives strengthen navigation capabilities while avoiding the inference overhead of intermediate textual reasoning and online geometric encoding. We also introduce **Guided-R2R** and **Guided-RxR**, two annotation-enriched navigation datasets that support more interpretable VLN research. GuidedNav establishes state-of-the-art performance among the compared monocular RGB policies, improving success rates by 8.1% and 13.1% relative to the strongest prior results on R2R-CE and RxR-CE val-unseen, respectively, while achieving faster inference. Extensive experiments further demonstrate the effectiveness and real-world transferability of our approach.

1 INTRODUCTION

Consider the instruction "Walk towards the yellow chair, then walk past the dining table, and enter the room" (Figure 1). Moving forward may be appropriate at each stage, although the active instruction span and relevant landmark differ. The action label alone therefore does not specify which evidence should support a decision. This ambiguity matters in vision-and-language navigation (VLN), where an agent follows extended instructions through changing visual observations (Anderson et al., 2018; Ku et al., 2020). Recent vision-language policies predict actions directly from instructions and RGB video (Zhang et al., 2024; 2025a; Cheng et al., 2025), but their behavioral supervision only indirectly constrains the task and visual information used to act.

Grounding the current objective is only part of the problem: the policy must also relate 2D observations to the 3D layout through which it moves. Historical views provide broader spatial context, while nearby observations reveal local structure relevant to the next movement. Existing work strengthens these capabilities through progress reasoning, region-level grounding, and geometric memory (Wang et al., 2026b; Guo et al., 2026; Li et al., 2023; Zeng et al., 2026). Intermediate text generation and online geometric encoding can, however, add inference cost. Auxiliary reasoning and future-state prediction provide additional training signals (Wang et al., 2025b; Zhang et al., 2026b), motivating a complementary question: at the moment of action generation, does the policy attend to the active instruction span and its visual referents, and encode the spatial information needed to proceed?

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.

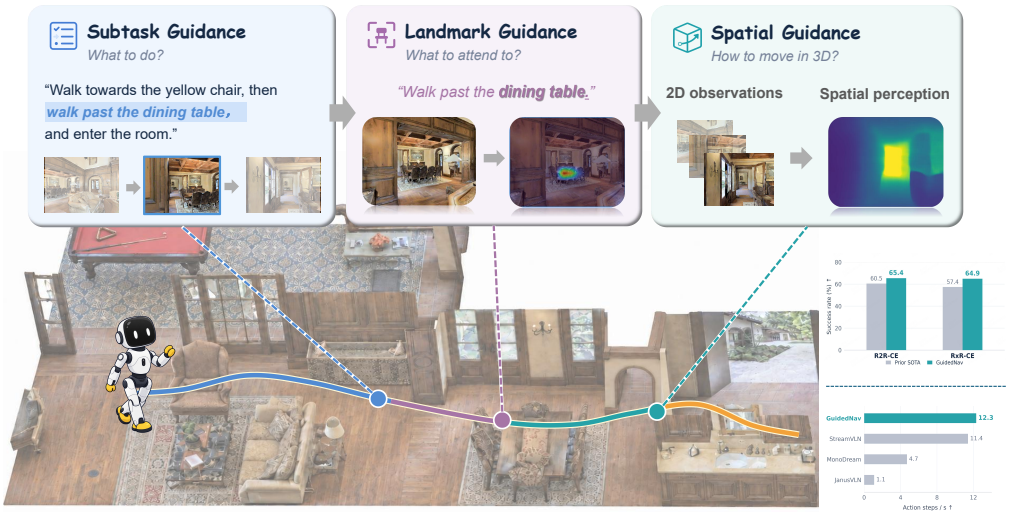

Figure 1: **From Instruction Progress to Grounded Navigation.** Along the illustrated route, moving forward can mean approaching a chair, passing a dining table, or entering a room. The relevant instruction span, landmark, and spatial context change even when the action remains the same. The three panels illustrate the evidence targeted by GuidedNav's training-time guidance.

We introduce **GuidedNav**, a unified framework that strengthens navigation representations through training-time auxiliary guidance. The position immediately preceding the first action token provides a natural supervision point: its hidden state predicts the action sequence's beginning, while its native attention gathers evidence from instructions and observations. Subtask and landmark guidance direct this attention toward the active instruction span and associated image regions, grounding action prediction in the current objective. Spatial guidance aligns current-frame representations with long-horizon geometric targets and supervises local temporal fusion with geometric information from nearby views. At inference, the policy retains the learned fusion but discards the geometric teacher and auxiliary projectors, enabling faster, direct action prediction without intermediate reasoning.

We introduce **Guided-R2R** and **Guided-RxR**, two annotation-enriched datasets linking navigation steps to active subtask spans and landmark masks. GuidedNav improves success rates by 8.1% and 13.1% relative to the strongest prior models on R2R-CE and RxR-CE, respectively, while achieving a 10.7× per-action inference speedup. Our contributions are threefold:

- We propose GuidedNav, a unified framework that explicitly guides navigation representations through subtask attention, landmark grounding, and spatial supervision, strengthening direct action prediction without intermediate reasoning or an online geometric encoder.
- We introduce Guided-R2R and Guided-RxR through a landmark-centric data engine, linking navigation steps to active subtask spans and landmark masks to support representation learning and more interpretable VLN research.
- We achieve state-of-the-art navigation performance with faster inference. Extensive benchmark evaluations, component ablations, representation analyses, and real-robot experiments validate the effectiveness of our guidance and its transfer to real-world environments.

2 Related Work

Vision-and-language navigation. VLN has evolved from navigation graphs to continuous action execution (Anderson et al., 2018; Ku et al., 2020; Krantz et al., 2020). Graph- and map-based methods combine grounding with planning (Chen et al., 2022; An et al., 2023; Wang et al., 2023b; An et al., 2024), while pretrained vision-language policies enable RGB-based action prediction and robot deployment (Zhang et al., 2024; 2025a; Cheng et al., 2025). GuidedNav focuses on the task and spatial representations supporting these decisions.

Instruction execution and visual grounding. Progress monitoring and auxiliary reasoning extend action supervision (Ma et al., 2019; Zhu et al., 2020); sub-instruction methods and Landmark-RxR

align instructions with trajectories and landmarks (Hong et al., 2020; He et al., 2021). Recent policies generate reasoning or progress descriptions (Zhou et al., 2024; Lin et al., 2025; Wang et al., 2026b; Guo et al., 2026), or learn from auxiliary reasoning and future-state prediction (Wang et al., 2025b; Zhang et al., 2026b). KERM and GELA strengthen region/entity grounding (Li et al., 2023; Cui et al., 2023); HINT aligns gradient-based explanations with human annotations (Selvaraju et al., 2019), and GuidedVLA specializes action-decoder attention heads (Jia et al., 2026). GuidedNav instead supervises native pre-action attention to the active subtask span and landmarks, without adding specialized heads or generating intermediate text.

Spatial representations for navigation. Maps organize observations for spatial reasoning (An et al., 2024; 2023; Wang et al., 2023b); StreamVLN models temporal context through streaming attention (Wei et al., 2026), while JanusVLN retains VGGT-based geometric memory at inference (Zeng et al., 2026; Wang et al., 2025a). Spatial Forcing transfers geometry through training-time representation alignment (Li et al., 2026). GuidedNav extends this principle to sequential navigation with history-conditioned targets and a local objective that supervises temporal fusion, incorporating both temporal scales without online teachers.

3 METHOD

3.1 PROBLEM FORMULATION

We consider vision-and-language navigation in continuous environments (Krantz et al., 2020). At step t, the agent predicts a short action sequence from instruction I, sampled historical RGB observations H_t, and current observation o_t:

$$A_t \sim \pi_\theta(\cdot \mid I, H_t, o_t). \tag{1}$$

Actions comprise forward movement, left/right turns, and stop; observations update after execution until stopping. Here $A_t = (a_{t,1}, \ldots, a_{t,m_t})$ contains up to four primitive actions ($m_t \leq 4$), terminating early when stop is predicted. We learn representations connecting the active instruction segment, landmarks, and spatial context for action prediction.

3.2 GUIDEDNAV FRAMEWORK

GuidedNav augments an autoregressive vision-language policy with training-time guidance (Figure 2): subtask and landmark objectives supervise native language and visual attention, while spatial guidance combines long-horizon geometry with nearby observations. Inference predicts actions directly.

Native pre-action attention. We read attention at the existing position immediately preceding the first action token, which predicts the beginning of the action sequence. Following the notation in Figure 2, Q_p denotes the query at this pre-action position, while K_I and K_t denote keys of instruction tokens and current-frame visual tokens. For a selected set of attention heads $\mathcal{B}$, we compute

$$\alpha_D = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathrm{softmax}\left(\frac{Q_p^{(b)} (K_D^{(b)})^\top}{\sqrt{d_b}} \right), \qquad D \in \{I, t\}, \tag{2}$$

where d_b is the head dimension. The query and keys retain the backbone's positional encoding, and softmax is applied separately within each token group before averaging across heads. Both domains exclude action tokens; the readout introduces no query tokens or attention layers.

