FantasyVLN: Unified Multimodal Chain-of-Thought Reasoning for Vision-and-Language Navigation

Jing Zuo[1,3*§] Lingzhou Mu[2,3*§] Fan Jiang[3*‡] Chengcheng Ma[3] Mu Xu[3] Yonggang Qi[1†]

[1]Beijing University of Posts and Telecommunications [2]Tsinghua University [3]Fantasy AIGC Team

jiangfan0576@gmail.com[‡] qiyg@bupt.edu.cn[†]

Abstract

Achieving human-level performance in Vision-and-Language Navigation (VLN) requires an embodied agent to understand textual instructions, perceive visual observations, and reason over long action sequences. Recent works, such as NavCoT and NavGPT-2, demonstrate the potential of Chain-of-Thought (CoT) reasoning for improving interpretability and long-horizon planning. Moreover, multimodal extensions like OctoNav-R1 and CoT-VLA further validate CoT as a promising pathway toward human-like navigation reasoning. However, existing approaches face critical drawbacks: purely textual CoTs lack visual perception and easily overfit to sparse annotated reasoning steps, while multimodal CoTs incur severe token inflation by generating imagined visual observations, making real-time navigation impractical. In this work, we propose FantasyVLN, a unified implicit reasoning framework that preserves the benefits of CoT reasoning without explicit token overhead. Specifically, imagined visual tokens are encoded into a compact latent space using a pretrained Visual AutoRegressor (VAR) during CoT reasoning training, and the model jointly learns from textual, visual, and multimodal CoT modes under a unified multi-CoT strategy. At inference, our model performs direct instruction-to-action mapping while still enjoying reasoning-aware representations. Extensive experiments on LH-VLN show that our approach achieves reasoning-aware yet real-time navigation, improving success rates and efficiency while reducing inference latency by an order of magnitude compared to explicit CoT methods. Code is available at https://github.com/Fantasy-AMAP/fantasy-vln.

1. Introduction

Vision-and-Language Navigation (VLN) aims to enable an embodied agent to follow natural-language instructions and navigate complex visual environments [1, 6, 8, 22]. Solving this task requires understanding textual instructions, perceiving visual observations, and performing long-horizon reasoning to plan a sequence of actions. Especially for real-world navigation scenarios involving multi-stage and long-horizon trajectories [13], robust multimodal reasoning is critical. As illustrated in Fig. 1, this requires effectively integrating linguistic intent with visual context over extended temporal dependencies. Despite the progress made by recent multimodal large models, achieving effective multimodal reasoning in VLN remains challenging due to the language–vision gap and the need for interpretable yet sample-efficient reasoning mechanisms.

The recent success of large language models (LLMs) has inspired the integration of Chain-of-Thought (CoT) reasoning into embodied navigation to improve interpretability and long-horizon decision-making. Methods such as NavCoT [11], NavGPT-2 [33] and OmniNav [24] employ step-by-step textual reasoning to decompose navigation instructions or generate intermediate subgoals. However, their reasoning remains confined to the textual modality, typically by translating observations into captions, thereby limiting true multimodal reasoning essential for successful navigation. This limitation is compounded by the difficulty of annotating CoT supervision in VLN, as highlighted by EvolveNav [10], where multiple valid action sequences often exist. Moreover, explicitly supervised CoT reasoning tends to overfit training distributions and generalize poorly to unseen environments.

Lately, works such as CoT-VLA [31], VISTA [7], RBF++ [2], and OctoNav-R1 [5] have extended CoT reasoning into visual or multimodal domains for improved generalization. While this multimodal CoT paradigm marks an important step forward, it also introduces new challenges for long-horizon navigation. In particular, modeling reasoning chains across language and vision requires the model

*Equal contribution.
†Corresponding author.
‡Project Leader.
§Work done during internship at Fantasy AIGC Team.

