2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

VISTA: Generative Visual Imagination for Vision-and-Language Navigation

Yanjia Huang Mingyang Wu Renjie Li Zhengzhong Tu
Department of Computer Science and Engineering, Texas A&M University

***Abstract*— Vision-and-Language Navigation (VLN) tasks agents with locating specific objects in unseen environments using natural language instructions and visual cues. Many existing VLN approaches typically follow an 'observe-and-reason' schema, that is, agents observe the environment and decide on the next action to take based on the visual observations of their surroundings. They often face challenges in long-horizon scenarios due to limitations in immediate observation and vision-language modality gaps. To overcome this, we present VISTA, a novel framework that employs an 'imagine-and-align navigation strategy. Specifically, we leverage the generative prior of pre-trained diffusion models for dynamic visual imagination conditioned on both local observations and high-level language instructions. A Perceptual Alignment Filter module then grounds these goal imaginations against current observations, guiding an interpretable and structured reasoning process for action selection. Experiments show that VISTA sets new state-of-the-art results on Room-to-Room (R2R) and RoboTHOR benchmarks, *e.g.*, +3.6% increase in Success Rate on R2R. Extensive ablation analysis underscores the value of integrating forward-looking imagination, perceptual alignment, and structured reasoning for robust navigation in long-horizon environments.**

Key Words: Vision-and-Language Navigation, Diffusion Models, Vision Language Models

I. INTRODUCTION

"Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world." — Albert Einstein

Humans rarely navigate by sight alone. When sitting in sofa instructed by your friend to "go to kitchen and grab a soda", we don't only scan our surroundings; instead we imagine what the kitchen looks like, anticipate its layout, and mentally simulate possible paths, even we are in a friend's house. This cognitive loop between imagination, perception, and reasoning enables us to act with foresight, adapt to ambiguity, and recover from uncertainty. Despite rapid progress in Vision-and-Language Navigation (VLN), most existing agents operate without such a mechanism [1]–[3]. They treat navigation as a sequence of local predictions, reacting to immediate observations and static instructions without explicitly modeling what lies beyond their current view [4]–[7].

In this paper, we argue that explicit goal imagination is crucial for embodied navigation, enabling them to anticipate future states and make more informed decisions. Integrating visual imagination into navigation presents significant challenges [8]–[16]. One key difficulty is to determine when to rely on high-level semantic instructions, especially under trajectory uncertainty. Moreover, aligning imagined goals with current observations requires robust and interpretable.

To address these challenges, we propose **VISTA**—a cognitively inspired VLN system equipping the agent to imagine, verify, and act through a structured, interpretable framework. Specifically, we introduce the Adaptive Imagination Scheduler (AIS), a mechanism that dynamically decides between instruction-driven and observation-driven imagination modes based on trajectory confidence and visual-semantic alignment. Crucially, rather than treating these generated visuals as uninterpretable priors [17]–[19], we apply a Perceptual Alignment Filter (PAF) module that aligns imagined goals with current observations via cross view attention. This yields interpretable visual attention maps that highlight goal-relevant regions — effectively letting the agent see where it should look before deciding where to go. Finally, all information is passed to a navigational reasoning module that performs Chain-of-Thought (CoT) [20] action prediction, closing the loop with language, vision, and decision-making. The AIS and PAF components together form a robust solution, effectively enabling forward looking imagination to improve navigation decisions.

Our contributions are summarized as follows:

- We present a unified closed-loop VLN framework—VISTA, integrating proactive imagination, visual-semantic alignment, and structured Chain-of-Thought reasoning, resulting in robust and interpretable navigation behaviors.
- We introduce two modules: Adaptive Imagination Scheduler (AIS), dynamically selecting between instruction-driven and observation-driven imagination modes based on trajectory uncertainty and visual-semantic alignment; and Perceptual Alignment Filter (PAF) module, explicitly grounding visual imagination with real-time observations to enhance interpretability and navigation accuracy.
- We validate our approach on standard VLN benchmarks (R2R [1], RoboTHOR [21]), demonstrating significant gains in success rate, navigation efficiency, and interpretability.

We clarify that "forward-looking imagination" in VISTA does not mean photo-realistic prediction of exact future frames. Instead, imagined scenes serve as exploratory visual priors that help the agent reason about plausible but currently unseen regions. These priors are explicitly grounded by the Perceptual Alignment Filter (PAF), which rejects inconsistent content and preserves robustness.