3.2.1 SUBTASK ATTENTION GUIDANCE

An extended instruction contains several subtasks, only some of which are relevant to the current decision. Let $\mathcal{S}_t$ denote the token indices of the active subtask span, obtained from the annotations in Section 3.3. We encourage the pre-action position to attend to this span through

$$\mathcal{L}_{\mathrm{Subtask}} = -c_t \log \left(\sum_{i \in \mathcal{S}_t} \alpha_{I,i} \right), \tag{3}$$

where c_t balances the contribution of subtask intervals, preventing long intervals from dominating the span objective (Appendix A.2). This objective increases attention to the active span while allowing

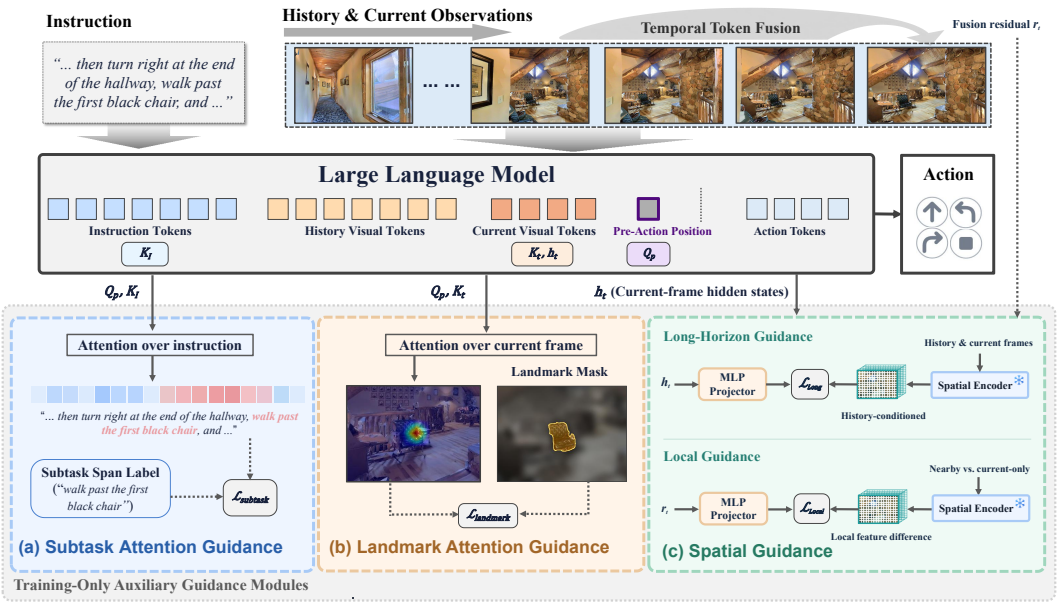

Figure 2: **Architecture of GuidedNav.** Subtask and landmark guidance supervise native pre-action attention toward the active instruction span and relevant visual regions. Spatial guidance aligns current-frame representations with history-conditioned geometric features and supervises local fusion using geometric feature differences from nearby observations. All auxiliary objectives and the frozen spatial encoder are used only during training.

the policy to distribute attention among its tokens. Subtask labels provide training supervision and are not required as policy inputs.

3.2.2 Landmark Attention Guidance

To connect instruction execution with visual evidence, we guide attention toward landmarks associated with the active subtask. Accepted landmark masks are combined and area-pooled onto the current visual-token grid, yielding patch occupancies $m_{t,j} \in [0, 1]$. We maximize the attention mass supported by these regions:

$$\mathcal{L}_{\text{Landmark}} = -\log\left(\sum_{j=1}^{N_t} m_{t,j}\alpha_{t,j}\right), \tag{4}$$

where N_t is the number of current-frame visual tokens. The mask weights preserve fractional patch coverage rather than forming a normalized target distribution. Together, the two attention objectives connect the active instruction segment and its visual referents to the same action-predicting position.

3.2.3 Spatial Guidance

A single RGB view provides incomplete depth and occlusion cues; sampled history and nearby views offer complementary context. We use a frozen VGGT teacher for long-horizon alignment and local fusion supervision (Figure 8).

Long-horizon spatial guidance. Frozen VGGT (Wang et al., 2025a) jointly encodes sampled history and the current frame. Its final aggregated patch features f_t^{long} are aligned to the policy's current-frame grid after adding positional embeddings; Appendix A.2 specifies the preprocessing. An MLP P_{long} aligns current-frame states $h_{t,j}$ from hidden-state index 27 (h_t in Figure 2) with these history-conditioned targets:

$$\mathcal{L}_{\text{Long}} = \frac{1}{N_t}\sum_{j=1}^{N_t}\left[1 - \cos\left(P_{\text{long}}(h_{t,j}), f_{t,j}^{\text{long}}\right)\right]. \tag{5}$$

History supplies context to both networks; alignment covers only the current frame's full patch grid, independently of landmark masks.

Temporal token fusion. Before the language model, we refine current visual-encoder tokens $z_{t,j}$ using up to three immediately preceding observations. After aligning token-grid resolutions, cosine similarity with a spatial-distance penalty selects mutual best matches $\mathcal{M}_t(j)$, retaining at most 25% of the grid per past frame (Appendix A.2):

$$r_{t,j} = \frac{\sum_{(u,i)\in\mathcal{M}_t(j)} w_{t-u}(z_{u,i} - z_{t,j})}{\sum_{(u,i)\in\mathcal{M}_t(j)} w_{t-u}}, \qquad \widetilde{z}_{t,j} = z_{t,j} + g\, r_{t,j}. \tag{6}$$

Here w_{t-u} weights temporal distance, g is a learned bounded scalar gate, and unmatched tokens have $r_{t,j} = 0$. Fusion preserves token count and spatial arrangement.

Local spatial guidance. To supervise fusion, we compare current-frame teacher features from a local temporal window, f_t^{local}, and the current frame alone, f_t^{single}. Aligned to the current patch grid, their normalized difference Δf_t captures the teacher-feature change induced by nearby views (Equation 9, Appendix A.2). A separate projector P_{local} aligns the pre-gate residual with this target over patches with valid matches, denoted by $\mathcal{J}_t$:

$$\mathcal{L}_{\text{Local}} = \frac{1}{|\mathcal{J}_t|} \sum_{j\in\mathcal{J}_t} \left[1 - \cos\left(P_{\text{local}}(r_{t,j}), \Delta f_{t,j}\right)\right]. \tag{7}$$

We omit this term without valid matches. Long-horizon guidance supervises current representations; local guidance supervises fusion residuals. Both use only current and past observations.

3.3 Landmark-Centric Data Engine

We construct Guided-R2R and Guided-RxR by augmenting R2R and English RxR training instructions with step-aligned subtask spans and landmark masks. R2R spans come from FGR2R (Hong et al., 2020), mapped to continuous trajectories using recorded poses; RxR spans are extracted by gpt-oss-120b and temporally refined by GLM-4.5V. GLM-4.5V selects landmarks and verifies GroundingDINO boxes (Liu et al., 2024); SAM2 (Ravi et al., 2025) propagates accepted prompts within each subtask interval. Unreliable annotations are excluded from the corresponding auxiliary loss, while original observations, instructions, and action targets are retained. Appendix B and Figure 9 detail the pipeline.

3.4 Training and Inference

GuidedNav combines action-token cross-entropy with attention guidance and spatial guidance, where $\mathcal{L}_{\text{Spatial}} = \lambda_{\text{long}}\mathcal{L}_{\text{Long}} + \lambda_{\text{local}}\mathcal{L}_{\text{Local}}$:

$$\mathcal{L} = \mathcal{L}_{\text{Action}} + \eta(s)\big[\lambda_{\text{sub}}\mathcal{L}_{\text{Subtask}} + \lambda_{\text{lm}}\mathcal{L}_{\text{Landmark}} + \mathcal{L}_{\text{Spatial}}\big]. \tag{8}$$

where $\eta(s)$ gradually introduces auxiliary supervision at the beginning of training. Each auxiliary term is averaged over its valid training examples, and teacher features are detached from optimization. At inference, GuidedNav retains the learned policy and temporal fusion, while discarding the geometry teacher and auxiliary projectors. It consumes RGB observations and instructions without requiring subtask labels, landmark masks, or intermediate textual reasoning.

4 Experiments

4.1 Experimental Setup

Datasets and three-stage training. We evaluate on R2R-CE and RxR-CE val-unseen (Anderson et al., 2018; Ku et al., 2020; Krantz et al., 2020), using English guide instructions for RxR. Training starts from Qwen3-VL (Bai et al., 2025) and proceeds in three stages. Stage 1 fully fine-tunes the model on approximately 0.53M R2R and 1.90M RxR expert examples, together with 0.79M visual instruction-tuning (VQA) examples from LLaVA-Video (Zhang et al., 2025d) and VSI (Yang et al., 2025). Stage 2 continues full fine-tuning on 2.97M DAgger-collected navigation examples (Ross et al., 2011) mixed with VQA data. Stage 3 applies LoRA fine-tuning to Guided-R2R and Guided-RxR. Throughout the experiments, **Base Model** denotes the same Stage-2 checkpoint followed by

Table 1: **Monocular RGB Navigation on R2R-CE and RxR-CE Val-Unseen.** External data counts samples beyond standard R2R/RxR-CE data.