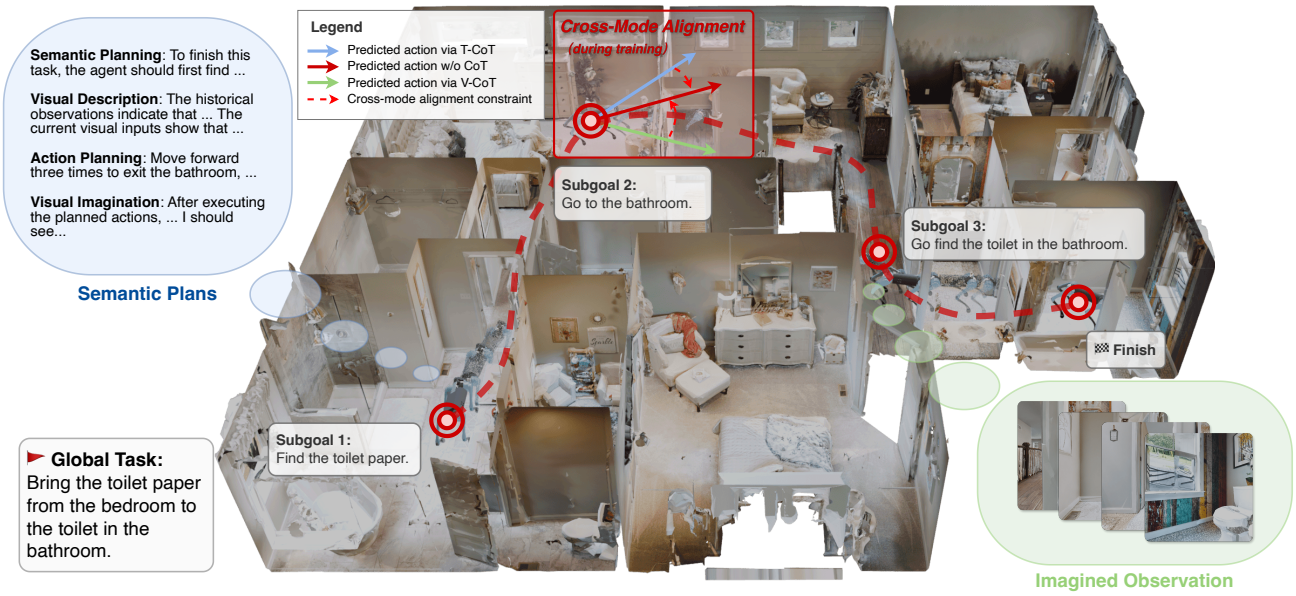

Figure 1. Aligning semantic and visual Reasoning. Complex real-world navigation tasks are typically multi-stage and long-horizon. Addressing them requires both textual and visual reasoning to jointly enhance semantic planning and visual perception. A critical challenge is how to effectively align these two distinct reasoning capabilities within a unified framework.

to iteratively generate and interpret imagined intermediate observations at each step, leading to severe token inflation. A typical reasoning step spanning 5–7 actions expands into over 3k–5k tokens, an order of magnitude larger than purely textual CoTs (usually $<$500 tokens). This sequence explosion drastically increases both training and inference latency, rendering real-time navigation infeasible even on high-end GPUs.

To address these challenges, we propose a unified implicit reasoning framework that retains the benefits of CoT-style reasoning while eliminating its explicit token overhead during inference. The key idea is twofold: (i) During training, we encode the imagined observation tokens generated by multimodal CoT reasoning into a compact latent space using a pretrained Visual AutoRegressive (VAR) model. This significantly reduces sequence length and training cost without compromising the richness of visual reasoning. (ii) At inference, the agent performs direct instruction-to-action mapping while still leveraging reasoning-aware representations, inspired by the *train-with-CoT, infer-without-CoT* paradigm of Aux-Think [16].

Concretely, we introduce a unified multi-CoT training strategy that jointly learns from textual-only, visual-only, and textual–visual CoT modes using a special tag token to indicate each mode. This design unifies both the input format and model parameters within a single framework. During training, we align the action predictions from CoT-based reasoning modes with those from direct prediction (without CoT), enforcing modality-invariant reasoning representations. Consequently, the model learns implicit reasoning capabilities that generalize effectively without explicit CoT supervision or overfitting to training distributions.