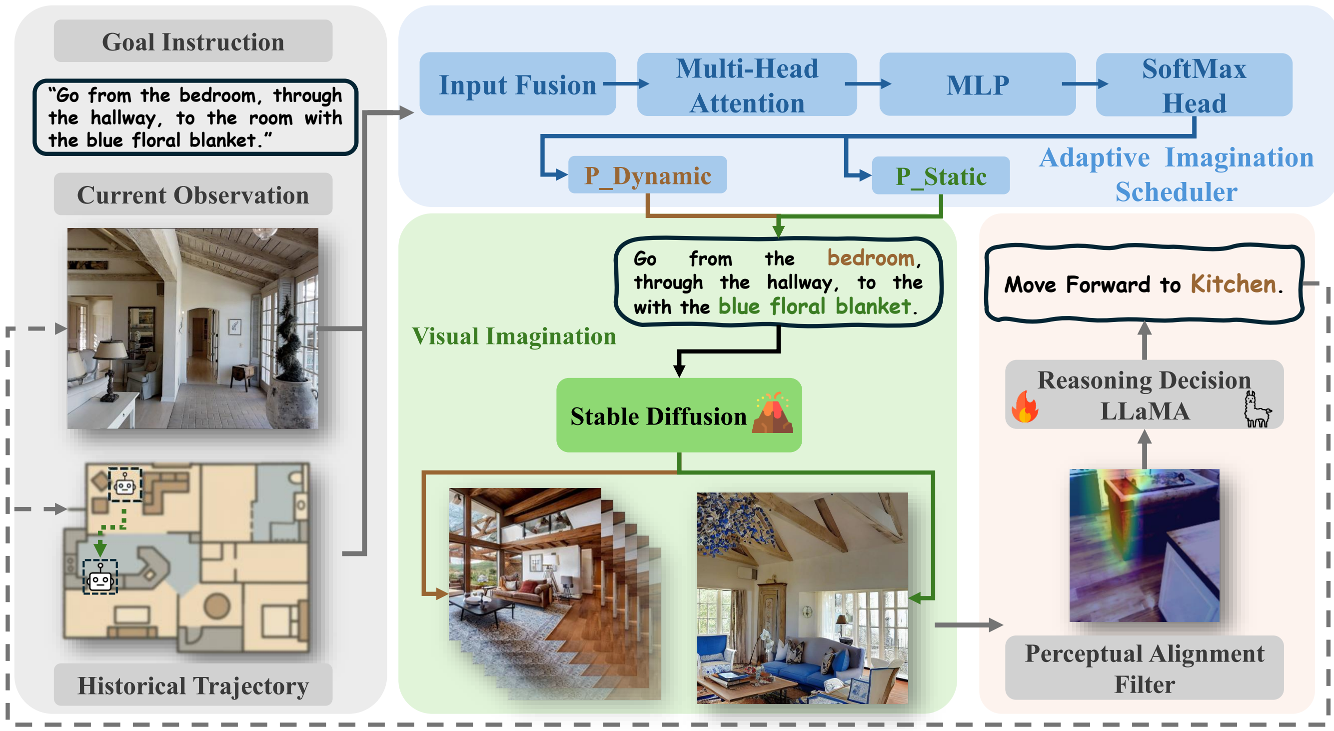


Fig. 1: Overview of **VISTA**—At each step, the agent predicts what it expects to see, imagines that goal visually, aligns the prediction with reality, and reasons through language to decide how to move. This loop is guided by an adaptive scheduler that balances global task intent with local observations, enabling dynamic goal prediction. By grounding imagination in perception and embedding structured reasoning into decision-making, VISTA offers a forward-looking and interpretable approach to vision-and-language navigation. Colors indicate scheduler modes (brown: instruction driven; green: observation drive). Images are all offline priors; PAF filters them against current observations.

II. Related Work

A. Vision-and-Language Navigation (VLN)

Vision-and-Language Navigation (VLN) aims to guide agents through unfamiliar environments using natural language instructions. This task has been enabled by large-scale simulators (e.g., Matterport3D [22]) and benchmark datasets such as R2R [1] and REVERIE [23], which focus on low-level navigation and high-level object interaction respectively. To further enhance generalization to unseen environments, recent approaches leverage large-scale pre-training with vision-language objectives [6], [24]–[26].

Recent works have explored incorporating large language models (LLMs) as navigation backbones [27]–[29] to enhance VLN agents. For instance, NavGPT [10] treats navigation as a language-driven process, converting observations and history into prompts for LLM interpreting and reasoning. DiscussNav leverages dialogue-style interactions with LLM experts to enable decision-making, while MapGPT [28] introduces an online-updated, language-formalized map to support long-range planning and adaptive behaviors. NaviLLM [30] further explores fine-tuning LLMs for embodied tasks. Despite these advances, current VLN agents still struggle to adapt to realistic, commonsense rich environments and often lack interpretability and structured reasoning in decision-making. Unlike these methods, we integrates visual imagination and perceptual alignment into structured reasoning enabling more interpretable and goal driven navigation decisions.

B. Visual Imagination in VLN

Recent research has explored visual imagination through generative models to enhance VLN agents by predicting future observations. Pathdreamer [31] uses GANs to hallucinate panoramic views from visual history, but its outputs are restricted to heuristic-based search and lack policy integration. PanoGen [32] and DiffPano [33] use diffusion models to extrapolate goal views. PanoGen++ [34] further adapts diffusion models for indoor navigation via LoRA fine-tuning and structured inpainting. However, these approaches typically rely on offline generation and treat imagined content as auxiliary information rather than a core component of decision making. Currently, look-ahead exploration with neural radiance fields and prompt-driven imagination for VLN have been studied. Unlike these works, VISTA integrates imagination *inside* the decision loop through an explicit visual bottleneck (PAF) and a scheduler (AIS), providing grounded priors rather than standalone previews.

