Dual-Anchoring: Addressing State Drift in Vision-Language Navigation

Kangyi Wu[1†], Pengna Li[1†], Kailin Lyu[2‡], Xi Lin[3‡], Lin Zhao[4§], Qingrong He[4], Jinjun Wang[15], and Jianyi Liu[1*]

[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Johns Hopkins University
[4] Joy Future Academy, JD
[5] Xpeng Inc.

Abstract. Vision-Language Navigation (VLN) requires an agent to navigate through 3D environments by following natural language instructions. While recent Video Large Language Models (Video-LLMs) have largely advanced VLN, they remain highly susceptible to *State Drift* in long scenarios. In these cases, the agent's internal state drifts away from the true task execution state, leading to aimless wandering and failure to execute essential maneuvers in the instruction. We attribute this failure to two distinct cognitive deficits: *Progress Drift*, where the agent fails to distinguish completed sub-goals from remaining ones, and *Memory Drift*, where the agent's history representations degrade, making it lose track of visited landmarks. In this paper, we propose a Dual-Anchoring Framework that explicitly anchors the instruction progress and history representations. First, to address progress drift, we introduce Instruction Progress Anchoring, which supervises the agent to generate structured text tokens that delineate completed versus remaining sub-goals. Second, to mitigate memory drift, we propose Memory Landmark Anchoring, which utilizes a Landmark-Centric World Model to retrospectively predict object-centric embeddings extracted by the Segment Anything Model, compelling the agent to explicitly verify past observations and preserve distinct representations of visited landmarks. Facilitating this framework, we curate two extensive datasets: 3.6 million samples with explicit progress descriptions, and 937k grounded landmark data for retrospective verification. Extensive experiments in both simulation and real-world environments demonstrate the superiority of our method, achieving a 15.2% improvement in Success Rate and a remarkable 24.7% gain on long-horizon trajectories. To facilitate further research, we will release our code, data generation pipelines, and the collected datasets.

Keywords: Vision-Language Navigation · Video-LLMs · World Model

† Equal contribution.
‡ These authors contributed equally.
§ Project Leader.
* Corresponding author.

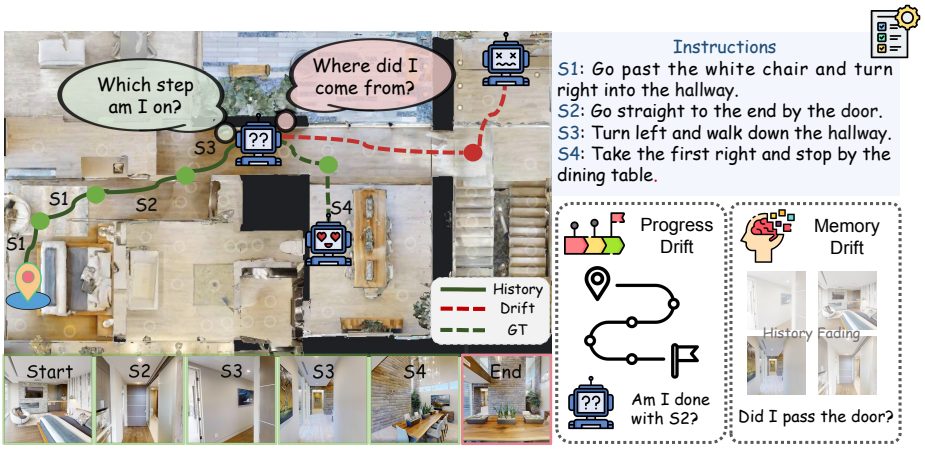

Fig. 1: The Challenge of State Drift. As the trajectory extends, the agent's predicted path deviates from the ground truth due to internal state decoupling. This manifests as the Progress Drift (confusion about which instruction step is active) and Memory Drift (failure to remember visited landmarks).

1 Introduction

Vision-Language Navigation (VLN) has emerged as a central challenge in embodied artificial intelligence [14,27], requiring an agent to follow natural language instructions to navigate in a previously unseen environment. Due to its profound potential for real-world applications in domestic service [15, 48], intelligent personal assistants [40, 41], etc, this task has garnered significant attention and witnessed rapid development in recent years.

Most VLN works follow a perception-to-action pipeline: the agent conditions on egocentric observations and the instruction to predict an action, iterating until reaching the goal. While effective for short instructions, this stepwise policy is brittle under long, compositional instructions. As navigation proceeds, agents may exhibit severe **State Drift**: their task state becomes progressively unreliable, and the agent gradually loses a consistent sense of (i) where it is in the instruction and (ii) where it has been in the environment. In other words, the internal state increasingly drifts away from the true task state, leading to erratic behaviors. As illustrated in Fig. 1, we attribute State Drift to two coupled failure modes: (1) **Progress Drift**, where instruction stage is mis-tracked and the boundary between completed and remaining sub-goals becomes blurred; and (2) **Memory Drift**, where history representations degrade over long trajectories, causing the agent to lose track of visited landmarks.

Recent Video-LLMs [17, 32] have advanced VLN [13, 29, 30, 34, 39, 47, 54] by leveraging powerful pre-trained representations. However, they remain prone to state drift in long-horizon scenarios. Fundamentally, these models rely on next-action prediction, a result-oriented objective that enforces correct local moves

but neglects the agent's internal cognitive process. This supervision allows the agent to "guess" the right action based on shallow local cues. However, without explicit constraints on the internal states, they will inevitably decouple from physical reality as navigation proceeds.

To address these challenges, we argue that robust navigation needs explicit state anchoring for instruction progress and visual history. For the former, we use structured language as a compact representation of instruction execution status. Instead of relying on implicit attention to infer progress, the agent explicitly maintains a "mental checklist", structurally delineating completed sub-goals from remaining ones. For the latter, we introduce a retrospective constraint: robust state grounding requires not only forward-looking control but also explicit verification of past observations. We therefore compel the model to preserve distinct, landmark-centric history representations to firmly anchor its current position.

In this paper, we propose the Dual-Anchoring Framework, which regularizes the Video-LLM's internal state through two complementary branches. First, to explicitly track instruction progress, we introduce **Instruction Progress Anchoring**. By designing a Progress Annotation Generation pipeline, we curate 3.6 million navigation samples where each action is prefixed with a structured progress description, explicitly delineating completed versus remaining sub-goals. Then, we utilize Instruction-Aware Co-training to force the model to articulate a concrete linguistic state before executing any control. Second, to enforce historical verification, we propose **Memory Landmark Anchoring**. To realize this, we devise a Landmark Frame Mining method to collect 937k grounded landmark data. Then, we design a Landmark-Centric World Model that operates retrospectively: it reconstructs the high-level semantic features of the most recently passed landmark from output tokens of history frames. This module acts as a "rear-view mirror," compelling the model to preserve distinct, grounded representations of its history.