To this end, our contributions are summarized as follows: (i) We propose the first unified implicit CoT reasoning framework that integrates textual, visual, and multimodal CoT paradigms within a single model. Unlike prior explicit CoT methods, our approach trains with diverse reasoning modes but performs inference without generating CoT sequences, achieving reasoning-aware yet real-time navigation. (ii) We introduce a gating-based multi-CoT learning mechanism that allows seamless switching among reasoning modes and direct action prediction under shared parameters. By aligning CoT-driven and direct action predictions, our model learns consistent, modality-invariant reasoning representations. (iii) To reduce the token explosion in multimodal reasoning, we compress imagined observation tokens into a compact latent space using a pretrained Visual Autoregressive (VAR), improving training efficiency while preserving textual–visual reasoning capacity. (iv) Extensive experiments on the challenging LH-VLN benchmark demonstrate that our method substantially improves navigation success and efficiency in multi-stage and long-horizon scenarios, while reducing inference latency by an order of magnitude compared to explicit CoT approaches.

2. Related Works

2.1. Vision-and-Language Navigation

Early VLN research often simplifies the task into a decision-making process within discrete environments based on

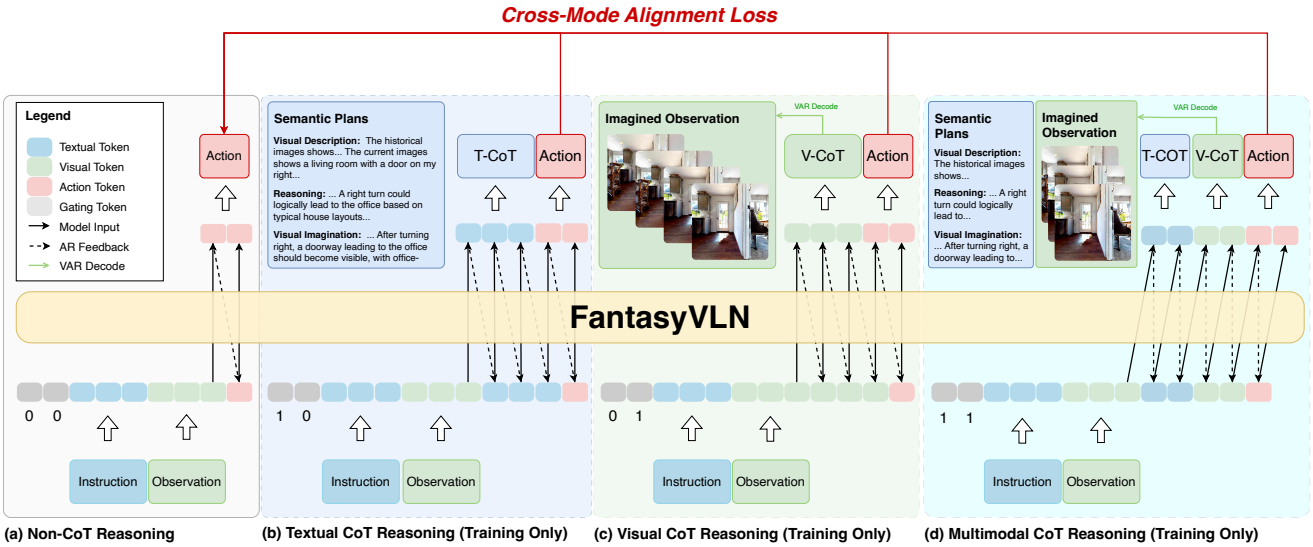

Figure 2. Unified multimodal Chain-of-Thought reasoning framework. Within a shared representation space, a single model supports four reasoning modes: (a) non-CoT reasoning for real-time inference, (b) textual CoT for semantic planning, (c) compact visual CoT for latent future imagination, and (d) multimodal CoT integrating both modalities. A flexible gating mechanism facilitates seamless transitions among the four modes, while an alignment constraint enforces representation consistency during training.

panoramic views, as seen in benchmarks like R2R [1] and RxR [9]. Methods from this period typically decompose the pipeline into distinct modules, optimizing them via imitation [12, 15, 21] or reinforcement learning [18, 23] with module-specific auxiliary objectives. However, these methods struggle to generalize to continuous and unseen environments (e.g., VLN-CE [8]) due to the discrete-to-real gap and the lack of unified representations in modular designs. To address these limitations, recent efforts have shifted towards end-to-end policy learning, wherein large-scale pre-trained vision-language models are post-trained on offline expert video trajectories. For example, NaVid [25] demonstrates that fine-tuning VLMs on monocular videos enables purely visual navigation. Uni-NaVid [26] integrates several embodied tasks and planning within a cohesive architecture, while NaVILA [3] further extends this approach to legged robots. Despite achieving groundbreaking progress in basic scenarios, these models often fall short in complex, long-horizon tasks (e.g., LH-VLN) since they lack high-level reasoning capabilities. To address this shortcoming, CoT reasoning has emerged as a pivotal paradigm for Embodied AI [27, 31]. Specifically within VLN, NavGPT [33] leverages the zero-shot CoT abilities of GPT-4, whereas Aux-Think [16] incorporates auxiliary CoT supervision to internalize deduction patterns during training. Nevertheless, current CoT-based approaches predominantly restrict their thought processes to a single modality, leaving the vast potential of multimodal CoT largely untapped. In this paper, we adhere to the continuous environment setting and pioneer a systematic investigation into multimodal CoT reasoning for VLN.