C. Chain of Thoughts Reasoning

Building upon the integration of LLMs into VLN agents, recent research has further emphasized the need for advanced reasoning strategies to strengthen decision-making. Among them, Chain-of-Thought (CoT) prompting [35] introduces an in-context learning approach that decompose complex reasoning tasks into a sequence of simpler, interpretable sub-steps. Extending the linear reasoning paradigm, Tree-of-Thoughts (ToT) [36] adopts a branching structure that enables the parallel exploration of multiple reasoning paths. Self-consistency [37] further augments CoT by sampling diverse reasoning chains and aggregating the outputs through majority voting. Inspired by reasoning strategies, C-Instructor [38] employs CoT prompting to guide LLMs in identifying key landmarks and generating navigation instructions. InstructNav [39] introduces a Dynamic Chain-of-Navigation, aligned with CoT reasoning, to unify diverse instruction types and convert linguistic planning into executable navigation trajectories. NavCoT [40] further introduces a trainable CoT framework that incorporates world model reasoning and formalized supervision. Our method, VISTA, rather than using LLMs as monolithic planners, we embeds visual imagination and alignment within the CoT framework, creating visually grounded reasoning steps.

III. METHOD

We propose a cognitively inspired navigation framework operating in a closed-loop reasoning cycle, where the agent actively imagines future goals, aligns these predictions with real-time visual input, and reasons over actions. Specifically, at each navigation step, an Adaptive Imagination Scheduler selects modes based on two metrics. The chosen mode generates a visual hypothesis of the next goal through Visual Imagination Module , which is then spatially aligned with current observations via Perceptual Alignment Filter. Finally, the agent integrates this aligned visual information with instructions and navigation history through Navigational Chain-of-Thought Reasoning to select next action. This structured imagination to alignment to reasoning loop enables the agent to proactively anticipate goals, ground predictions visually, and make interpretable navigation decisions, shown in Figure 1.

a) Notation.: At step t, the agent observes O_t and follows instruction I. The imagination module produces $\hat{I}_t^{\text{SD}}$ (text-conditioned) and $\hat{I}_t^{\text{inpaint}}$ (observation-inpainted). AIS relies on two scalars: (i) trajectory uncertainty $u_t = \alpha\,\mathscr{H}(\pi(\cdot \mid \mathscr{H}_t)) + (1-\alpha)\,\mathrm{dev}(\mathscr{P}_t)$, where $\mathscr{H}$ is the entropy of the action distribution over the last K steps and dev is the path deviation w.r.t. the shortest path; (ii) visual semantic similarity $s_t = \cos\big(\phi(\hat{I}_t^{\text{SD}}), \phi(O_t)\big)$ with a frozen image encoder $\phi(\cdot)$. AIS selects Dynamic if $u_t > \tau_u$ and $s_t < \tau_s$, otherwise Static. The thresholds τ_u, τ_s are picked on the validation set and we found performance stable within small perturbations.

A. Adaptive Imagination Scheduler

A key challenge in goal anticipation is deciding when an agent should rely on high-level instructions and when to adapt based on immediate visual observations. Instructions provide semantic guidance but can lack precision, especially in unfamiliar settings, where visual observations are detailed but potentially "noisy" without context. To balance these two sources effectively, we propose an **Adaptive Imagination Scheduler (AIS)**. It dynamically switches between **static** (instruction driven) and **dynamic** (observation driven) goal predictions.

Specifically, in static mode, AIS extracts semantic entities (e.g., "kitchen", "stairs", "sofa") from instructions with language models. In dynamic mode, it leverages last trajectory history and current visual observations to propose immediate goals. AIS determines the prediction mode by evaluating two metrics: *trajectory uncertainty* u_t, measured by action entropy and path deviation. And *visual semantic similarity* score s_t which computed as the cosine similarity between embeddings of imagined and observed images. Formally, we define the mode selection as:

$$z_t = \begin{cases} \text{static}, & \text{if } u_t < \tau_u \text{ and } s_t > \tau_s \\ \text{dynamic}, & \text{otherwise} \end{cases} \tag{III.1}$$

where τ_u and τ_s are predefined thresholds. AIS thus enables the agent to flexibly integrate high-level intent with real-time perceptual insights, facilitating robust and context-aware navigation decisions.

B. Visual Imagination Module

To visually anticipate future goals, our agent constructs scene hypotheses using a two-stage vision-language pipeline. First, a candidate scene—either from training data or previously imagined—is passed through Qwen2-VL [41] to generate a descriptive caption (e.g., *"A spacious living room with a blue sofa, wooden floors, and large windows."*). This caption is then fed into a LoRA-finetuned Stable Diffusion model [42], [43] to synthesize a corresponding visual scene. To enhance contextual relevance, we additionally create an inpainted version I_t^{inpaint} by filling masked regions of the current observation with imagined content guided by the same caption. This approach produces semantically rich, visually coherent hypotheses that serve as proactive priors for subsequent alignment and decision-making steps. VISTA does not aim to photo-realistically predict the exact next camera view. The imagined scenes act as exploratory priors to expose plausible goal hypotheses beyond the current field of view. These priors are filtered by PAF against O_t so that irrelevant generations are down-weighted before decision making.