Our main contributions are summarized as follows:

- We identify State Drift as a critical bottleneck in VLN and propose the Dual-Anchoring Framework, a unified solution that explicitly regularizes the agent's internal state, greatly eliminating erratic behaviors.
- We introduce Instruction Progress Anchoring, backed by a scalable pipeline that synthesizes 3.6 million structured samples to enforce explicit instruction state tracking. Complementarily, we design a Memory Landmark Anchoring that collects 937k grounded landmark data and leverages SAM-based retrospective prediction to anchor the agent's visual memory.
- We conduct extensive experiments on VLN-CE benchmarks, where our method achieves state-of-the-art (SOTA) performance. Furthermore, real-world deployment demonstrates the framework's superior robustness and generalization in complex, unseen environments.

2 Related Works

2.1 Vision Language Navigation.

Vision-Language Navigation in Continuous Environments (VLN-CE) [25] requires agents to interpret instructions and execute low-level actions in physics-realistic 3D spaces. Unlike graph-based settings [10,11,19,57], VLN-CE presents significant challenges in aligning continuous perception with long-term planning. The emergence of Video Large Language Models (VideoLLM) [4,5,17,32] has bifurcated recent solutions into Hierarchical Planners and End-to-End Navigators. Hierarchical methods [13] decouple reasoning and control, treating the VLM as a "slow system" for mid-level plans. Conversely, End-to-End models [47,54] fine-tune Video-LLMs to map streaming observations directly to actions, leveraging the base model's strong generalization. However, both approaches remain susceptible to state drift during extended trajectories. Without explicit state management, the agent's internal representation tends to decouple from the physical reality. In this work, we mitigate this drift by enforcing explicit regularization on the decision process: we supervise the agent to articulate structured progress descriptions for semantic alignment, and leverage an object-centric objective to predict historical features for visual grounding.

2.2 Multi-task Training for VLN.

Standard VLN training relies on human-annotated navigation datasets like R2R [3] and RxR [26]. However, the scarcity of such high-quality trajectory data limits the agent's ability to generalize across diverse unseen environments. To address this, the community has embraced Multi-task Training, broadly categorized into two paradigms. The first focuses on Broad Visual Generalization, where agents are co-trained on massive synthetic environments [13] or real-world human videos [12] to master diverse visual distributions. The second emphasizes Reasoning Alignment, employing auxiliary tasks like Hierarchical Planning [34] or Chain-of-Thought (CoT) generation [21] to rationalize the perception-action loop. However, these auxiliary tasks remain predominantly action-centric or instantaneous, focusing primarily on interpreting the current scene for immediate decision-making. Consequently, they often neglect the explicit alignment between the accumulated history and the instruction, leaving agents vulnerable to Progress Drift. Our Instruction Progress Anchoring shifts focus from local perception to global status tracking, strictly enforcing a structural distinction between "Done" and "Remaining" parts to lock the decision process. Similar auxiliary supervision strategies have been explored more broadly in multimodal learning, aiming to stabilize optimization and improve robustness beyond direct end-task prediction [28, 31, 33, 35–37, 49, 50].

2.3 World Model for VLN.

World models serve as internal representations of the environment, enabling agents to perform mental simulation by predicting future states. Early approaches,

such as NWM [6] and AstraNav-World [20], operate as Generative Simulators, synthesizing high-fidelity future video streams to learn general physical priors. However, pixel-level generation is computationally prohibited for real-time onboard inference. To address this efficiency bottleneck, recent works like LS-NWM [56], NavMorph [51], and NavForesee [34] shift towards Feature-level Prediction, forecasting compact latent dynamics or semantic features instead of raw images. However, these methods remain fundamentally future-centric. They prioritize foresight—anticipating what will happen next—while neglecting the explicit maintenance of what has already happened. In long trajectories, this lack of historical grounding leaves agents vulnerable to Memory Drift, where the history representation fades. Diverging from this foresight-dominant paradigm, we propose a Landmark-Centric World Model that prioritizes hindsight. Instead of hallucinating uncertain futures, we leverage the Segment Anything Model to reconstruct the object-centric features of previously visited landmarks. This retrospective objective compels the agent to anchor its internal state to the concrete visual history, effectively mitigating perceptual aliasing.

3 Method

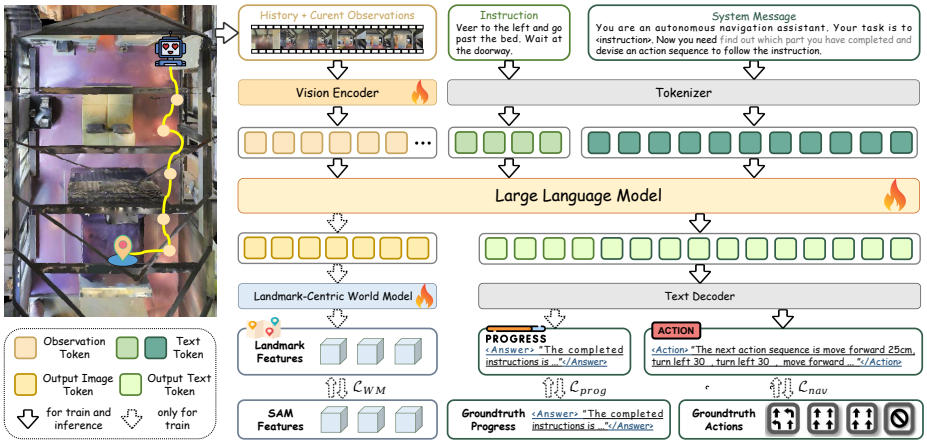

Fig. 2: Overview of the Dual-Anchoring Framework. The model processes natural language instructions and streaming egocentric observations through a unified Video-LLM backbone. Beyond standard action prediction, the framework incorporates two auxiliary tasks: generating structured progress descriptions to ensure semantic alignment, and reconstructing dense SAM features to anchor the visual memory.

3.1 Problem Formulation and Overall Architecture

Problem Formulation. We formulate the Vision-Language Navigation (VLN) task as a sequential decision-making process in a continuous 3D environment.