2.2. Chain-of-Thought Reasoning

Chain-of-Thought reasoning empowers large language models to tackle complex problems by explicitly formulating intermediate deduction steps [20]. Subsequent variants, such as Self-Consistency [19] and Least-to-Most Prompting [32], further augment both robustness and explainability. More recently, this paradigm has been extended to vision-language models [29]. Based on the modality of the intermediate outputs, existing approaches can be broadly categorized into three types: Textual CoT, Visual CoT, and Multimodal CoT. In particular, Textual CoT [30] in VLMs largely mirrors the conventional text-based format of standard LLMs. Conversely, Visual CoT methods, such as CoT-VLA [31] and DreamVLA [28], synthesize future frames prior to action execution in manipulation tasks, whereas Multimodal CoT [4] jointly predicts paired textual and visual states for embodied scenarios. To the best of our knowledge, FantasyVLN stands as the first unified framework to seamlessly integrate these three CoT paradigms.

3. Methods

3.1. Overview

We propose FantasyVLN, a framework that intrinsically integrates multimodal reasoning while enabling implicit reasoning for highly efficient inference. As illustrated in Figure 2, FantasyVLN internalizes diverse CoT reasoning patterns across modalities via end-to-end joint training. Crucially, it enhances the non-CoT reasoning mode through a

cross-mode alignment constraint. This design leverages the advantages of both textual and visual CoT reasoning without the inference latency typically associated with explicit CoT generation. Furthermore, we perform visual CoT reasoning within the latent space of the VAR model [14], significantly improving both training and inference efficiency compared to conventional pixel-space methods.

Below, we detail the problem setup, compact visual CoT reasoning, unified multimodal CoT reasoning and cross-mode alignment constraint.

3.2. Problem Setup

We formulate VLN as a sequential decision-making process, aiming to develop an embodied agent π_θ that can navigate continuous 3D environments $\mathcal{E}$ following natural language instructions $\mathcal{I}$. Let s_0 denote the initial state (i.e., location and orientation), and $\mathcal{S}$ denote the action space. At each timestep t, the agent π_θ receives multi-view visual observations $\mathcal{O}_t \in \mathcal{E}$. Given the partial observability of the environment, the agent predicts future actions $\mathcal{A}_{t+1} \in \mathcal{S}$ conditioned on the given instruction $\mathcal{I}$ and the accumulated historical observations $\{\mathcal{O}_{\leq t}\}$. Subsequently, the predicted actions $\mathcal{A}_{t+1}$ are executed, transitioning the agent π_θ to a new state according to the underlying environment dynamics. This sequential interaction process continues until a stop action is executed or the maximum step limit T is reached.

3.3. Compact Visual CoT reasoning

Existing visual CoT methods [28, 31] primarily decode intermediate reasoning steps in the pixel space, which makes a single reasoning step require predicting hundreds or even thousands of tokens. To prevent the visual CoT from becoming a computational bottleneck of the entire framework, we propose compact visual CoT, which decodes visual reasoning steps in the VAR latent space. To this end, we adopt a pretrained VAR model [14] as the image decoder of a VLM and construct a joint vocabulary for them. As shown in Fig. 2 (c), the VLM takes the instruction $\mathcal{I}$ and visual observations $\{\mathcal{O}_{\leq t}\}$ as inputs, and predicts latent future images $\widehat{\mathcal{V}}_{t+1}$ and actions $\widehat{\mathcal{A}}_{t+1}$, where $\widehat{\mathcal{V}}_{t+1}$ are early-scale VAR latents. Subsequently, the VAR model further decodes $\widehat{\mathcal{V}}_{t+1}$ into pixel future images $\widehat{\mathcal{O}}_{t+1}$ via next-scale prediction. We freeze the VAR model during training, while the VLM learns to predict actions conditioned on latent future imagination. During inference, we use only the VLM to perform visual CoT-based navigation without explicit VAR decoding, which largely improves reasoning efficiency.