C. Perceptual Alignment Filter

Given its imagined visual hypothesis, the agent must verify whether its prediction aligns with reality—and if so, where to look. To this end, we introduce a perceptual alignment filter that compares the agent's imagined goal scenes with its current egocentric observation, producing an attention map

Prompt

Input: Instruction: "Walk with the family photo on your right. Continue straight into the bathroom. Walk past the bathtub. Stop in the hall between the bathroom and toilet doorways."

Observations:

A. Go back to <a hallway with a window and a small plant>
B. Go forward to <a bathroom with a bathtub and gray tiled walls>
C. Turn left into <a living room with a blue couch and a wall-mounted photo>
D. Turn right into <a small hallway between two white doors>

GPT-4o: (History: Step 2. Turned left into a hallway with a photo on the right.)
Thought: The instruction asks to walk with the family photo on the right, continue into the bathroom, pass the bathtub, and stop between bathroom and toilet doors.
From the history, I've already walked with the photo on the right and am now in the hallway. To continue, I should enter the bathroom and find the bathtub. Among the options, C mentions a bathroom and a bathtub, so it seems to be the right choice.
Action: C ❌

LLaMA2: (History: Step 2. Turned left into hallway, photo observed)
Thought: The instruction has been partially fulfilled (photo seen).
Next, I should go into the bathroom and stop after passing the bathtub. Option C mentions both bathroom and bathtub, so I will choose that.
Action: C ❌

VISTA(ours): (History: Step 2. Walked past photo; entered hallway; next target = bathroom → then hallway between two doors)
Imagination: a narrow hallway with two white doors (bathroom and toilet) on either side.
Filtered Observation:
- C matches bathroom but does not lead to hallway.
- D explicitly describes a hallway between two white doors → matches imagined goal.
- F is toilet room only, not the in-between space.
Action: D ✅

Fig. 2: Chain-of-Thought (CoT) reasoning comparison on a navigation step. Given the same instruction, observation candidates, and history, we compare zero-shot GPT-4, LLaMA2 baseline, and our VISTA agent. While other methods rely on textual cues only, our agent grounds its reasoning in visual imagination and attention alignment. VISTA successfully identifies the correct observation (D) by matching imagined scene with actual perception. We show one consistent pipeline trace per row (instruction, imagination prompt, PAF attention, CoT summary, action).

that highlights likely goal regions. The filter operates over three visual inputs: the current observation O_t, the imagined scene I_t^{SD}, and the inpainted image I_t^{inpaint}. Each image is encoded through a shared ResNet-18 backbone, and their features are fused via a multi-head attention mechanism. The fused representation is decoded into a spatial attention map $A_t \in [0,1]^{H\times W}$ that assigns soft relevance scores to each pixel in O_t:

$$A_t = \sigma\left(f_\theta\left(\phi(O_t), \phi(I_t^{\text{SD}}), \phi(I_t^{\text{inpaint}})\right)\right) \tag{III.2}$$

To supervise the attention map, we use pseudo ground-truth masks A_t^{GT} derived from grounding models. Specifically, we extract entity-level keywords from captions generated by Qwen, and localize them in O_t using Grounded-SAM [44]. These masks indicate where the imagined object should appear in the real observation.

We construct a dataset of 30,000 quadruples—$(O_t, I_t^{\text{SD}}, I_t^{\text{inpaint}}, A_t^{\text{GT}})$—spanning diverse scene types and object categories. The filter is trained using a combination of binary cross-entropy and soft Dice loss, with Adam optimizer and a learning rate of 1×10^{-4}. All images are resized to 256×256, and the model is trained for 50 epochs with early stopping. This alignment mechanism enables the agent to reconcile imagination with perception transforming generative predictions into spatially grounded attention. Crucially, the resulting mask A_t serves as a visual bottleneck: it filters irrelevant information, enforces goal specific focus, and enhances the interpretability of downstream decisions.

PAF is trained with weak supervision from caption-driven detections to encourage cross-modal grounding. This does not reduce PAF to a trivial copy of O_t: the module explicitly learns to align imagined content with observations and suppress mismatches. Our ablations comparing variants with or without PAF support that PAF contributes beyond textual descriptions.

D. Navigational Chain-of-Thought Decision

After aligning its imagined goal with current observations, the agent must determine the next navigation step. Unlike traditional flat action predictions, we adopt a structured, language-grounded Chain-of-Thought (CoT) reasoning [35], [45] approach. At each timestep, our policy model explicitly decomposes decision-making into an interpretable, step-wise reasoning process that integrates multiple modalities—including task instructions, observation captions, visual attention maps, and navigation history—into structured prompts. Our CoT reasoning consists of three stages: ❶ Goal grounding: Infer what the agent is currently trying to find (e.g., "a room with a blue sofa") based on the global instruction. ❷ Perceptual verification: Interpret the observation and attention map to determine whether this goal appears in the current scene. ❸ Decision justification: Based on alignment, select the best next action and justify it.

To generate interpretable reasoning traces, we fine-tune a LLaMA2-7B [46] model with LoRA adapters to output both intermediate reasoning steps and final actions. We represent visual context succinctly via textual summaries derived from attention maps (e.g., ”the mask highlights the left side of the hallway”). Crucially, unlike previous CoT approaches that primarily served post-hoc explanations, our CoT reasoning is directly embedded into the navigation control loop, bridging abstract instructions with concrete visual perception. This explicit, step-wise reasoning enhances interpretability, improves navigation robustness, and supports more transparent decision-making processes. A complete CoT reasoning example is illustrated in Figure 2.