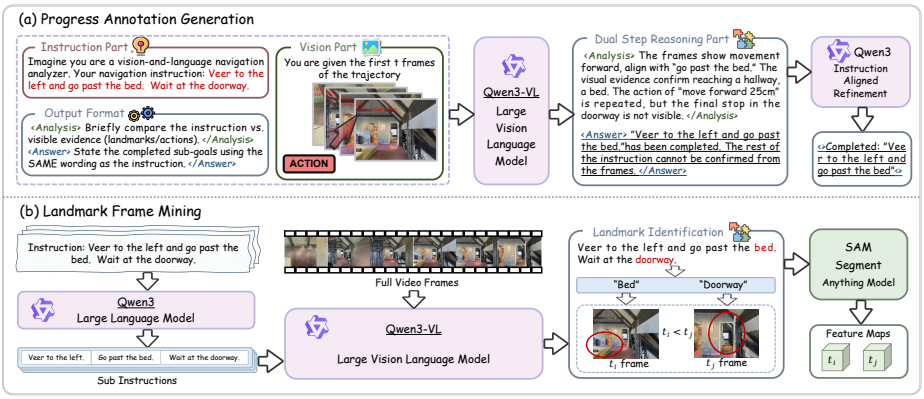

Fig. 3: Overview of the data generation pipeline. (a) A multimodal LLM performs dual-step reasoning to generate 3.6 million progress descriptions that are strictly aligned with instruction wording. (b) A hierarchical approach decomposes instructions into atomic sub-instructions and identifies temporal landmarks. The Segment Anything Model then extracts 937k object-centric features to serve as retrospective ground truth.

The agent is initialized at a starting position and provided with a natural language instruction $\mathcal{I} = \{w_1, w_2, \ldots, w_L\}$. At each time step t, the agent perceives the environment via a monocular RGB observation o_t. Based on the instruction $\mathcal{I}$ and the history of observations, the agent predicts a low-level navigational action $a_t \in \mathcal{A}$ (e.g., `move_forward`, `turn_left`, `turn_right`,`stop`). Success is measured by whether the agent stops within a threshold distance of the target.

We adopt the architecture of StreamVLN [47] as our backbone. This framework extends standard Video-LLMs to support streaming input by maintaining a history context $\mathcal{H}_t$ that aggregates history information. The agent predicts the next action a_t autoregressively by conditioning on the instruction and this history context:

$$P(a_t|\mathcal{I}, \mathcal{H}_t) = \text{VideoLLM}_\theta(\mathcal{I}, \mathcal{H}_t), \tag{1}$$

where θ is the trainable parameter. Despite the efficiency of this streaming architecture, relying solely on implicit hidden states of $\mathcal{H}_t$ to track long-horizon progress remains challenging, motivating the need for explicit state anchoring.

Overall Architecture. The framework is illustrated in Fig. 2. To explicitly combat the State Drift dilemma in VLN, we propose the **Dual-Anchoring Framework**. While maintaining the efficient streaming backbone of StreamVLN for low-latency action generation, we augment the training architecture with Instruction Progress Anchoring and Memory Landmark Anchoring to prevent the agent's mental state from decoupling from the physical reality. The former is designed as a Co-Training Task while the latter as auxiliary head, thus incurring no additional computation for deployment.

3.2 Instruction Progress Anchoring

To mitigate Progress Drift, the agent must explicitly maintain a mental checklist of its execution status. However, standard VLN datasets only provide coarse-grained trajectory-level instructions, lacking step-by-step state annotations. To bridge this gap, we introduce an automated pipeline to synthesize fine-grained progress description data at scale.

Progress Annotation Generation. We leverage a powerful Multimodal LLM (Qwen3-VL) as an offline annotator to generate textual progress summaries y_t^{prog} given a ground-truth trajectory $\tau = \{(o_t, a_t)\}_{t=0}^{T}$ and its instruction $\mathcal{I}$. As illustrated in Fig. 3, the pipeline operates in four sequential steps. *1) Visual Kinematics Prompting:* To bridge the gap between static frames and dynamic navigation, we explicitly overlay the frame index and executed action text (e.g., "Turn Left 15°") onto the top-left corner of each image frame. This explicitly informs the annotator of the agent's kinematics between timestamps. *2) Interval Sampling:* We perform interval sampling along the trajectory with a stride of n. For each sampled step t, the annotator receives the instruction $\mathcal{I}$ and the processed visual history sequence $o_{0:t}$. *3) Dual-Step Reasoning:* To ensure high-quality annotation, we prompt the model to perform a chain-of-thought process: first analyzing the visual evidence against the instruction , and then generating a concise summary of *Completed Sub-goals* using the original wording. *4) Instruction-Aligned Refinement:* Finally, to eliminate intermediate reasoning and possible hallucinations, we prompt the LLM to distill the analysis into a single, concise sentence. This step strictly constrains the output to be a verbatim prefix of the instruction. In this way, we collect a large-scale dataset comprising 3.6 million progress description samples.

Instruction-Aware Co-training. During training, we treat this as a co-training task, where the ground-truth action is prefixed with the synthesized progress descriptions. In this case, we add "find out which part you have completed" in the Prompt. The generation objective at step t is modified to maximize $P(y_t^{prog}, a_t | \mathcal{I}, \mathcal{H}_t)$, forcing the agent to articulate its status before action.

3.3 Memory Landmark Anchoring

While Instruction-Aware Co-training explicitly track instruction progress, the agent still suffers from Memory Drift—forgetting distinct landmarks and falling into perceptual aliasing. Therefore, we propose Memory Landmark Anchoring that enforces *Retrospective Grounding*.

Landmark Frame Mining. Constructing supervision for this objective requires identifying *when* specific landmarks mentioned in the instruction appear in the video. To realize this, we design a two-stage extraction process. The pipeline is illustrated in Fig 3 (b). **First**, in the Decomposition stage, we prompt the LLM (Qwen3) to decompose the complex instruction $\mathcal{I}$ into a sequence of atomic

sub-goals $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, where each s_k contains a specific action or landmark. **Second**, during Temporal Grounding, we feed the full video and $\mathcal{S}$ to the MLLM (Qwen3-VL) to identify the *first* appearance frame $t_{\text{lm}}^{(k)}$ for the landmark in each s_k. To eliminate hallucination, we enforce a strict ordering constraint: the appearance times must be strictly increasing ($t_{\text{lm}}^{(i)} < t_{\text{lm}}^{(j)}$ for $i < j$), and any annotation violating this temporal logic is filtered out. For any given time step t during navigation, this process allows us to retrieve the most recently passed landmark frame index t^* (where $t^* \leq t$). To construct the dense supervisory signal, we extract the high-resolution spatial feature map $F_{SAM}(o_{t^*})$ from this identified frame using the Segment Anything Model (SAM) [22]. This object-centric feature serves as the retrospective ground truth for our world model.