Method	R2R Val-Unseen				RxR Val-Unseen				External Training Data
	NE↓	OSR↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑	
NaVid$_{[RSS'24]}$ (Zhang et al., 2024)	5.47	49.1	37.4	35.9	–	–	–	–	953K
Sim2Real$_{[CoRL'24]}$ (Wang et al., 2025c)	5.95	55.8	44.9	30.4	8.79	36.7	25.5	18.1	0K
Uni-NaVid$_{[RSS'25]}$ (Zhang et al., 2025a)	5.58	53.3	47.0	42.7	6.24	48.7	40.9	–	3577K
InternVLA-N1$_{[Tech.\ Rep.'25]}$ (Intern Robotics, 2025)	4.89	60.6	55.4	52.1	6.41	49.5	41.8	62.6	51200K
StreamVLN$^{*}_{[ICRA'26]}$ (Wei et al., 2026)	6.05	53.8	45.5	41.6	–	–	–	–	10033K
NaVILA$^{*}_{[RSS'25]}$ (Cheng et al., 2025)	5.37	57.6	49.7	45.5	–	–	–	–	12574K
NaVILA$_{[RSS'25]}$ (Cheng et al., 2025)	5.22	62.5	54.0	49.0	6.77	49.3	44.0	58.8	13132K
Aux-Think$_{[NeurIPS'25]}$ (Wang et al., 2025b)	6.08	60.0	54.8	46.9	6.24	52.2	40.2	–	1600K‡
MonoDream$_{[AAAI'26]}$ (Wang et al., 2026a)	5.45	61.5	55.8	49.1	6.38	49.4	40.9	–	0K‡
JanusVLN$^{†}_{[ICLR'26]}$ (Zeng et al., 2026)	5.17	58.0	52.8	49.2	6.46	51.4	44.3	59.1	0K
StreamVLN$_{[ICRA'26]}$ (Wei et al., 2026)	4.98	64.2	56.9	51.9	6.22	52.9	46.0	61.9	∼26330K
NavFoM$_{[ICLR'26]}$ (Zhang et al., 2026a)	5.01	64.9	56.2	51.2	5.51	57.4	49.4	60.2	∼10480K
JanusVLN$_{[ICLR'26]}$ (Zeng et al., 2026)	4.78	65.2	60.5	56.8	6.06	56.2	47.5	62.1	10692K
GuidedNav (Ours)	**3.83**	**71.5**	**65.4**	**59.1**	**4.40**	**64.9**	**54.3**	**69.2**	3765K

†: VLN-CE expert trajectories only, without additional training data. StreamVLN* uses EnvDrop data; NaVILA* excludes human-following data. ‡: source-reported data counts use different definitions: Aux-Think includes additional RxR, DAgger, and web data, whereas MonoDream excludes simulator-generated data. "–" denotes an unavailable metric or data count, not zero.

one epoch of LoRA fine-tuning on the original R2R and RxR data without our designed guidance modules, providing the comparison for GuidedNav.

Training configuration. Each stage runs for one epoch with global batch size 256, AdamW, cosine learning-rate decay, 3% warmup, and weight decay 0.01. Stages 1 and 2 update the language model, multimodal projector, and vision tower, with peak language-model learning rates of 2×10^{-5} and 1×10^{-5}, respectively; the projector and vision tower use 5×10^{-6} in both stages. Stage 3 uses rank-32 LoRA with scaling factor 16 and a peak learning rate of 2×10^{-4} for both Base Model and GuidedNav. Training uses BF16 and DeepSpeed ZeRO-2. See Appendix A.1 for more details.

Metrics and evaluation protocol. We report navigation error (NE), oracle success rate (OSR), success rate (SR), and success weighted by path length (SPL), together with normalized dynamic time warping (nDTW) on RxR (Ilharco et al., 2019). NE is measured in meters and the remaining metrics in percentages. Evaluation uses greedy decoding with up to eight context images and four-action execution chunks. Each primitive action moves forward by 0.25 m or rotates by $15°$, with a maximum of 500 actions per episode. Table 6 in Appendix A lists the remaining simulator settings.

4.2 Main Results

Results on VLN-CE benchmarks. Table 1 shows that GuidedNav advances the performance frontier for monocular RGB navigation. Relative to the strongest prior result for each metric, it improves SR by 8.1% on R2R-CE and 13.1% on RxR-CE, alongside SPL gains of 4.0% and 9.9%. The SR comparisons use JanusVLN (Zeng et al., 2026) on R2R and NavFoM (Zhang et al., 2026a) on RxR, corresponding to gains of 4.9 and 7.5 percentage points. Thus the improvement extends beyond reaching more goals to completing routes more efficiently. On RxR, nDTW also improves by 10.5% over the best prior RGB result, indicating closer adherence to the instructed route rather than endpoint success alone. This joint improvement is consistent with guidance linking the current objective to visual evidence and spatial context throughout navigation. Additional comparisons and trajectory visualizations appear in Appendices E and D.

Performance across route lengths. Figure 3 compares GuidedNav with Base Model across four path-length groups. SR changes little and SPL is slightly lower in the two shorter groups, whereas both metrics improve in the 8.43–10.43 m and > 10.43 m groups. The benefit is therefore concentrated where the policy must maintain the instruction-to-scene connection over more decisions; on shorter routes, the small SPL decrease suggests that guidance does not uniformly improve path efficiency.

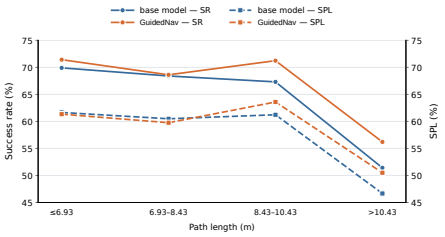

Figure 3: **Performance across Path-Length Groups** on R2R-CE val-unseen.

Table 2: **Inference Efficiency** on a single NVIDIA RTX 5090. Latency is amortized per executed primitive action; throughput is measured in action steps per second.

Method	Params (B)	ms/step↓	steps/s↑
StreamVLN (Wei et al., 2026)	8.03	87	11.4
JanusVLN (Zeng et al., 2026)	9.31	864	1.1
MonoDream (Wang et al., 2026a)	2.00	211	4.7
GuidedNav (Ours)	8.77	**81**	**12.3**

Table 3: Ablation studies on R2R-CE val-unseen. SG denotes Spatial Guidance. The upper block evaluates individual modules against Base Model, defined in Section 4.1; the lower block progressively builds Spatial Guidance. The complete configuration in each block is shaded in blue.

Configuration	NE↓	OSR↑	SR↑	SPL↑
Base Model	4.87	63.8	59.3	56.1
Overall Guidance Ablation				
Base + Subtask Attention Guidance	4.34	**72.3**	62.9	55.2
Base + Landmark Attention Guidance	4.38	72.1	62.6	55.8
Base + Spatial Guidance	4.33	68.5	62.4	57.2
GuidedNav (All)	**3.83**	71.5	**65.4**	**59.1**
Detailed Spatial Guidance Ablation				
Base + Long-Horizon SG	4.35	68.4	61.8	56.6
Base + Long-Horizon SG + Gated Fusion	**4.30**	68.0	62.0	56.7
Base + Long-Horizon SG + Gated Fusion + Local SG	4.33	**68.5**	**62.4**	**57.2**

4.3 Inference Efficiency

Table 2 reports measurements on a single NVIDIA RTX 5090, amortized over the primitive actions executed from each prediction. GuidedNav achieves 12.3 steps/s at 81 ms per step, yielding 10.7× and 2.6× per-step speedups over JanusVLN and MonoDream, respectively, and 6.9% lower latency than StreamVLN. The central efficiency advantage is architectural: geometric learning is moved into training, so deployment retains temporal fusion but discards the teacher and auxiliary projectors. This allows the policy to benefit from spatial guidance without repeatedly encoding geometry or generating intermediate reasoning.

4.4 Ablation Study

The upper block of Table 3 shows complementary effects of attention and spatial guidance under the same Stage-3 protocol. Subtask, landmark, and spatial guidance improve SR over Base Model by 3.6, 3.3, and 3.1 percentage points, respectively. Attention guidance gives the largest OSR gains, suggesting better access to the evidence needed to reach the goal vicinity, whereas spatial guidance also improves SPL by 1.1 points. Combining the three signals improves SR by 6.1 points and SPL by 3.0 points over Base Model, outperforming each individual module on both metrics.

The lower block isolates the progressive spatial design. Beyond long-horizon guidance, gated fusion adds 0.2 SR and 0.1 SPL points; supervising the local residual adds another 0.4 and 0.5 points. The larger incremental SPL gain from local supervision suggests that the content of the fused temporal information matters more than fusion alone. Appendix C compares alternative supervision strategies: direct attention guidance gives the highest SR for both subtask and landmark learning, supporting supervision at the action-predicting position without requiring textual intermediate outputs.

Table 4: **Attention Localization and Spatial Probing.** Subtask and landmark localization are evaluated on R2R-CE val-unseen; EBPG is computed at the patch level, and Rank Acc. measures near–far depth ordering. Protocols appear in Appendices C.3 and C.4.