E. Runtime & Complexity

The reported $\sim$4.2s per SD-generated image refers to *offline precomputation* used to build a pool of imagination candidates. Online navigation *does not* regenerate at every step: imagination is triggered sparsely by AIS (every Δt steps by default) and cached. The control loop (observation $\rightarrow$ PAF $\rightarrow$ CoT) runs at $\sim$8 FPS on our hardware. Hence there is no contradiction: the heavy image synthesis cost is amortized and does not bound the online step rate. We release prompts and settings to facilitate replication.

IV. Experiments

A. Experimental Setup

We evaluate our approach on the Room-to-Room (R2R) dataset [1], built within the Matterport3D simulator [22]. The environment consists of 90 real-world indoor scenes, with 10,567 panoramic viewpoints connected as a navigation graph. Each panorama is rendered as 36 discretized single-view images, providing a diverse egocentric perspectives at each location. The dataset provides with 7,189 trajectories, each annotated with three natural language instructions and a ground-truth path. It is split into: train (4675 trajectories), val-seen (340 trajectories), val-unseen (783 trajectories), and test (1391 trajectories) [49]. To support imagination-guided navigation, we curate a large-scale auxiliary dataset—R2R-Imagine—which provides visual predictions conditioned on both instructions and predicted goal concepts. In total, we generate over **250,000** images (256*256), covering imagined goal scenes across all R2R splits. These images are produced via a two-stage vision-language process (see Section III-B) Image generation takes approximately 4.2 seconds per image on a single NVIDIA A100 GPU. The full R2R-Imagine dataset—including all instruction-aligned, caption-conditioned imaginations—will be released publicly upon acceptance to encourage further research on generative visual priors for navigation. Visual goal captions are generated using Qwen2-VL, and converted into imagined goal images via LoRA-finetuned Stable Diffusion (v1.5). LoRA is trained

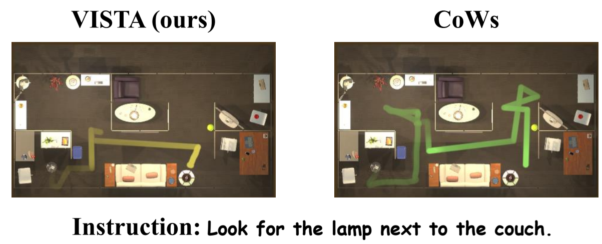


Fig. 3: **RoboTHOR instruction-following:** qualitative comparison between VISTA and CoWs.

using 5k paired goal-caption samples from the MP3D environment, with a learning rate of 5e-5 and 50 epochs. Training takes approximately 18 hours on 2 NVIDIA A100 GPUs with 80GB memory, using a batch size of 32 and sequence length of 30.

Evaluation Metrics We employ standard VLN metrics [50]: ❶ Success Rate (SR) measures the percentage of episodes where the agent stops within 3 meters of the goal; ❷ Success weighted by Path Length (SPL) penalizes longer trajectories by comparing the agent’s path to the shortest possible route; ❸ Navigation Error (NE) is the shortest-path distance between the agent’s final position and the goal location, measured in meters; ❹ Trajectory Length (TL) is the total distance traveled by the agent during an episode. Higher SR and SPL indicate better performance, while lower NE and TL reflect more accurate and efficient navigation. During inference, the control loop (observation $\rightarrow$ PAF $\rightarrow$ CoT) runs at 8 FPS on an A100. Imagination is precomputed offline and, when needed, sparsely refreshed under AIS with caching; it does not bound the online step rate.

B. Main Results

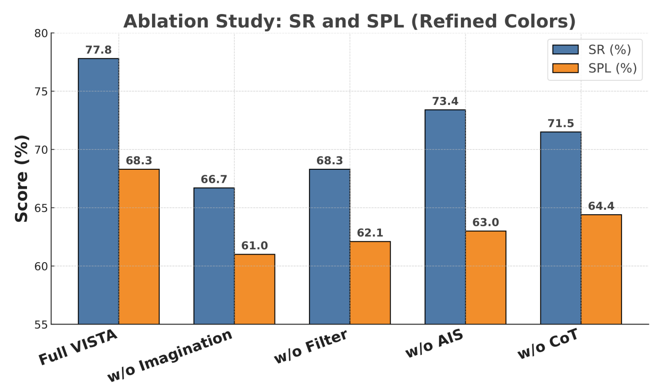
Table I presents a comprehensive comparison on the Room-to-Room benchmark. VISTA consistently achieves state-of-the-art performance across both Validation and Test Unseen splits, outperforming prior models in SR, SPL, and NE. On Validation Unseen, VISTA attains 77.8% SR and 68.3 SPL, surpassing the previous best (DUET-Imagine) by **+3.2% SR** and **+7.5 SPL**. On Test Unseen, VISTA maintains a strong 74.9% SR, indicating that its improvements generalize robustly to held-out environments. These improvements stem from the structured integration of AIS, PAF, and CoT, enabling anticipatory and interpretable decision making. As shown in Figure 4b, VISTA maintains higher success rates, more efficient paths, and more stable confidence across increasing path lengths. This indicates that our model scales robustly to long-horizon tasks by combining global imagination, local grounding, and structured reasoning. To further validate the generalization capability, we additionally evaluate VISTA on RoboTHOR [5], [21] Table II. Compared to competitive baselines such as VLTNet and ESC, VISTA achieves significant improvements, reaching 28.8% SWPL and 43.1% SR. Qualitative comparisons Figure 3 demonstrate that our approach effectively grounds instructions into visual

TABLE I: Performance on R2R (higher is better for SR/SPL; lower is better for TL/NE).