Landmark-Centric World Model. To effectively mitigate Memory Drift, the agent must maintain a continuous awareness of where it has been. Our World Model achieves this by compelling the agent to reconstruct the dense spatial features of the most recently passed landmark.

To realize that, we utilize a **Learnable Spatial Query Decoder**. Let $X_t \in \mathbb{R}^{N \times d_{llm}}$ denote the output sequence of the VideoLLM at step t, which encapsulates both historical and current visual semantics. To reduce computational overhead, we first project X_t into a compact latent space via a linear layer with LayerNorm:

$$\hat{X}_t = \text{LayerNorm}(X_t W_{in}) \in \mathbb{R}^{N \times d_{attn}}. \tag{2}$$

Next, a set of learnable spatial queries $Q_{spa} \in \mathbb{R}^{(H \times W) \times d_{attn}}$ is initialized, where each query acts as a pixel-wise anchor corresponding to the target spatial resolution $H \times W$. These queries retrieve highly relevant local spatial cues from $\hat{X}_t$ via cross-attention:

$$Z = \text{Softmax}\left(\frac{Q_{spa}\hat{X}_t^T}{\sqrt{d_{attn}}}\right)\hat{X}_t \in \mathbb{R}^{(H \times W) \times d_{attn}}, \tag{3}$$

where Z adaptively aggregates the required spatial information while preserving semantic heterogeneity. Subsequently, Z is linearly projected to the target channel dimension d_{sam} and reshaped into a 2D spatial format $F_t \in \mathbb{R}^{d_{sam} \times H \times W}$.

Finally, the objective is to minimize the Mean Squared Error (MSE) between the predicted feature map $\hat{F}_t$ and the frozen SAM feature $F_{SAM}(o_{t^*})$:

$$\mathcal{L}_{WM} = ||F_t - F_{SAM}(o_{t^*})||_2^2. \tag{4}$$

This objective acts as a "rear-view mirror", compelling the agent's internal state to retain dense and distinct object information from its past trajectory, effectively preventing the memory from decaying.

3.4 Data Composition and Training Strategy

To ensure both the robustness and generalizability of our model, we compile a comprehensive training set combining standard navigation benchmarks, our synthesized anchoring data, and general vision-language data.

Data Composition. As summarized in Table 1, our training mixture consists of four main categories. The Base Navigation data includes 180K trajectories from R2R, RxR, and EnvDrop, alongside 155K ones from a ScaleVLN subset. To explicitly enforce state anchoring, we augment these trajectories with our self-collected State Anchoring data, comprising 3.6M progress descriptions and 937K grounded landmark SAM features. Finally, the dataset is supplemented with 240K DAgger rollouts for robust policy learning, as well as 400K VideoQA pairs and 230K image-text pairs to preserve general vision-language capabilities.

Table 1: Overview of Training Data Composition and Utilization.

Dual Category	Source / Dataset	Scale	Objective	Stage
Base Navigation	R2R, RxR, EnvDrop	180K	$\mathcal{L}_{nav}$	1, 2
	ScaleVLN (HM3D Subset)	155K	$\mathcal{L}_{nav}$	1, 2
State Anchoring	**Progress Description Data**	**3.6M**	$\mathcal{L}_{prog}$	**1, 2**
	Grounded Landmark Data	**937K**	$\mathcal{L}_{WM}$	**1, 2**
Expert Correction	DAgger Rollouts	240K	$\mathcal{L}_{nav}$	2
Generalist VL	VideoQA (LLaVA-Video, ScanQA)	400K	Next-Token	2
	Interleaved Image-Text (MMC4)	230K	Next-Token	2

Two-Stage Training Pipeline. We adopt a two-stage training paradigm to enhance the agent's capabilities progressively:

Stage 1: Navigation Data Pre-training. The model is first trained on the aggregated navigation datasets augmented with our Dual-Anchoring objectives. The overall loss function is a weighted sum of the standard navigation action loss $\mathcal{L}_{nav}$, the progress generation loss $\mathcal{L}_{prog}$, and the world model MSE loss $\mathcal{L}_{WM}$:

$$\mathcal{L}_{Stage1} = \mathcal{L}_{nav} + \lambda_{prog}\mathcal{L}_{prog} + \lambda_{WM}\mathcal{L}_{WM}. \tag{5}$$

Stage 2: DAgger and Generalist Fine-tuning. To mitigate exposure bias and enhance robustness, we employ Data Aggregation (DAgger). We sample trajectories using the Stage 1 policy, collecting corrective expert actions to form a DAgger dataset of approximately 240K samples. In this stage, we co-train the model on a mixture of the navigation data from Stage 1, the newly collected DAgger data, and *General Vision-Language Data* to prevent catastrophic forgetting of pre-trained knowledge. The general data includes VideoQA samples (LLaVA-Video-178K, ScanQA) and interleaved image-text data (MMC4). The Dual-Anchoring objectives ($\mathcal{L}_{prog}$ and $\mathcal{L}_{WM}$) remain strictly active for all navigation-related batches during this final fine-tuning stage.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate our proposed method on two continuous VLN benchmarks: **R2R-CE** [25] and **RxR-CE** [26]. R2R-CE is based on the Matterport3D simulator and contains 90 scenes with 5.6k trajectories. RxR-CE is a larger dataset with more complex paths and fine-grained instructions, which is particularly suitable for evaluating the agent's ability to handle long-horizon navigation and instruction following. We report standard metrics for VLN-CE: Success Rate (**SR**), Success weighted by Path Length (**SPL**), Oracle Success Rate (**OSR**), and Navigation Error (**NE**).

Implementation Details. Our model is built upon StreamVLN [47] with LLaVA-Video [55] as the backbone. We use the AdamW optimizer with a learning rate of 1e-5. All experiments are conducted on 32 NVIDIA H200 GPUs. Detailed dataset statistics and training hyperparameters are provided in the Appendix.

Table 2: Comparison with state-of-the-art methods on R2R-CE and RxR-CE validation unseen splits. * indicates methods using the waypoint predictor from [19]. The **best** and second-best results are highlighted.