3.4. Unified Multimodal CoT Reasoning

We further proposed a unified multimodal CoT reasoning framework named FantasyVLN, which integrates textual, compact visual, and textual-visual paired CoT reasoning within a single VLN model.

Textual CoT Reasoning in VLN. Conditioned on the instruction $\mathcal{I}$ and visual observations $\mathcal{O}_{\leq t}$, the agent first generates textual reasoning chains $\mathcal{T}_t$ followed by predicting the subsequent actions $\mathcal{A}_{t+1}$. We define the textual reasoning chain $\mathcal{T}_t$ as a structured cognitive process, wherein the agent first decomposes the textual instruction $\mathcal{I}$ into several executable sub-tasks, subsequently infers the currently active sub-task based on visual observations $\mathcal{O}_{\leq t}$, and finally deduces the optimal strategy to be adopted. By explicitly articulating this step-by-step decision-making process, textual CoT reasoning significantly enhances the agent's capability to resolve complex navigation tasks.

Compact Visual CoT Reasoning in VLN. The agent first anticipates future visual observations $\mathcal{O}_{t+1}$, and subsequently deduces the corresponding future actions $\mathcal{A}_{t+1}$ conditioned on these synthesized visual states. Such predictive visual modeling compels the agent to internalize spatial geometry and scene dynamics, thereby fostering more robust visual representations for navigating complex environments. Notably, this entire process is instantiated via our proposed compact visual CoT, which compresses $\mathcal{O}_{t+1}$ into latent tokens $\mathcal{V}_{t+1}$ to guarantee high reasoning efficiency.

Multimodal CoT Reasoning in VLN. Following [4], we conceptualize multimodal reasoning chains as a synergistic integration of textual and visual chains, wherein the agent concurrently generates paired textual–visual reasoning steps $\widehat{\mathcal{M}}_t = [\widehat{\mathcal{T}}_t, \widehat{\mathcal{V}}_t]$ and predicts future actions $\widehat{\mathcal{A}}_{t+1}$. This multimodal CoT reasoning provides an additional decision pattern conditioned on joint textual and visual reasoning steps, which further improves navigation robustness.

Unified Reasoning framework. To seamlessly integrate the aforementioned reasoning modes within a unified framework, we introduce a gating mechanism wherein two binary signals, $g_\mathcal{T}$ and $g_\mathcal{V}$, explicitly govern the activation of the textual and visual reasoning pathways, respectively. By combining these control signals with the standard navigation inputs, the agent dynamically adopts the corresponding reasoning modes. It autoregressively formulates the designated reasoning chain $\widehat{\mathcal{R}}_{t+1}$ prior to predicting the future action $\widehat{\mathcal{A}}_{t+1}$:

$$[\widehat{\mathcal{R}}_{t+1}, \widehat{\mathcal{A}}_{t+1}] = \pi_\theta\big(\mathcal{I}, \{\mathcal{O}_{\leq t}\}, g_\mathcal{T}, g_\mathcal{V}\big), \tag{1}$$

where

$$\widehat{\mathcal{R}}_{t+1} = \begin{cases} \text{None}, & \text{if } (g_\mathcal{T}, g_\mathcal{V}) = (0,0), \\ \widehat{\mathcal{T}}_{t+1}, & \text{if } (g_\mathcal{T}, g_\mathcal{V}) = (1,0), \\ \widehat{\mathcal{V}}_{t+1}, & \text{if } (g_\mathcal{T}, g_\mathcal{V}) = (0,1), \\ \widehat{\mathcal{M}}_{t+1}, & \text{if } (g_\mathcal{T}, g_\mathcal{V}) = (1,1). \end{cases} \tag{2}$$