	Subtask Localization		*Landmark Localization*		*Spatial Probing*		
Model	AUPRC (%) ↑	AUROC (%) ↑	PG (%) ↑	EBPG (%) ↑	AbsRel ↓	RMSE (m) ↓	Rank Acc. (%) ↑
Base Model	34.5	53.2	15.2	13.4	0.25	0.82	85.9
GuidedNav	**56.3**	**76.1**	**49.1**	**42.6**	**0.19**	**0.64**	**91.7**

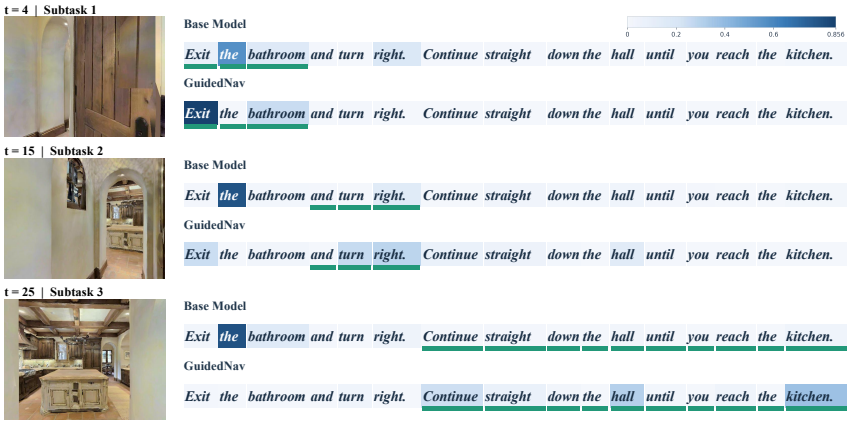

Figure 4: **Subtask Attention across Navigation Steps.** We compare the Base Model and GuidedNav at sampled steps. Green underlines mark the GT active subtask, while darker blue indicates higher native pre-action attention.

4.5 Representation Analysis

Task and landmark localization. We visualize native attention from the position immediately preceding the first action token. Attention is normalized separately over instruction tokens and current-frame visual tokens, then averaged across the selected heads. The resulting scores highlight words in the instruction and form heatmaps on the visual-token grid. Table 4 shows gains of 21.8 and 22.9 points in subtask AUPRC and AUROC, alongside landmark PG and EBPG improvements of 33.9 and 29.2 points. The subtask gains indicate better separation of the active instruction span from completed or upcoming subtasks. The landmark gains extend beyond placing the attention peak inside the target: more attention mass is also assigned to the annotated regions, suggesting more consistent selection of task-relevant visual evidence. Figure 4 illustrates how this emphasis changes with instruction progress, while Figure 5 shows stronger localization of the relevant landmarks. Together, these observations suggest that guidance helps organize evidence around the current navigation objective rather than merely producing a fixed preference for salient words or objects. These maps describe relative attention within each token group; they do not directly measure the causal contribution of individual tokens to an action. Further details appear in Appendix C.

Spatial information in frozen representations. We assess spatial information by fitting depth probes to frozen navigation features, allowing the learned representations to be examined. For visualization, a separate convolutional head reads current-frame features and predict a depth map. This head is used only for analysis and does not participate in action prediction. GuidedNav reduces probe AbsRel by 24.0% and RMSE by 22.0%, while improving near–far ranking accuracy by 5.8 points (Table 4). The joint improvement matters because depth errors and ranking accuracy capture complementary properties: the former assess metric reconstruction, whereas the latter tests whether the representation preserves relative spatial ordering. Figure 6 provides a qualitative view of the recovered scene structure, with lower per-example depth errors for GuidedNav in both displayed cases. These results suggest that spatial guidance makes geometry more accessible within the policy itself, even after the geometric teacher is removed. Details are shown in Appendix C.4.



Figure 5: **Landmark Grounding at the Pre-Action Position.** Reference masks (left), Base Model attention (middle), and GuidedNav attention (right) for stairs and chairs.

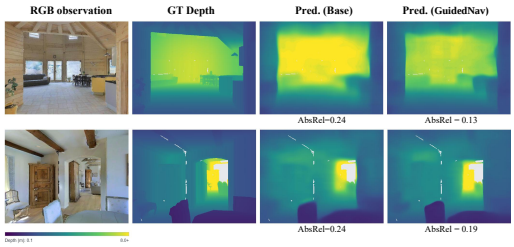

Figure 6: **Depth Decoded from Frozen Navigation Representations.** RGB, reference depth, and Base Model/GuidedNav predictions, with per-example AbsRel and depth in meters.

4.6 Real-World Navigation

On an AgiBot X2 humanoid with a forward-facing RGB camera, GuidedNav navigates without real-world fine-tuning. Figure 7 shows the robot passing the orange chairs and stopping at the coffee machine: attention shifts from the intermediate landmark to the destination as the instruction progresses. This qualitative transfer illustrates the learned instruction–landmark connection beyond simulation, while preserving direct action prediction. Deployment details and additional cases appear in Appendix D.2.

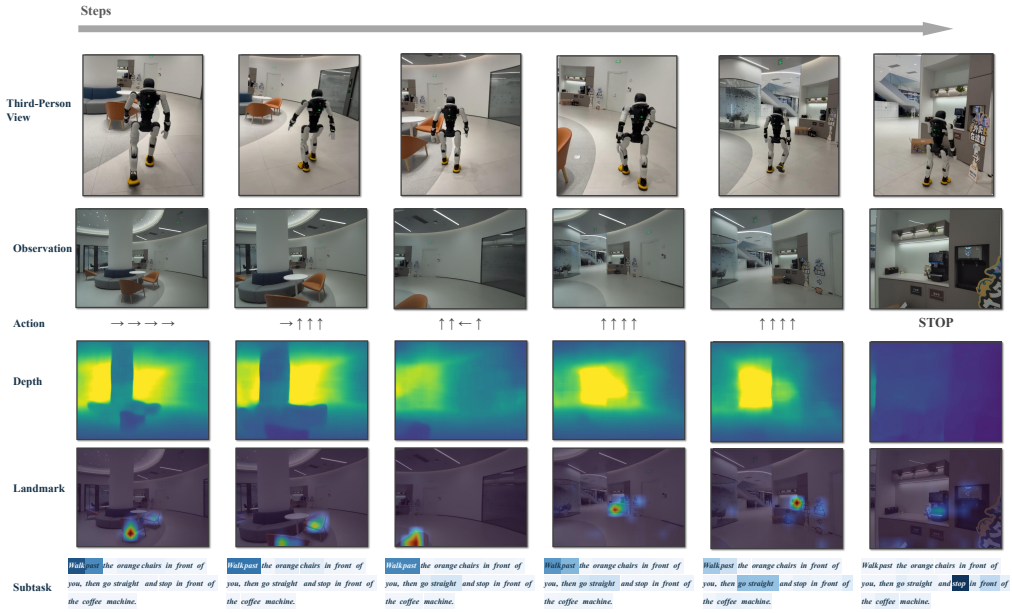

Figure 7: **Real-World Navigation without Real-World Fine-Tuning.** Six decisions show third-person views, RGB observations, actions, decoded depth, and pre-action attention. Depth is for analysis only; attention is normalized within each displayed token group.

5 Conclusion

We presented GuidedNav, which unifies instruction progress, landmark grounding, and spatial geometry through training-time representation guidance. Supervising native pre-action attention and transferring geometry across long-horizon and local contexts strengthen the policy while preserving direct action prediction. Guided-R2R and Guided-RxR supply aligned subtask and landmark supervision. State-of-the-art monocular RGB results, representation analyses, and real-world transfer demonstrate the promise of this approach: richer navigation representations without an online geometric teacher or intermediate textual reasoning.

AI USE STATEMENT

We used large language models to assist with manuscript drafting, translation, language refinement, and code implementation. We also used generative AI tools to assist with schematic illustrations, and employed gpt-oss-120b and GLM-4.5V in the annotation pipeline described in Section 3.3. The authors are responsible for verifying all AI-assisted content and for the accuracy and integrity of the final manuscript, implementation, and reported results.

REFERENCES

Dong An, Zun Wang, Yangguang Li, Yi Wang, Yicong Hong, Yan Huang, Liang Wang, and Jing Shao. 1st place solutions for RxR-Habitat vision-and-language navigation competition (CVPR 2022). *arXiv preprint arXiv:2206.11610*, 2022. URL https://arxiv.org/abs/2206.11610.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://github.com/MarSaKi/VLN-BEVBert.

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*, 2024. URL https://github.com/MarSaKi/ETPNav.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K. Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23568–23576, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34526.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16537–16547, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Think_Global_Act_Local_Dual-Scale_Graph_Transformer_for_Vision-and-Language_Navigation_CVPR_2022_paper.html.

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/0deb1c54814305ca9ad266f53bc82511-Abstract.html.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2412.04453.

Yibo Cui, Liang Xie, Yakun Zhang, Meishan Zhang, Ye Yan, and Erwei Yin. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12043–12053, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cui_Grounded_Entity-Landmark_Adaptive_Pre-Training_for_Vision-and-Language_Navigation_ICCV_2023_paper.html.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the International Conference on Machine Learning*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://arxiv.org/abs/1406.2283.

Wenxuan Guo, Xiuwei Xu, Yichen Liu, Xiangyu Li, Hang Yin, Huangxing Chen, Wenzhao Zheng, Jianjiang Feng, Jie Zhou, and Jiwen Lu. AwareVLN: Reasoning with self-awareness for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4065–4075, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Guo_AwareVLN_Reasoning_with_Self-awareness_for_Vision-Language_Navigation_CVPR_2026_paper.html.