Method	Sensors				R2R-CE				RxR-CE		
	Pano.	Odo.	Depth	S-RGB	NE↓	OSR↑	SR↑	SPL↑	NE↓	SR↑	SPL↑
HPN+DN* [23]	✓	✓	✓		6.31	40.0	36.0	34.0	–	–	–
CMA* [18]	✓	✓	✓		6.20	52.0	41.0	36.0	8.76	26.5	22.1
Sim2Sim* [24]	✓	✓	✓		6.07	52.0	43.0	36.0	8.76	26.5	22.1
GridMM* [45]	✓	✓	✓		5.11	61.0	49.0	41.0	–	–	–
DreamWalker* [43]	✓	✓	✓		5.53	59.0	49.0	44.0	–	–	–
Reborn* [2]	✓	✓	✓		5.40	57.0	50.0	46.0	5.98	48.6	42.0
ETPNav* [1]	✓	✓	✓		4.71	65.0	57.0	49.0	5.64	54.7	44.8
HNR* [44]	✓	✓	✓		4.42	67.0	61.0	51.0	5.50	56.3	46.7
AG-CMTP [8]	✓	✓	✓		7.90	39.0	23.0	19.0	–	–	–
R2R-CMTP [8]	✓	✓	✓		7.90	38.0	26.0	22.0	–	–	–
Instruc-Nav [38]	✓	✓	✓		6.89	–	31.0	24.0	–	–	–
LAW [42]		✓	✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0
CM2 [16]		✓	✓	✓	7.02	41.0	34.0	27.0	–	–	–
WS-MGMap [9]		✓	✓	✓	6.28	47.0	38.0	34.0	–	–	–
AO-Planner [7]	✓		✓		5.55	59.0	47.0	33.0	–	–	–
Seq2Seq [25]			✓	✓	7.77	37.0	25.0	22.0	12.10	13.9	11.9
CMA [25]			✓	✓	7.37	40.0	32.0	30.0	–	–	–
NaVID [54]				✓	5.47	49.0	37.0	35.0	–	–	–
Uni-NaVID [53]				✓	5.58	53.5	47.0	42.7	6.24	48.7	40.9
NaVILA [12]				✓	5.22	62.5	54.0	49.0	6.77	49.3	44.0
NavFoM [52]				✓	5.01	64.9	56.2	51.2	5.51	57.4	49.4
StreamVLN [47]				✓	4.98	64.2	56.9	51.9	6.22	52.9	46.0
DualVLN [46]				✓	**4.05**	**70.7**	64.3	58.5	4.58	61.4	51.8
Ours				✓	4.15	69.2	**65.6**	**62.1**	**4.42**	**61.7**	**53.3**

4.2 Experiments on Simulated Environments

Comparison with State-of-the-Arts. As summarized in Table 2, our Dual-Anchoring Framework establishes a new state-of-the-art on both benchmarks.

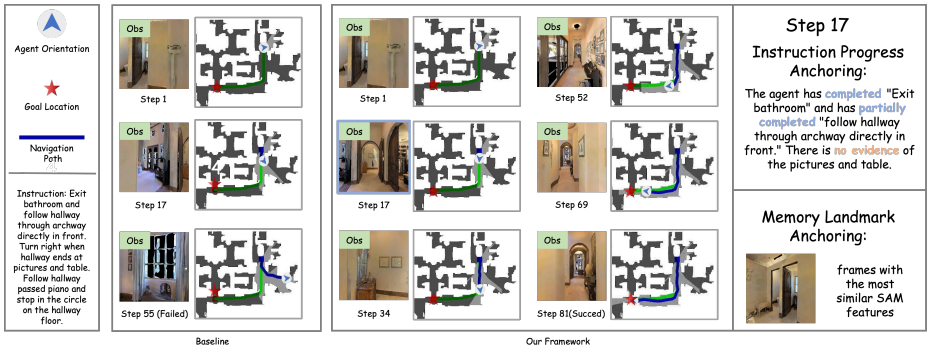

Fig. 4: Qualitative Visualization in Simulated Environment. We compare the navigation trajectories of the Baseline and our model on an R2R-CE unseen episode. While the baseline turns prematurely due to state drift, our model maintains the correct heading. The right panel showcases the internal states of our agent: the Instruction Progress Anchoring correctly identifies the current sub-goal as only "partially completed" , and the Memory Landmark Anchoring successfully reconstructs the visual features of the previously passed "bathroom" to ground the current spatial context.

On R2R-CE, our method significantly outperforms the strong StreamVLN [47] baseline, improving Success Rate (SR) from 56.9% to 65.6% and SPL from 51.9% to 62.1%. This demonstrates that while implicit hidden states in prior Video-LLMs tend to decouple from physical reality, our explicit regularization effectively synchronizes the agent's cognitive process with the environment. For the more challenging RxR-CE dataset, which features longer trajectories and complex instructions, the benefits of our approach are even more pronounced. While long horizons typically exacerbate State Drift, our framework maintains robustness, achieving an 8.8% absolute gain in SR (52.9% $\rightarrow$ 61.7%) over StreamVLN. These results confirm that anchoring internal states to both instruction progress and visual landmarks is crucial in long-horizon navigation.

Visualization. To demonstrate how our framework mitigates State Drift, we visualize a challenging R2R-CE episode in Fig. 4. The instruction requires the agent to "*exit the bathroom, go straight to the end of the hallway, and turn left.*" At Step 17, when encountering a deceptive opening before the hallway's true end, the baseline suffers from Progress Drift, which assumes the navigation has reached the turning point, leading to a premature turn left and ultimate failure. In contrast, our Dual-Anchoring agent maintains a synchronized internal state. Although the auxiliary heads are typically discarded during inference to minimize latency, we employ them here as probes to decode and visualize the agent's latent consistency. The Instruction Progress Anchoring acts as a semantic stabilizer, explicitly signaling that the "go straight" phase is still in progress, which effectively suppresses the distractor action. Simultaneously, the Memory Landmark Anchoring provides a "rear-view mirror" effect, retrospectively grounding the agent by retrieving the starting bathroom's features. This suggests that ex-

plicit state anchoring could prevent the decision-making process from decoupling from physical reality, ensuring robust long-horizon execution.

Table 3: Ablation study of key components on R2R-CE Val Unseen. We evaluate the Instruction Progress Anchoring (short for IPA) and Memory Landmark Anchoring (short for MLA) under two training data scales. To ensure fairness, the number of navigation samples remains identical across variants within the same setting.