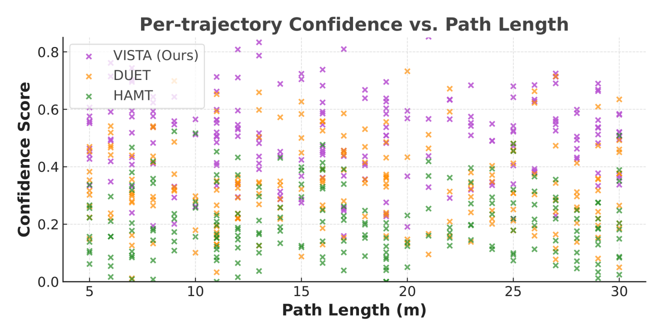
Models	Validation Unseen				Test Unseen			
	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑
EnvDrop [24]	10.70	5.22	52.0	48.0	**11.66**	5.23	51.0	47.0
PREVALENT [47]	**10.19**	4.71	58.0	53.0	**10.51**	5.30	54.0	51.0
RecBERT [48]	12.01	3.93	63.0	57.0	12.35	4.09	63.0	57.0
HAMT [19]	11.46	3.62	66.2	61.5	12.27	3.93	65.0	60.0
HAMT-Imagine [49]	11.80	**3.58**	**67.26**	**62.02**	12.66	**3.89**	**65.0**	**60.0**
DUET [18]	13.94	3.31	71.52	60.41	14.73	3.65	69.0	59.0
DUET-Imagine [49]	14.35	**3.19**	**72.12**	**60.48**	15.35	**3.52**	**71.0**	**60.0**
PanoGen [32]	13.40	3.03	74.2	64.3	14.38	**3.31**	**71.7**	**61.9**
NavCoT [45]	**9.95**	6.26	40.23	36.64	—	—	—	—
VISTA (ours)	**13.26**	**2.92**	**77.8**	**68.3**	**14.20**	**3.77**	**74.9**	**66.7**

TABLE II: **RoboTHOR** comparison (higher is better).

Model	SWPL↑	SR↑
CLIP-Ref. [5]	2.4	2.7
CLIP-Patch [5]	10.6	20.3
CLIP-Grad. [5]	9.7	15.2
MDETR [5]	8.4	9.3
OWL [5]	17.2	27.5
ESC [51]	22.2	38.1
VLTNet [6]	17.1	33.2
VISTA (ours)	**28.8**	**43.1**



(a) Removing imagination or alignment modules significantly drops navigation success.



(b) Trajectory confidence. VISTA maintains more stable confidence across long paths.

Fig. 4: **Ablation and planning analysis.** (a) Effect of removing key modules; (b) Robustness of VISTA's confidence vs. path length.

observations, highlighting its interpretability and robustness across different embodied navigation scenarios.

Despite significant performance improvements, VISTA occasionally encounters navigation failures. Typical failure scenarios include: 1. Inaccurate Visual Imagination: When the visual imagination module generates scenes significantly diverging from the actual environment, the subsequent perceptual alignment misleads action selection, causing navigation errors (e.g., selecting incorrect rooms or hallways). 2. Attention Misalignment: The perceptual alignment filter sometimes produces overly broad or misplaced attention masks, particularly in visually cluttered or ambiguous scenes, reducing localization accuracy and causing inefficient navigation paths. 3. Over reliance on Instruction-driven Mode: Occasionally, AIS prematurely selects static imagination mode in highly ambiguous environments, neglecting real-time visual cues and resulting in suboptimal paths or failed navigation.

V. Qualitative Traces

Success. Instr: "Go to the kitchen table and stop by the window." AIS=Static. Prompted imagination suggests a table near a bright window; PAF highlights the central rectangular surface in O_t; CoT infers "turn left, go forward two steps, align with window". **Failure.** Instr: "Find the small desk behind the couch." AIS=Dynamic. Imagination proposes a desk near shelves; PAF attention is diffuse due to occlusion; CoT selects a detour and over-rotates.

A. Ablation Studies

We observe that removing the visual imagination module leads to the most severe performance drop, highlighting its role in enabling forward-looking behavior. Without predicted goal images, the agent struggles to reason beyond its current field of view, resulting in both lower success and increased localization error. In contrast, removing the chain-of-thought (CoT) reasoning has a milder impact on success rate but noticeably reduces path efficiency. This suggests that while the agent can still reach the goal, it lacks the structured decision process to do so efficiently. Similarly, dropping the perceptual alignment filter degrades spatial focus, as the agent loses the ability to localize its imagined target within the scene—leading to more erratic trajectories. These trends confirm that each module in VISTA supports a distinct cognitive capacity, and removing any one of them disrupts the closed-loop interplay between imagination, grounding, and reasoning. To further understand module roles, we analyze their contributions via RoboTHOR, Visual imagination significantly improves SR (+11.1%) on long paths (>15m), confirming its necessity for foresight in extended navigation.