Keji He, Yan Huang, Qi Wu, Jianhua Yang, Dong An, Shuanglin Sima, and Liang Wang. Landmark-RxR: Solving vision-and-language navigation with fine-grained alignment supervision. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0602940f23884f782058efac46f64b0f-Abstract.html.

Yicong Hong, Cristian Rodriguez, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 3360–3376, 2020. doi: 10.18653/v1/2020.emnlp-main.271. URL https://aclanthology.org/2020.emnlp-main.271/.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15439–15449, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Hong_Bridging_the_Gap_Between_Learning_in_Discrete_and_Continuous_Environments_CVPR_2022_paper.html.

Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, and Hao Tan. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3055–3067, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hong_Learning_Navigational_Visual_Representations_with_Semantic_Map_Supervision_ICCV_2023_paper.html.

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. URL https://arxiv.org/abs/1907.05446.

Intern Robotics. InternVLA-N1: An open dual-system vision-language navigation foundation model with learned latent plans. Technical report, Shanghai Artificial Intelligence Laboratory, 2025. URL https://internrobotics.github.io/internvla-n1.github.io/static/pdfs/InternVLA_N1.pdf.

Xiaosong Jia, Bowen Yang, Zuhao Ge, Xian Nie, Yuchen Zhou, Cunxin Fan, Yufeng Li, Yilin Chai, Chao Jing, Zijian Liang, Qingwen Bu, Haidong Cao, Chao Wu, Qifeng Li, Zhenjie Yang, Chenhe Zhang, Hongyang Li, Zuxuan Wu, Junchi Yan, and Yu-Gang Jiang. GuidedVLA: Specifying task-relevant factors via plug-and-play action attention specialization. In *Robotics: Science and Systems*, 2026. URL https://arxiv.org/abs/2605.12369.

Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 588–603, 2022. URL https://arxiv.org/abs/2204.09667.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, 2020. URL https://arxiv.org/abs/2004.02857.

Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15162–15171, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Krantz_Waypoint_Models_for_Instruction-Guided_Navigation_in_Continuous_Environments_ICCV_2021_paper.html.

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*, 2020. URL https://arxiv.org/abs/2010.07954.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=euMVC1DO4k.

Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. KERM: Knowledge enhanced reasoning for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2583–2592, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_KERM_Knowledge_Enhanced_Reasoning_for_Vision-and-Language_Navigation_CVPR_2023_paper.html.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. NavCoT: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. URL https://github.com/expectorlin/NavCoT.

Haoran Liu, Weikang Wan, Xiqian Yu, Minghan Li, Jiazhao Zhang, Bo Zhao, Zhibo Chen, Zhongyuan Wang, Zhizheng Zhang, and He Wang. NaVid-4D: Unleashing spatial intelligence in egocentric RGB-D videos for vision-and-language navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10607–10615, 2025. URL https://lhrrhl0419.github.io/NaVid4D/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06319.pdf.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. InstructNav: Zero-shot system for generic instruction navigation in unexplored environment. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2049–2060. PMLR, 2025. URL https://proceedings.mlr.press/v270/long25b.html. CoRL 2024.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.03035.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/45c1f6a8cbf2da59ebf2c802b4f742cd-Paper-Conference.pdf.

Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (LAW) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4018–4028. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.emnlp-main.328/.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Selvaraju_Taking_a_HINT_Leveraging_Explanations_to_Make_Vision_and_Language_ICCV_2019_paper.html.

Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. DREAMWALKER: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pp. 10873–10883, 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DREAMWALKER_Mental_Planning_for_Continuous_Vision-Language_Navigation_ICCV_2023_paper.html.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. URL https://arxiv.org/abs/1910.01279.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a. URL https://arxiv.org/abs/2503.11651.

Shuo Wang, Yongcai Wang, Wanting Li, Xudong Cai, Yucheng Wang, Maiyue Chen, Kaihui Wang, Zhizhong Su, Deying Li, and Zhaoxin Fan. Aux-Think: Exploring reasoning strategies for data-efficient vision-language navigation. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.11886.

Shuo Wang, Yongcai Wang, Zhaoxin Fan, Yucheng Wang, Maiyue Chen, Kaihui Wang, Zhizhong Su, Wanting Li, Xudong Cai, Yeying Jin, and Deying Li. MonoDream: Monocular vision-language navigation with panoramic dreaming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10074–10082, 2026a. doi: 10.1609/aaai.v40i12.37974. URL https://ojs.aaai.org/index.php/AAAI/article/view/37974.

Shuo Wang, Yucheng Wang, Guoxin Lian, Yongcai Wang, Maiyue Chen, Kaihui Wang, Bo Zhang, Zhizhong Su, Yutian Zhou, Wanting Li, Deying Li, and Zhaoxin Fan. Progress-Think: Semantic progress reasoning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4076–4086, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Progress-Think_Semantic_Progress_Reasoning_for_Vision-Language_Navigation_CVPR_2026_paper.html.

Zihan Wang and Gim Hee Lee. g3D-LF: Generalizable 3D-language feature fields for embodied tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14191–14202, 2025. URL https://arxiv.org/abs/2411.17030.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. GridMM: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15625–15636, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_GridMM_Grid_Memory_Map_for_Vision-and-Language_Navigation_ICCV_2023_paper.html.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Sim-to-real transfer via 3D feature fields for vision-and-language navigation. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2982–2995. PMLR, 2025c. URL https://proceedings.mlr.press/v270/wang25i.html. CoRL 2024.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. StreamVLN: Streaming vision-and-language navigation via SlowFast context modeling. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://github.com/InternRobotics/StreamVLN.

Jiahao Yang, Zihan Wang, Xiangyang Li, Xing Zhu, Yujun Shen, Yinghao Xu, and Shuqiang Jiang. GA-VLN: Geometry-aware BEV representation for efficient vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2605.22036.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025. URL https://arxiv.org/abs/2511.04670.

Xuan Yao, Junyu Gao, and Changsheng Xu. NavMorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5536–5546, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yao_NavMorph_A_Self-Evolving_World_Model_for_Vision-and-Language_Navigation

_in_Continuous_ICCV_2025_paper.html.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, Xing Wei, and Ning Guo. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.22548.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, 2016. URL https://arxiv.org/abs/1608.00507.

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 *Robotics: Science and Systems*, 2024. URL https://arxiv.org/abs/2402.15852.

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 *Robotics: Science and Systems*, 2025a. doi: 10.15607/RSS.2025.XXI.013. URL https://www.roboticsproceedings.org/rss21/p013.html.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, Yuxin Fan, Wenjun Li, Zhibo Chen, Fei Gao, Qi Wu, Zhizheng Zhang, and He Wang. Embodied navigation foundation model. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=kkBOIsrCXh.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. MapNav: A novel memory representation via annotated semantic maps for VLM-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13032–13056. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.acl-long.638/.

Lingfeng Zhang, Zeying Gong, Xiaoshuai Hao, Haoxiang Fu, Qiang Zhang, Mingliang Zhou, Hangjun Ye, Xiaojun Liang, Junwei Liang, and Wenbo Ding. FutureNav: Unified world-action modeling for vision-and-language navigation. *arXiv preprint arXiv:2606.30367*, 2026b. URL https://arxiv.org/abs/2606.30367.

Siqi Zhang, Yanyuan Qiao, Qunbo Wang, Zike Yan, Qi Wu, Zhihua Wei, and Jing Liu. COSMO: Combination of selective memorization for low-cost vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5511–5522, 2025c. URL https://arxiv.org/abs/2503.24065.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025d. URL https://arxiv.org/abs/2410.02713.

Gengze Zhou, Yicong Hong, and Qi Wu. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7641–7649, 2024. doi: 10.1609/aaai.v38i7.28597. URL https://ojs.aaai.org/index.php/AAAI/article/view/28597.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Zhu_Vision-Language_Navigation_With_Self-Supervised_Auxiliary_Reasoning_Tasks_CVPR_2020_paper.pdf.

A Training and Implementation Details

A.1 Three-Stage Training and Baseline Definition

Training starts from Qwen3-VL-8B-Instruct and follows Table 5. Stages 1 and 2 use expert and DAgger navigation data, respectively, each mixed with approximately 0.79M VQA examples from LLaVA-Video-178K and VSI-590K. Guidance is introduced only in Stage 3, using the original R2R and English RxR expert data enriched with our annotations (Appendix B); this stage excludes DAgger and VQA data.

Table 5: **Three-Stage Training Recipe.** Every stage uses one epoch and global batch size 256. Stages 1 and 2 use full fine-tuning; Stage 3 adapts the policy with rank-32 LoRA and scaling factor 16.

	Stage 1	Stage 2	Stage 3 (GuidedNav)
Initialization	Qwen3-VL-8B-Instruct	Stage 1 checkpoint	Stage 2 checkpoint
Navigation data	R2R + RxR experts	DAgger	Annotated R2R + RxR experts
VQA data	Included	Included	Not included
Policy update	Full parameters	Full parameters	LoRA adapters
Peak policy LR	2×10^{-5}	1×10^{-5}	2×10^{-4}
Our guidance	None	None	Subtask, landmark, spatial

Optimization. All stages use AdamW, cosine decay, 3% learning-rate warmup, weight decay 0.01, BF16, and DeepSpeed ZeRO-2. Stages 1 and 2 use Adam coefficients $(0.9, 0.999)$, maximum gradient norm 1.0, and a peak learning rate of 5×10^{-6} for the multimodal projector and vision tower. Their maximum sequence length is 8,192 tokens, with image pixel bounds of $28^2 \times 16$ and $28^2 \times 256$. The random seed is 42.