Components		R2R+RxR+EnvDrop				All Datasets			
IPA	**MLA**	**NE ↓**	**OS ↑**	**SR ↑**	**SPL ↑**	**NE ↓**	**OS ↑**	**SR ↑**	**SPL ↑**
		6.49	46.9	40.8	37.4	4.98	64.2	56.9	51.9
✓		6.27	51.9	45.4	41.1	4.72	66.5	60.8	55.1
	✓	6.01	53.6	47.7	43.3	4.51	67.0	62.5	57.3
✓	✓	**5.73**	**55.8**	**49.5**	**44.9**	**4.15**	**69.2**	**65.6**	**62.1**

Ablation Study of Components. To investigate the effect of our proposal, we conduct ablation studies on the R2R-CE validation unseen split under two training settings: 1) Standard Mixture (w/o dagger and VL data) to isolate methodological gains, and 2) All Datasets to validate robustness at scale. In Standard Mixture setting, we make the total number of training samples strictly identical. For example, if Instruction Progress Anchoring is added, we replace half of the base nav data in the baseline with our augmented Progress Description Data, adding no additional samples. In All Datasets setting, we don't control the number of training samples. Therefore, if Instruction Progress Anchoring is added, we add all Progress Description Data to the original training dataset.

As shown in Table 3, the baseline yields the worst results due to State Drift. Incorporating the Instruction Progress Anchoring significantly improves Success Rate (e.g., from 40.8% to 45.4%), confirming that explicit sub-goal prediction maintains semantic alignment. Meanwhile, the Memory Landmark Anchoring notably boosts SPL and reduces Navigation Error ($6.49 \rightarrow 6.01$), validating that retrospective landmark reconstruction effectively prevents trajectory drift. Finally, the Dual-Anchoring configuration achieves the best performance across both settings, demonstrating that our explicit state regularizations offer fundamental, complementary benefits that persist even with scaled-up data.

Analysis across Trajectory Lengths. We further analyze the performance across different trajectory lengths. The R2R-CE validation unseen set is divided into three subsets based on the trajectory geodesic distance: **Short** ($\in [3.85, 7.55)$ meters), **Medium** ($\in [7.55, 9.81)$ meters), and **Long** ($\in [9.81, 21.04]$ meters). As illustrated in Fig. 5, the baseline exhibits sharp degradation as navigation distance increases due to unmitigated drift. In contrast, our Dual-Anchoring agent

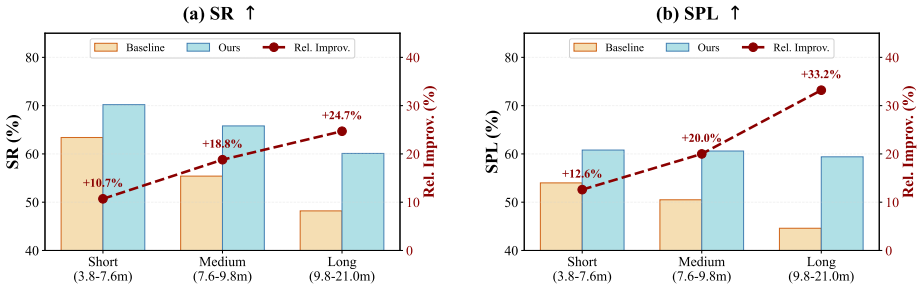

Fig. 5: Performance across different trajectory lengths. We report (a) SR and (b) SPL across Short, Medium, and Long episodes on R2R-CE validation unseen set. As trajectories become longer, our method provides increasingly substantial gains.

maintains superior stability, with the performance gap widening significantly on longer trajectories. Especially for the relative improvement in SPL, which escalates from +12.6% in Short episodes to a remarkable +**33.2%** in Long episodes. This trend confirms that explicit state anchoring—via both Instruction Progress and Memory Landmark modules—effectively counteracts the decoupling of internal states from physical reality, a benefit that becomes increasingly critical n long-horizon navigation scenarios.

Table 4: Quantitative Assessment of Selected Datasets. We evaluate the quality of auto-generated data using reference-free metrics. The "Baseline" column denotes the control settings: raw image inputs (w/o Visual Prompting) for Progress Description Data, and Random Sampling for Grounded Landmark Data.

Datasets	Baseline	Metric	Baseline	Ours	Δ
Progress Description	w/o Visual Prompting	Hallucination Rate (HR) (%) ↓	8.13	**6.04**	**-25.7%**
		Logical Consistency Score (LCS) (1-5) ↑	1.71	**4.26**	**+149%**
Grounded Landmark	Random Sampling	Landmark Presence Rate (LPR) (%) ↑	13.9	**75.6**	**+444%**

4.3 Quality Assessment of Self-Collected Dataset.

Since our framework relies on self-collected datasets, ensuring their reliability is paramount. We establish a rigorous evaluation protocol to quantify the quality of Progress Description Data and Grounded Landmark Data.

Progress Description Data. We employ an LLM-as-a-Judge paradigm to evaluate generated progress descriptions using two key metrics: 1) *Hallucination Rate (HR):* The percentage of descriptions containing entities not present in the original instruction. 2) *Logical Consistency Score (LCS):* A 1-5 Likert scale measuring whether the generated progress represents a logically contiguous prefix of the instruction without skipping steps. We calculate these under two distinct settings: feeding raw image inputs VS employing our *Visual Kinematics Prompting* (Sec. 3.2). As shown in Table 4, the setting without visual prompting struggles