TABLE III: **Validation Unseen Ablations.** Removing imagination and alignment significantly reduces SR and SPL.

Variant	SR	SPL	NE	TL
Full	**77.8**	**68.3**	**2.92**	**13.26**
w/o Img	66.7	61.0	3.98	12.55
w/o Filter	68.3	62.1	3.42	14.30
w/o AIS	73.4	63.0	3.27	13.87
w/o CoT	71.5	64.4	3.18	13.76

VI. Conclusion

We present **VISTA**, a closed-loop vision-and-language navigation framework that integrates visual imagination, perceptual grounding, and structured reasoning into a unified decision process. Central to our method is an adaptive imagination scheduler that dynamically selects between static and dynamic goal prediction based on navigation uncertainty and visual-semantic alignment. Imagined scenes are precomputed offline (and sparsely refreshed under AIS when necessary), aligned with egocentric observations through a perceptual filter, and used to guide chain-of-thought reasoning in action selection. Our results on R2R and RoboTHORdemonstrate state-of-the-art performance, particularly in long-horizon and visually ambiguous scenarios. Through ablation and qualitative analysis, we show that each module—imagination, filtering, and reasoning—contributes distinctly to navigation accuracy, interpretability, and robustness.

Beyond VLN, our approach offers a general framework for visually grounded, imagination-driven planning. We believe that coupling generation and reasoning will be increasingly important for embodied agents operating under uncertainty, and we hope VISTA provides a step in that direction.

VII. Limitations

While VISTA shows strong performance in long-horizon navigation tasks, several limitations remain. First, the fidelity of imagined goal scenes depends on the generative quality of the diffusion model, which may fail in cluttered or diverse environments, leading to misalignments. Second, the imagination and reasoning pipeline introduces nontrivial computational overhead, which may limit real-time deployment. Third, the Adaptive Imagination Scheduler relies on manually tuned thresholds that may require adaptation across environments. Finally, all evaluations are conducted in simulation; real-world deployment may face additional challenges such as sensor noise and domain shifts. Future work may explore lightweight generation modules, learned scheduling strategies, and transfer methods to improve generalization and deployment readiness.

References

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," 2018. [Online]. Available: https://arxiv.org/abs/1711.07280

[2] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," 2019. [Online]. Available: https://arxiv.org/abs/1811.10092

[3] P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang, and C. Liu, "Voronav: Voronoi-based zero-shot object navigation with large language model," 2024. [Online]. Available: https://arxiv.org/abs/2401.02695

[4] W. Pengying, M. Yao, W. Bingxian, H. Yi, M. Ji, Z. Shanghang, and L. Chang, "Voronav: Voronoi-based zero-shot object navigation with large language model," *arXiv preprint arXiv:2401.02695*, 2024. [Online]. Available: https://www.arxiv.org/abs/2401.02695

[5] G. Samir, Yitzhak, W. Mitchell, I. Gabriel, S. Ludwig, and S. Shuran, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," *arXiv preprint arXiv:2203.10421*, 2022. [Online]. Available: https://www.arxiv.org/abs/2203.10421

[6] W. Congcong, H. Yisiyuan, H. Hao, H. Yanjia, Y. Shuaihang, H. Yu, L. Hui, L. Yu-Shen, and F. Yi, "Zero-shot object navigation with vision-language models reasoning," *arXiv preprint arXiv:2410.18570*, 2024. [Online]. Available: https://www.arxiv.org/abs/2410.18570

[7] D. Yinpei, P. Run, L. Sikai, and C. Joyce, "Think, act, and ask: Open-world interactive personalized robot navigation," *arXiv preprint arXiv:2310.07968*, 2023. [Online]. Available: https://www.arxiv.org/abs/2310.07968

[8] W. Xin, H. Qiuyuan, C. Asli, G. Jianfeng, S. Dinghan, W. Yuan-Fang, W. William, Yang, and Z. Lei, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," *arXiv preprint arXiv:1811.10092*, 2018. [Online]. Available: https://www.arxiv.org/abs/1811.10092

[9] C. Jinyu, G. Chen, M. Erli, Z. Qiong, and L. Si, "Reinforced structured state-evolution for vision-language navigation," *arXiv preprint arXiv:2204.09280*, 2022. [Online]. Available: https://www.arxiv.org/abs/2204.09280

[10] Z. Gengze, H. Yicong, and W. Qi, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," *arXiv preprint arXiv:2305.16986*, 2023. [Online]. Available: https://www.arxiv.org/abs/2305.16986

[11] D. Vishnu, Sashank, S. Gunnar, P. Robinson, T. Jesse, and S. Gaurav, S., "Clip-nav: Using clip for zero-shot vision-and-language navigation," *arXiv preprint arXiv:2211.16649*, 2022. [Online]. Available: https://www.arxiv.org/abs/2211.16649