Base Model and component comparisons. Base Model applies one epoch of Stage-3 LoRA fine-tuning to the Stage-2 model on the same R2R and RxR expert data, without the proposed guidance modules. It shares GuidedNav's adapter rank, scaling factor, learning rate, global batch size, and epoch budget. This baseline is used for component ablations, route-length analysis, and representation comparisons. Component variants add the indicated modules; GuidedNav combines all three.

A.2 Spatial Architecture and Auxiliary Settings

Figure 8 summarizes long-horizon alignment and local refinement. The long-horizon teacher receives the current frame and up to seven sampled past observations; alignment covers the full current patch grid. The local branch uses the current frame and up to three preceding observations. Both branches use only observations available at time t.

Auxiliary supervision. Attention is read from decoder block 27 (one-based), heads 0 and 16 (zero-based), and normalized separately over instruction tokens and current-image patches. Spatial supervision uses the output hidden states of the same block. The weights $(\lambda_{\text{sub}}, \lambda_{\text{lm}}, \lambda_{\text{long}}, \lambda_{\text{local}})$ are $(0.01, 0.01, 0.5, 0.1)$, with 3% auxiliary warmup. Missing annotations disable only the corresponding losses.

Teacher features and grid alignment. Both branches use frozen VGGT-1B, concatenating frame and global features from its final aggregator output and removing camera and register tokens to obtain 2048-dimensional patch features. The teacher's 14-pixel patch grid is reshaped spatially and bilinearly resized to the policy's post-merger visual-token grid, obtained from its image-grid metadata and 2×2 merge factor. The long-horizon target adds VGGT's UV positional embedding with scale 0.1 before resizing; local and single-frame targets omit this embedding. Separate two-layer MLPs with GELU map the 4096-dimensional student features to 2048 dimensions.

Local geometric target. Features from the local window and current frame alone are independently normalized before subtraction, and the difference is normalized again:

$$\Delta f_{t,j} = \mathcal{N}\Big(\mathcal{N}(f_{t,j}^{\text{local}}) - \mathcal{N}(f_{t,j}^{\text{single}})\Big), \tag{9}$$

where $\mathcal{N}$ denotes ℓ_2 normalization.

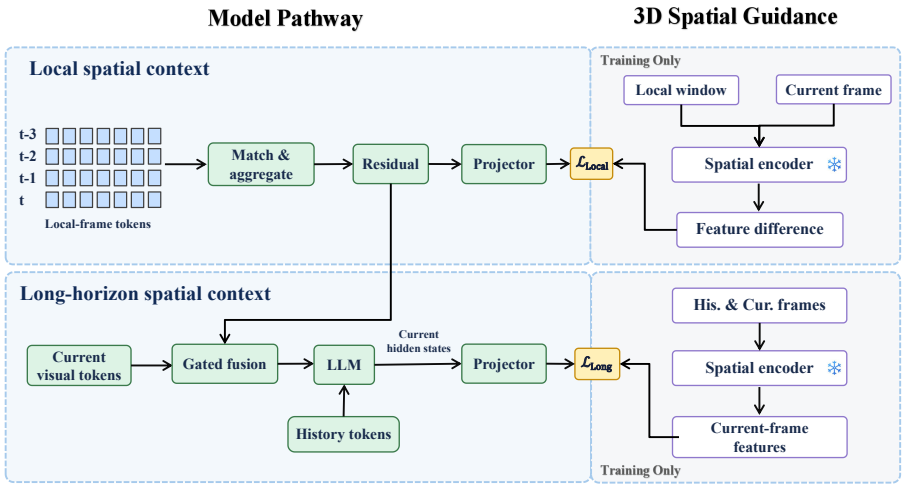

Figure 8: **Long-Horizon Spatial Guidance and Local Spatial Refinement.** The upper branch supervises matched-token residuals with $\mathcal{L}_{\text{Local}}$ and fuses them into current visual tokens. The lower branch aligns current hidden states with history-conditioned teacher features through $\mathcal{L}_{\text{Long}}$. The teacher and auxiliary projectors are used only during training; temporal fusion remains active at inference.

Sparse temporal matching. For a past-frame token i and current token j, matching uses detached features and the score

$$s_{u,i,j} = \cos(z_{u,i}, z_{t,j}) - 0.05\|q_i - q_j\|_2^2, \tag{10}$$

where q_i, q_j are normalized coordinates on the resized token grid. A pair is accepted only if each token is the other's highest-scoring match. We retain at most $\text{round}(0.25N_t)$ highest-scoring mutual pairs per past frame, with weights $0.2, 0.35, 0.5$ for $t-3, t-2, t-1$, omitting unavailable frames. Gradients pass through selected features but not discrete matching. The scalar gate is $g = 0.2\,\text{sigmoid}(a)$, initialized to 0.05. The local loss supervises the residual before gating; action and other downstream objectives train the gate through the fused tokens.

Subtask interval balancing. For each source dataset, let V count valid span-labeled examples, G count trajectory–subtask groups, and n_g count valid examples in group g. Equation 3 uses $c_t = V/(Gn_g)$, giving each interval equal total weight and each valid example unit weight on average. Landmark supervision is averaged over valid frame-level masks without interval balancing.

Table 6: **Simulator and Action Settings.** The common action settings apply to both datasets; collision-body dimensions shown below are the RxR-specific settings.

Setting	Value
Decoding	Greedy
Maximum sampled context images	8
Executed action chunk	Up to 4 primitives
Forward displacement	0.25 m
Turn angle	15°
Maximum episode actions	500
Sliding	Enabled
RGB camera height	1.25 m
RxR collision-body height / radius	0.88 m / 0.18 m

B LANDMARK-CENTRIC DATA ENGINE

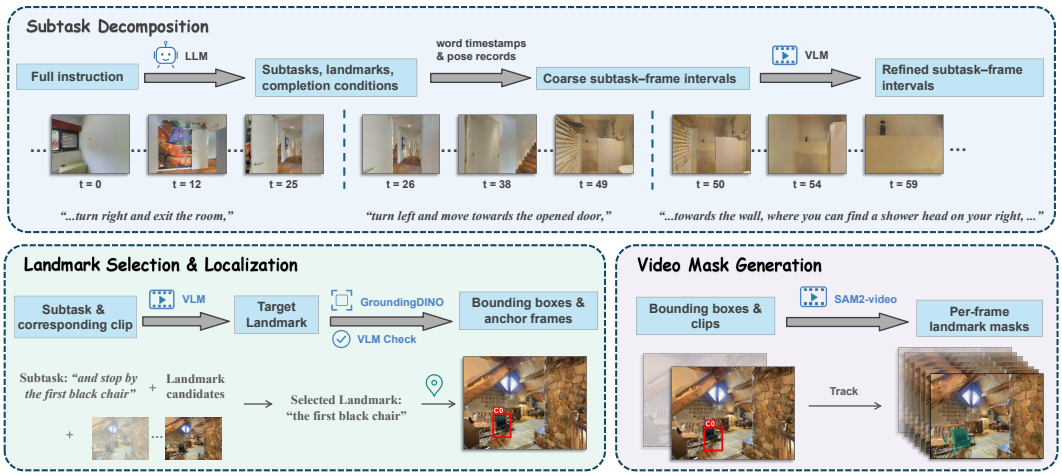

Figure 9: **Offline Subtask and Landmark Annotation.** Subtask intervals connect instruction spans to observation clips. GroundingDINO proposes landmark boxes, GLM-4.5V verifies instances and anchor frames, and SAM2 propagates accepted prompts within each interval. R2R reuses FGR2R spans; English RxR decomposition uses gpt-oss-120b, followed by temporal refinement with GLM-4.5V. These labels supervise training and are not policy inputs at inference.

Guided-R2R and Guided-RxR augment R2R and RxR (Anderson et al., 2018; Ku et al., 2020) with active subtask spans and landmark masks (Figure 9). R2R reuses FGR2R subtasks (Hong et al., 2020); gpt-oss-120b decomposes English RxR training instructions into spans, landmarks, and completion conditions. RxR word timestamps and guide poses provide coarse temporal alignment, which GLM-4.5V refines using trajectory overviews and frames near subtask boundaries. For each aligned clip, GLM-4.5V selects explicitly mentioned landmarks or visually supported structures implied by the action, verifies GroundingDINO proposals (Liu et al., 2024), and selects anchor frames. SAM2 (Ravi et al., 2025) propagates accepted box prompts within each interval. Masks are filtered by area, detection consistency, and temporal support; the original observations, instructions, and action targets are preserved.

Mapping R2R subtasks to continuous trajectories. We match exported episodes to FGR2R instructions and reference paths. Reference-path anchors are associated with their closest recorded continuous poses using Euclidean distance. Anchor-arrival indices are made nondecreasing when necessary; the FGR2R subpath boundaries then define the per-frame subtask assignments. Exports with missing poses, ambiguous matches, or incompatible path topology are excluded from this alignment.

Annotation access. Offline annotation can inspect an entire expert trajectory to establish temporal intervals. The policy and spatial teacher use only current and past observations at each decision. Samples without reliable spans or masks contribute to action supervision, with the unavailable auxiliary terms omitted.