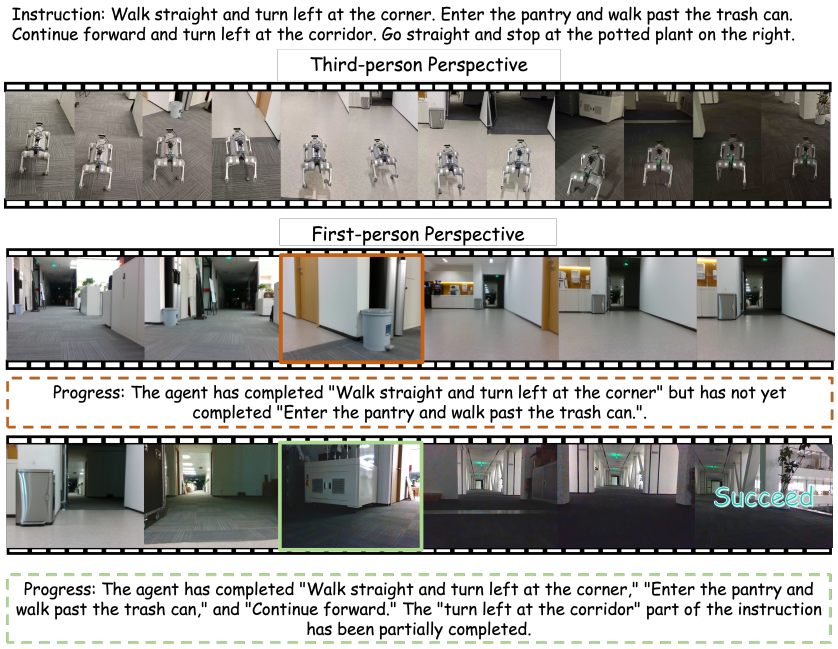

Fig. 6: Qualitative Visualization of Real-World Deployment. Top: Third-person perspective of the robot's trajectory. Bottom: First-person RGB stream. The **dashed boxes** display the real-time progress descriptions explicitly generated by our model, corresponding to the frames bordered in the same color.

to track the agent's dynamic movement, resulting in a low LCS of 1.71. In contrast, our method achieves a high LCS of **4.26** and a significantly reduced HR from 8.13% to 6.04%. This improvement confirms that overlaying explicit action history is critical for the annotator to align visual changes with instruction steps, thereby yielding high-quality, coherent, and faithful training data.

Grounded Landmark Data. To evaluate the temporal precision of the grounded landmark, we propose the Landmark Presence Rate (LPR), which extracts the landmark entity from the sub-instruction and prompts Qwen3-VL to verify whether the object is visually recognizable. We compare our mined frames against a Random Sampling baseline. As reported in Table 4, random frames contain the target landmark only 13.9% of the time. Conversely, our method achieves an LPR of 75.6%. This substantial gap confirms that our pipeline effectively localizes the semantic "high-light" moments of landmarks, thus providing valid retrospective signals.

4.4 Real-World Deployment.

Experimental Settings. Our real-world experiments are conducted on a Unitree Go2 quadruped robot equipped with a front-facing Intel RealSense D435i

RGB-D camera. The navigation system operates in a client-server architecture: RGB observations are continuously streamed from the D435i camera to a remote server powered by NVIDIA H20 GPUs. Upon receiving a natural language instruction, our model processes the streaming observations to predict the next low-level navigational action in real-time. The predicted action is then transmitted back to the Go2 robot, which executes it by calling the robot motion API. Note that the model is trained entirely in the Matterport3D simulator and deployed directly to the real world without any fine-tuning.

Qualitative Results. As illustrated in Fig. 6, our agent successfully anchors its visual observations to linguistic sub-goals despite the significant domain gap. The model explicitly generates progress descriptions (shown in dashed boxes) that accurately reflect its current status. For instance, in the intermediate stage (orange box), the agent correctly identifies that it has completed the "turn left" action but still needs to "enter the pantry". Similarly, in the near-goal stage (green box), it precisely recognizes the completion of the full instruction. This capability to distinguish between completed and remaining segments effectively mitigates progress drift in long-horizon tasks.

5 Conclusion

In this paper, we address State Drift in Vision-Language Navigation by proposing the Dual-Anchoring Framework, which explicitly anchors the agent's internal state to instruction progress and visual landmarks. To support this, we introduce Instruction Progress Anchoring and Memory Landmark Anchoring, backed by automated pipelines that generate 4.5 million synthesized samples. Extensive experiments show our method achieves state-of-the-art performance on R2R-CE and RxR-CE, with remarkable robustness in long-horizon trajectories and successful real-world deployment.

6 Acknowledgements

This work was supported by the National Key R&D Program of China (2024YFB3309303) and supported by the Provincial Key R&D Program of Shaanxi (2025YXYC026).

References

1. An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. An, D., Wang, Z., Li, Y., Wang, Y., Hong, Y., Huang, Y., Wang, L., Shao, J.: 1st place solutions for rxr-habitat vision-and-language navigation competition (cvpr 2022). arXiv preprint arXiv:2206.11610 (2022)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: 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)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15791–15801 (2025)
7. Chen, J., Lin, B., Liu, X., Ma, L., Liang, X., Wong, K.Y.K.: Affordances-oriented planning using foundation models for continuous vision-language navigation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 23568–23576 (2025)
8. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11276–11286 (2021)
9. Chen, P., Ji, D., Lin, K., Zeng, R., Li, T., Tan, M., Gan, C.: Weakly-supervised multi-granularity map learning for vision-and-language navigation. Advances in Neural Information Processing Systems **35**, 38149–38161 (2022)
10. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. Advances in neural information processing systems **34**, 5834–5847 (2021)
11. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: 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)
12. Cheng, A.C., Ji, Y., Yang, Z., Gongye, Z., Zou, X., Kautz, J., Bıyık, E., Yin, H., Liu, S., Wang, X.: Navila: Legged robot vision-language-action model for navigation. arXiv preprint arXiv:2412.04453 (2024)
13. Contributors, I.: InternNav: InternRobotics' open platform for building generalized navigation foundation models. `https://github.com/InternRobotics/InternNav` (2025)
14. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied ai: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(2), 230–244 (2022)