[12] L. Yuxing, L. Xiaoqi, C. Wenzhe, and D. Hao, "Discuss before moving: Visual language navigation via multi-expert discussions," *arXiv preprint arXiv:2309.11382*, 2023. [Online]. Available: https://www.arxiv.org/abs/2309.11382

[13] W. Liuyi, H. Zongtao, D. Ronghao, C. Huiyi, L. Chengju, and C. Qijun, "Causality-based cross-modal representation learning for vision-and-language navigation," *arXiv preprint arXiv:2403.03405*, 2024. [Online]. Available: https://www.arxiv.org/abs/2403.03405

[14] Alex, P. er, S. Cordelia, and S. Chen, "Episodic transformer for vision-and-language navigation," *arXiv preprint arXiv:2105.06453*, 2021. [Online]. Available: https://www.arxiv.org/abs/2105.06453

[15] W. Liuyi, L. Chengju, H. Zongtao, L. Shu, Y. Qingqing, C. Huiyi, and C. Qijun, "Pasts: Progress-aware spatio-temporal transformer speaker for vision-and-language navigation," *arXiv preprint arXiv:2305.11918*, 2023. [Online]. Available: https://www.arxiv.org/abs/2305.11918

[16] L. Jialu and B. Mohit, "Improving vision-and-language navigation by generating future-view image semantics," *arXiv preprint arXiv:2304.04907*, 2023. [Online]. Available: https://www.arxiv.org/abs/2304.04907

[17] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," 2023. [Online]. Available: https://arxiv.org/abs/2310.08576

[18] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. 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*, 2022, pp. 16 537–16 547.

[19] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," *Advances in neural information processing systems*, vol. 34, pp. 5834–5847, 2021.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[21] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar, and A. Farhadi, "Robothor: An open simulation-to-real embodied ai platform," 2020. [Online]. Available: https://arxiv.org/abs/2004.06799

[22] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," 2017. [Online]. Available: https://arxiv.org/abs/1709.06158

[23] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, "Reverie: Remote embodied visual referring expression in real indoor environments," 2020. [Online]. Available: https://arxiv.org/abs/1904.10151

[24] H. Tan, L. Yu, and M. Bansal, "Learning to navigate unseen environments: Back translation with environmental dropout," *arXiv preprint arXiv:1904.04195*, 2019.

[25] C. Liu, F. Zhu, X. Chang, X. Liang, Z. Ge, and Y.-D. Shen, "Vision-language navigation with random environmental mixup," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1644–1654.

[26] T.-J. Fu, X. E. Wang, M. F. Peterson, S. T. Grafton, M. P. Eckstein, and W. Y. Wang, "Counterfactual vision-and-language navigation via adversarial path sampler," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 71–86.

[27] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.16986

[28] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. K. Wong, "Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation," 2024. [Online]. Available: https://arxiv.org/abs/2401.07314

[29] Y. Long, X. Li, W. Cai, and H. Dong, "Discuss before moving: Visual language navigation via multi-expert discussions," 2023. [Online]. Available: https://arxiv.org/abs/2309.11382

[30] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, "Towards learning a generalist model for embodied navigation," 2024. [Online]. Available: https://arxiv.org/abs/2312.02010

[31] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson, "Pathdreamer: A world model for indoor navigation," 2021. [Online]. Available: https://arxiv.org/abs/2105.08756

[32] J. Li and M. Bansal, "Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation," 2023. [Online]. Available: https://arxiv.org/abs/2305.19195

[33] W. Ye, C. Ji, Z. Chen, J. Gao, X. Huang, S.-H. Zhang, W. Ouyang, T. He, C. Zhao, and G. Zhang, "Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2410.24203

[34] S. Wang, D. Zhou, L. Xie, C. Xu, Y. Yan, and E. Yin, "Panogen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation," *Neural Networks*, vol. 187, p. 107320, Jul. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2025.107320

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[36] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.

[37] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[38] X. Kong, J. Chen, W. Wang, H. Su, X. Hu, Y. Yang, and S. Liu, "Controllable navigation instruction generation with chain of thought prompting," in *European Conference on Computer Vision*. Springer, 2024, pp. 37–54.

[39] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, "Instructnav: Zero-shot system for generic instruction navigation in unexplored environment," *arXiv preprint arXiv:2406.04882*, 2024.

[40] B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang, "Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[41] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[44] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024. [Online]. Available: https://arxiv.org/abs/2401.14159

[45] B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang, "Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2403.07376

[46] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[47] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 137–13 146.

[48] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "A recurrent vision-and-language bert for navigation," *arXiv preprint arXiv:2011.13922*, 2020.

[49] A. Perincherry, J. Krantz, and S. Lee, "Do visual imaginations improve vision-and-language navigation agents?" 2025. [Online]. Available: https://arxiv.org/abs/2503.16394

[50] G. Ilharco, V. Jain, A. Ku, E. Ie, and J. Baldridge, "General evaluation for instruction conditioned navigation using dynamic time warping," *arXiv preprint arXiv:1907.05446*, 2019.

[51] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, "Esc: Exploration with soft commonsense constraints for zero-shot object navigation," 2023. [Online]. Available: https://arxiv.org/abs/2301.13166