B.1 GUIDED-R2R AND GUIDED-RXR EXAMPLES

Figures 10 and 11 link subtask spans, frame intervals, and landmark masks by color. The R2R hallway interval illustrates missing mask coverage: action supervision is retained, while landmark supervision is omitted.



Figure 10: **Guided-R2R Annotation Example.** Three subtask spans align hallway traversal, a right turn, and stopping by a black chair with their frame intervals. Representative observations show the selected landmarks and whether a mask was accepted.

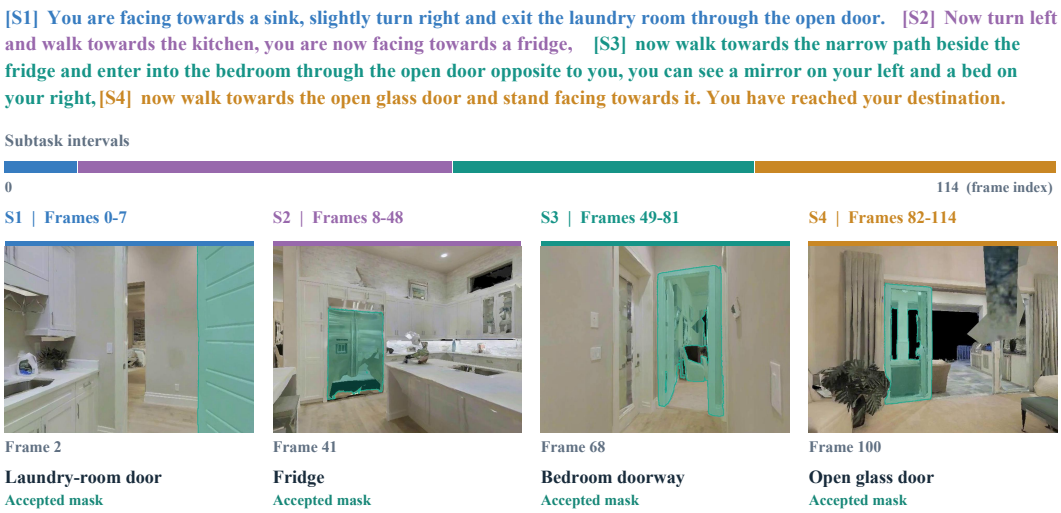

Figure 11: **Guided-RxR Annotation Example.** Four subtask spans follow a route from a laundry room through a kitchen and bedroom to a glass door. Matching colors link instruction text, frame intervals, and representative landmark masks.

C Further Analysis of the Guidance Modules

C.1 Subtask Attention Guidance

Table 7 compares textual supervision with native attention guidance. *Sequential Generation* predicts the active subtask before the action sequence, using teacher forcing and cross-entropy weights of 0.1 and 1.0 for subtask and action tokens, respectively, normalized by total token weight. *Auxiliary Generation* uses separate action and subtask forward passes through the shared policy, with independently normalized losses and a text-loss weight of 0.1 warmed up at the start of training. Its generated text is not fed to the action branch, and inference uses only the action pass. *Attention Guidance* supervises pre-action attention to the active span. It gives the highest SR and lowest NE, while Auxiliary Generation gives the highest OSR and Base Model the highest SPL. Figure 4 visualizes attention as instruction progress changes.

Table 7: **Subtask Supervision Strategies on R2R-CE Val-Unseen.**

Method	SR ↑	SPL ↑	OSR ↑	NE ↓
Base Model	59.3	**56.1**	63.8	4.87
Sequential Generation	60.4	55.4	66.7	4.43
Auxiliary Generation	62.0	55.6	**72.5**	4.55
Attention Guidance	**62.9**	55.2	72.3	**4.34**

C.2 Landmark Attention Guidance

Table 8 compares landmark supervision strategies. Our objective increases attention mass within accepted landmark regions without fixing its distribution over patches. It gives the highest SR and OSR and lowest NE, while Base Model retains the highest SPL. Figure 12 illustrates localization relative to the reference masks.

Table 8: **Landmark Attention Supervision on R2R-CE Val-Unseen.** We compare Distribution Matching, Access Regularization, and our Attention Guidance objective. The latter maximizes attention mass supported by the landmark mask. SR, SPL, and OSR are percentages; NE is in meters. Bold denotes the best value in each column.

Method	SR ↑	SPL ↑	OSR ↑	NE ↓
Base Model	59.3	**56.1**	63.8	4.87
Distribution Matching	61.6	55.2	71.2	4.48
Access Regularization	62.4	55.8	71.5	4.40
Attention Guidance (Ours)	**62.6**	55.8	**72.1**	**4.38**

C.3 Quantitative Attention Localization

Table 4 evaluates correspondence with automatically annotated instruction spans and landmarks. For subtasks, active-span tokens are positives and attention weights provide scores for area under the precision–recall curve (AUPRC) and receiver operating characteristic curve (AUROC) (Davis & Goadrich, 2006). Pointing Game (PG) (Zhang et al., 2016) checks whether the peak-attention location lies inside a landmark, while patch-level Energy-Based Pointing Game (EBPG) (Wang et al., 2020) measures the fraction of attention mass supported by the landmark mask. All four metrics improve for GuidedNav (Table 4). Attention is normalized separately over instruction tokens and current-image patches, so each metric describes localization within its corresponding token group.



Figure 12: **Landmark Attention Localization.** Instruction excerpts and reference masks are compared with Base Model and GuidedNav pre-action attention for stairs, chairs, and an oven. Heatmaps cover the full current image.

C.4 Spatial Representation Probing

We fit depth probes independently to frozen Base Model and GuidedNav representations; the probes are excluded from action prediction. Table 4 reports absolute relative error (AbsRel) and root mean squared error (RMSE) (Eigen et al., 2014), together with near–far ranking accuracy, the fraction of correctly ordered depth pairs inspired by relative-depth evaluation (Chen et al., 2016). Lower errors and higher ranking accuracy indicate that geometric information is more accessible under the fitted probes.

Dense depth visualization. For qualitative visualization, a separate convolutional head decodes frozen current-frame features from decoder block 27 into 512×384 depth maps. The visualization decoder is trained for one epoch on 10,000 R2R training frames, with 1,000 held-out frames from disjoint scenes. Figure 6 compares the decoded scene structure with reference depth. The visualization decoder is an analysis tool and adds no cost to navigation inference.

D Additional Navigation Visualizations

D.1 Simulation Rollouts

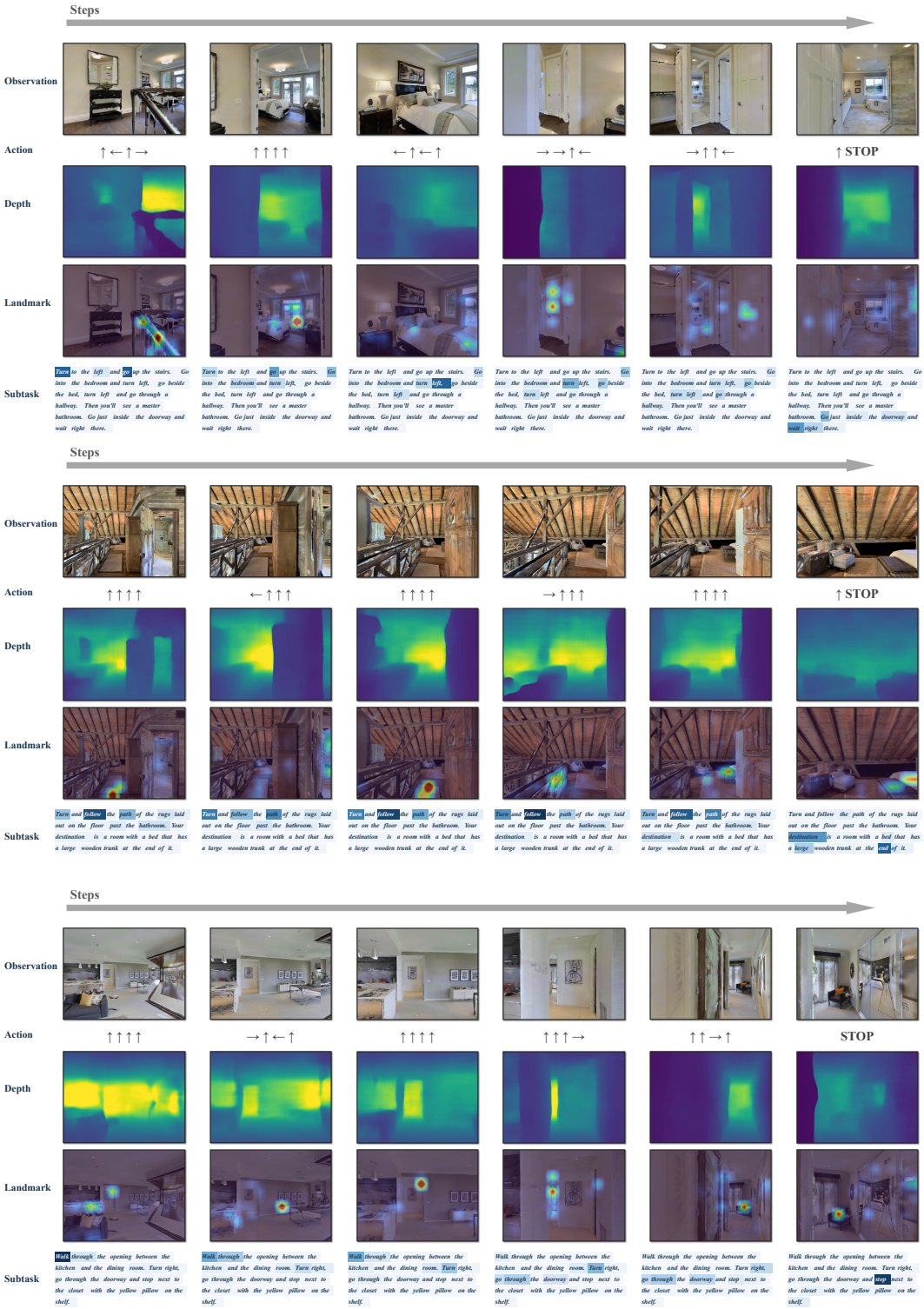

Figure 13: **Successful R2R-CE Rollouts: Cases 1-3.** Columns follow six sampled decisions along each trajectory. Rows show observations, predicted action chunks, decoded depth, landmark attention, and instruction attention; darker instruction highlighting indicates greater relative attention.