15. Forlizzi, J., DiSalvo, C.: Service robots in the domestic environment: a study of the roomba vacuum in the home. In: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction. pp. 258–265 (2006)
16. Georgakis, G., Schmeckpeper, K., Wanchoo, K., Dan, S., Miltsakaki, E., Roth, D., Daniilidis, K.: Cross-modal map learning for vision and language navigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15460–15470 (2022)
17. Guo, X., Chai, W., Li, S.Y., Wang, G.: Llava-ultra: Large chinese language and vision assistant for ultrasound. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 8845–8854 (2024)
18. Hong, Y., Wang, Z., Wu, Q., Gould, S.: 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)
19. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: A recurrent vision-and-language bert for navigation. arXiv preprint arXiv:2011.13922 (2020)
20. Hu, J., Chen, J., Bai, H., Luo, M., Xie, S., Chen, Z., Liu, F., Chu, Z., Xue, X., Ren, B., et al.: Astranav-world: World model for foresight control and consistency. arXiv preprint arXiv:2512.21714 (2025)
21. Huang, T., Li, D., Yang, R., Zhang, Z., Yang, Z., Tang, H.: Mobilevla-r1: Reinforcing vision-language-action for mobile robots. arXiv preprint arXiv:2511.17889 (2025)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
23. Krantz, J., Gokaslan, A., Batra, D., Lee, S., Maksymets, O.: Waypoint models for instruction-guided navigation in continuous environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15162–15171 (2021)
24. Krantz, J., Lee, S.: Sim-2-sim transfer for vision-and-language navigation in continuous environments. In: European conference on computer vision. pp. 588–603. Springer (2022)
25. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
26. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. arXiv preprint arXiv:2010.07954 (2020)
27. Li, P., Wu, K., Fu, J., Zhou, S.: Regnav: Room expert guided image-goal navigation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4860–4868 (2025)
28. Li, P., Wu, K., Huang, W., Zhou, S., Wang, J.: Camera-aware label refinement for unsupervised person re-identification. arXiv preprint arXiv:2403.16450 (2024)
29. Li, P., Wu, K., Xu, S., Li, F., Li, H., Zhao, L., Lyu, K., Chen, L., Yang, Z.X., Zheng, N.: Spaact: Spatially-activated transition learning with curriculum adaptation for vision-language navigation. arXiv preprint arXiv:2604.27620 (2026)
30. Li, P., Wu, K., Xu, S., Li, F., Zhao, L., Chen, L., Yang, Z.X., Zheng, N.: Think before go: Hierarchical reasoning for image-goal navigation. arXiv preprint arXiv:2604.17407 (2026)

31. Li, P., Wu, K., Zhou, S., Huang, Q., Wang, J.: Pseudo labels refinement with intra-camera similarity for unsupervised person re-identification. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 366–370. IEEE (2023)
32. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
33. Lin, X., Li, J., Wu, K., Tang, J., He, Q., Zhao, L.: Morn: Metacognitive object-goal regulation for resource-rational long-horizon navigation. arXiv preprint arXiv:2605.16932 (2026)
34. Liu, F., Xie, S., Luo, M., Chu, Z., Hu, J., Wu, X., Xu, M.: Navforesee: A unified vision-language world model for hierarchical planning and dual-horizon navigation prediction (2025), https://arxiv.org/abs/2512.01550
35. Liu, Y., Fu, J., Wu, Y., Wu, K., Li, P., Wu, J., Zhou, S., Xin, J.: Mind the gap: Aligning vision foundation models to image feature matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20313–20323 (2025)
36. Liu, Y., Fu, P., Li, H., Qi, Y., Jiang, C., Fu, J., Liu, Z., Qin, B., Luo, Z., Luan, J., et al.: Elva: Exploring ranking-driven universal multimodal retrieval. arXiv preprint arXiv:2606.20280 (2026)
37. Liu, Y., Huang, Q., Hui, S., Fu, J., Zhou, S., Wu, K., Li, P., Wang, J.: Semantic-aware representation learning for homography estimation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2506–2514 (2024)
38. Long, Y., Cai, W., Wang, H., Zhan, G., Dong, H.: Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. arXiv preprint arXiv:2406.04882 (2024)
39. Lyu, K., Wu, K., Li, P., Hu, X., Si, Q., Miao, C., Yang, N., Wang, Z., Xiao, L., Hu, L., et al.: Himemvln: Enhancing reliability of open-source zero-shot vision-and-language navigation with hierarchical memory system. arXiv preprint arXiv:2603.14807 (2026)
40. Martinez-Martin, E., del Pobil, A.P.: Personal robot assistants for elderly care: an overview. Personal assistants: Emerging computational technologies pp. 77–91 (2017)
41. Mišeikis, J., Caroni, P., Duchamp, P., Gasser, A., Marko, R., Mišeikienė, N., Zwilling, F., De Castelbajac, C., Eicher, L., Früh, M., et al.: Lio-a personal robot assistant for human-robot interaction and care applications. IEEE Robotics and Automation Letters **5**(4), 5339–5346 (2020)
42. Raychaudhuri, S., Wani, S., Patel, S., Jain, U., Chang, A.: 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 (2021)
43. Wang, H., Liang, W., Van Gool, L., Wang, W.: Dreamwalker: Mental planning for continuous vision-language navigation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10873–10883 (2023)
44. Wang, Z., Li, X., Yang, J., Liu, Y., Hu, J., Jiang, M., Jiang, S.: Lookahead exploration with neural radiance representation for continuous vision-language navigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2024)
45. Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Gridmm: Grid memory map for vision-and-language navigation. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 15625–15636 (2023)

46. Wei, M., Wan, C., Peng, J., Yu, X., Yang, Y., Feng, D., Cai, W., Zhu, C., Wang, T., Pang, J., et al.: Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. arXiv preprint arXiv:2512.08186 (2025)
47. Wei, M., Wan, C., Yu, X., Wang, T., Yang, Y., Mao, X., Zhu, C., Cai, W., Wang, H., Chen, Y., et al.: Streamvln: Streaming vision-and-language navigation via slowfast context modeling. arXiv preprint arXiv:2507.05240 (2025)
48. Wisspeintner, T., Van Der Zant, T., Iocchi, L., Schiffer, S.: Robocup@ home: Scientific competition and benchmarking for domestic service robots. Interaction Studies **10**(3), 392–426 (2009)
49. Wu, K., Li, P., Fu, J., Li, Y., Wu, Y., Liu, Y., Wang, J., Zhou, S.: Event-equalized dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8417–8427 (2025)
50. Wu, K., Li, P., Fu, J., Wu, Y., Liu, Y., Zhou, S., Wang, J.: Cem-net: Cross-emotion memory network for emotional talking face generation. arXiv preprint arXiv:2508.12368 (2025)
51. Yao, X., Gao, J., Xu, C.: Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. arXiv preprint arXiv:2506.23468 (2025)
52. Zhang, J., Li, A., Qi, Y., Li, M., Liu, J., Wang, S., Liu, H., Zhou, G., Wu, Y., Li, X., et al.: Embodied navigation foundation model. arXiv preprint arXiv:2509.12129 (2025)
53. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint arXiv:2412.06224 (2024)
54. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. arXiv preprint arXiv:2402.15852 (2024)
55. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
56. Zhang, Z., Zhang, H., Chen, X., Huang, K., Shi, C., Lu, H.: Latent-space autoregressive world model for efficient and robust image-goal navigation. arXiv e-prints pp. arXiv–2511 (2025)
57. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7641–7649 (2024)