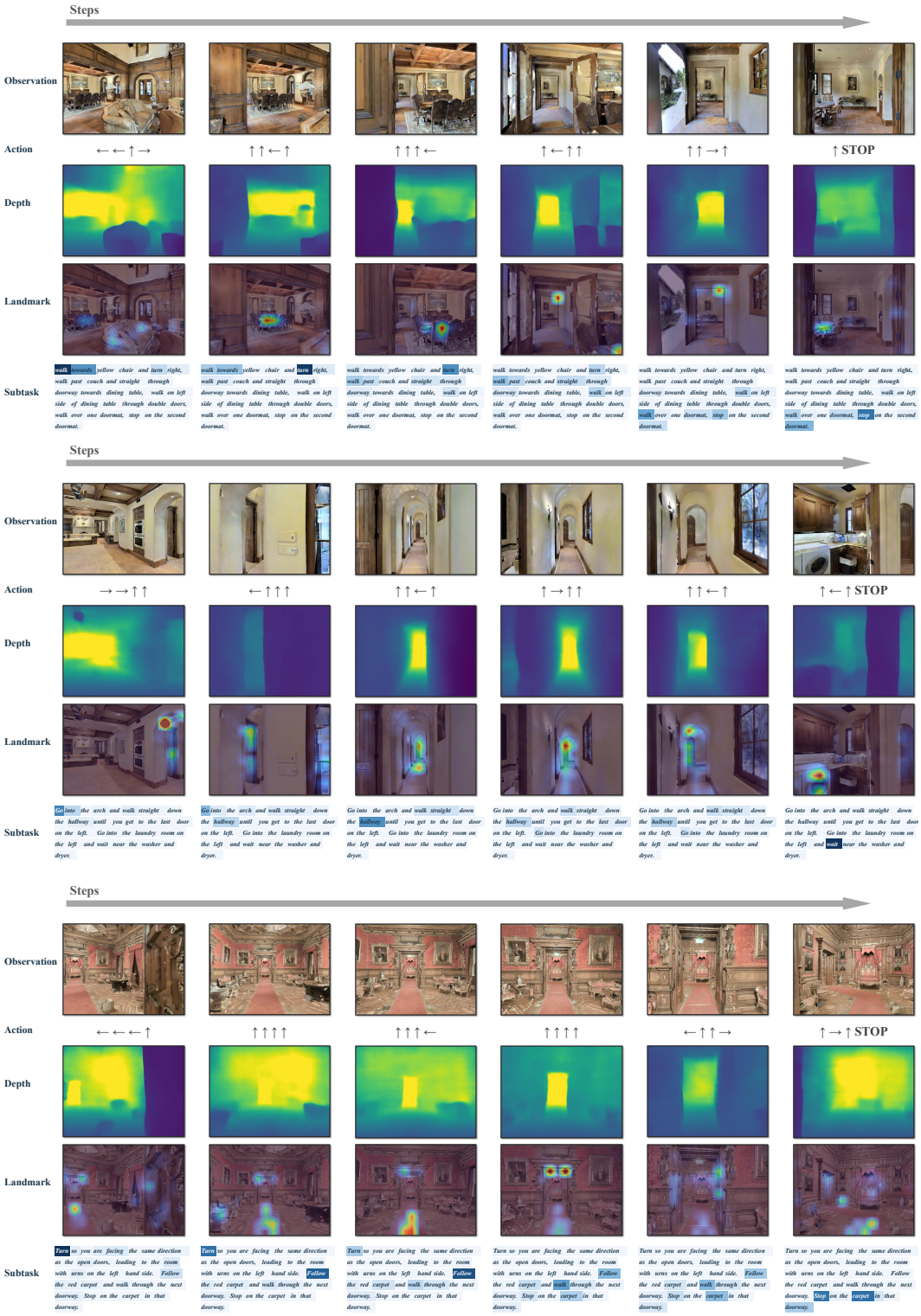

Figure 14: **Successful R2R-CE Rollouts: Cases 4-6.** Six sampled decisions per trajectory, following the conventions in Figure 13.

D.2 Real-World Setup and Rollouts

GuidedNav runs on an AgiBot X2 humanoid with a forward-facing RGB camera, receiving instructions and streaming egocentric observations. An off-board NVIDIA RTX 4090 server transmits predicted actions to the robot; no real-world fine-tuning is used. Depth is decoded only for visualization with the separate probe described in Appendix C.4, and attention uses the readout in Appendix A. Figure 7 shows navigation past orange chairs to a coffee machine; Figure 15 adds two trajectories in other scenes. These examples provide qualitative evidence of transfer to real observations.



Figure 15: **Additional Real-World Navigation Examples.** Six sampled decisions per trajectory show observations, predicted actions, decoded depth, and native pre-action attention. The visualization follows the conventions in Figure 7.

E Comparisons with Additional Observation Modalities

Table 9 compares methods using panoramic views or additional sensory inputs. Their observation configurations differ from GuidedNav's monocular RGB input; monocular RGB results are reported in Table 1. All baseline navigation results are taken from the respective papers.

Table 9: **Additional Navigation Comparisons.** Methods using panoramic observations or additional sensory inputs. P: panoramic RGB; R: monocular RGB; O: odometry; D: depth. NE is in meters; all other metrics are percentages.

(a) R2R-CE val-unseen

Method	Inputs	NE↓	OSR↑	SR↑	SPL↑
HPN+DN (Krantz et al., 2021)	P+O+D	6.31	40.0	36.0	34.0
CMA (Hong et al., 2022)	P+O+D	6.20	52.0	41.0	36.0
Sim2Sim (Krantz & Lee, 2022)	P+O+D	6.07	52.0	43.0	36.0
VLN↻BERT (Hong et al., 2022)	P+O+D	5.74	53.0	44.0	39.0
Ego^2-Map (Hong et al., 2023)	P+O+D	5.54	56.0	47.0	41.0
DreamWalker (Wang et al., 2023a)	P+O+D	5.53	59.0	49.0	44.0
GridMM (Wang et al., 2023b)	P+O+D	5.11	61.0	49.0	41.0
Reborn (An et al., 2022)	P+O+D	5.40	57.0	50.0	46.0
InstructNav (Long et al., 2025)	P+O+D	6.89	–	31.0	24.0
COSMO (Zhang et al., 2025c)	P	–	56.0	47.0	40.0
AO-Planner (Chen et al., 2025)	P+D	5.55	59.0	47.0	33.0
LAW (Raychaudhuri et al., 2021)	O+D+R	6.83	44.0	35.0	31.0
MapNav (Zhang et al., 2025b)	O+D+R	4.93	53.0	39.7	37.2
g3D-LF (Wang & Lee, 2025)	O+D+R	5.70	59.5	47.2	34.6
Seq2Seq (Krantz et al., 2020)	D+R	7.77	37.0	25.0	22.0
NaVid-4D (Liu et al., 2025)	D+R	5.99	55.7	43.8	37.1
NavMorph (Yao et al., 2025)	D+R	5.75	56.9	47.9	33.2
GA-VLN (Yang et al., 2026)	O+D+R	4.80	67.6	61.0	55.2

(b) RxR-CE val-unseen

Method	Inputs	NE↓	SR↑	SPL↑	nDTW↑
CMA (Hong et al., 2022)	P+O+D	8.76	26.5	22.1	47.0
VLN↻BERT (Hong et al., 2022)	P+O+D	8.98	27.0	22.6	46.7
Reborn (An et al., 2022)	P+O+D	5.98	48.6	42.0	63.3
AO-Planner (Chen et al., 2025)	P+D	7.06	43.3	30.5	50.1
LAW (Raychaudhuri et al., 2021)	O+D+R	10.90	8.0	8.0	38.0
Seq2Seq (Krantz et al., 2020)	D+R	12.10	13.9	11.9	30.8
NavMorph (Yao et al., 2025)	D+R	8.85	30.8	22.8	44.2
GA-VLN (Yang et al., 2026)	O+D+R	5.88	55.4	45.2	–

Results are from the respective papers. "–" denotes an unavailable metric, not zero; methods without reported RxR results are omitted from panel (b). Comparable external-training-data counts are unavailable for these entries; GA-VLN uses additional navigation datasets. Monocular RGB-only results appear in Table 1.