Joint Training of Diverse Reasoning Modes. We adopt end-to-end joint training strategy to construct a unified representation space across different reasoning modes. Specifically, during the training phase, we stochastically sample the gating signals $(g_{\mathcal{T}}, g_{\mathcal{V}})$ for each training instance, dynamically routing the forward pass through different reasoning pathways. This strategy compels the model to internalize diverse CoT capabilities within a shared parameter space. The overall optimization objective is formulated as a weighted combination of the task-specific losses corresponding to each activated mode:

$$\begin{aligned}\mathcal{L}_{\text{Joint}} = &(\neg g_{\mathcal{T}} \wedge \neg g_{\mathcal{V}})\, \mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{A}}_{t+1}, \mathcal{A}_{t+1}\big) \\ &+ (g_{\mathcal{T}} \wedge \neg g_{\mathcal{V}})\, \mathcal{L}_{\text{CE}}\big([\widehat{\mathcal{T}}_{t+1}, \widehat{\mathcal{A}}_{t+1}], [\mathcal{T}_{t+1}, \mathcal{A}_{t+1}]\big) \\ &+ (\neg g_{\mathcal{T}} \wedge g_{\mathcal{V}})\, \mathcal{L}_{\text{CE}}\big([\widehat{\mathcal{V}}_{t+1}, \widehat{\mathcal{A}}_{t+1}], [\mathcal{V}_{t+1}, \mathcal{A}_{t+1}]\big) \\ &+ (g_{\mathcal{T}} \wedge g_{\mathcal{V}})\, \mathcal{L}_{\text{CE}}\big([\widehat{\mathcal{M}}_{t+1}, \widehat{\mathcal{A}}_{t+1}], [\mathcal{M}_{t+1}, \mathcal{A}_{t+1}]\big),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ denotes the causal cross-entropy loss.

3.5. Cross-Mode Alignment Constraint

However, despite joint training, different reasoning pathways can still yield divergent navigation decisions. We address this by enforcing decision alignment between auxiliary modes and a primary reference mode during training, which simultaneously fosters robust unified multimodal representations. Given the real-time execution demands of the VLN task, we designate the non-CoT reasoning pathway as this primary reference, thereby circumventing the latency overhead associated with decoding massive explicit reasoning tokens. Let $\widehat{\mathcal{A}}_{t+1}, \widehat{\mathcal{A}}^{\mathcal{T}}_{t+1}, \widehat{\mathcal{A}}^{\mathcal{V}}_{t+1}$, and $\widehat{\mathcal{A}}^{\mathcal{M}}_{t+1}$ denote the action predictions from the non-CoT, textual, compact visual, and multimodal reasoning modes, respectively. In each training iteration, we first optimize the non-CoT reasoning mode using the following objective:

$$\mathcal{L}_{\text{non-CoT}} = \mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{A}}_{t+1}, \mathcal{A}_{t+1}\big), \tag{4}$$

where

$$\widehat{\mathcal{A}}_{t+1} = \pi_\theta\big(\mathcal{I}, \{o_{\le t}\}, g_{\mathcal{T}} = 0, g_{\mathcal{V}} = 0\big). \tag{5}$$

We then extract the soft targets $\widetilde{\mathcal{A}}_{t+1}$ by executing the forward pass (5) with the updated agent π_θ. Finally, we formulate the cross-mode aligned joint objective for unified multimodal CoT reasoning as follows:

$$\mathcal{L}^*_{\text{Joint}} = \mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{CoT}}, \tag{6}$$

where

$$\begin{aligned}\mathcal{L}_{\text{Align}} = &\mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{A}}^{\mathcal{T}}_{t+1}, \widetilde{\mathcal{A}}_{t+1}\big) \\ &+ \mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{A}}^{\mathcal{V}}_{t+1}, \widetilde{\mathcal{A}}_{t+1}\big) + \mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{A}}^{\mathcal{M}}_{t+1}, \widetilde{\mathcal{A}}_{t+1}\big),\end{aligned} \tag{7}$$

and

$$\begin{aligned}\mathcal{L}_{\text{CoT}} = &\mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{T}}_{t+1}, \mathcal{T}_{t+1}\big) \\ &+ \mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{V}}_{t+1}, \mathcal{V}_{t+1}\big) + \mathcal{L}_{\text{CE}}\big(\widehat{\mathcal{M}}_{t+1}, \mathcal{M}_{t+1}\big).\end{aligned} \tag{8}$$

We alternately minimize the non-CoT objective (4) and the cross-mode aligned joint objective (6) until convergence. Throughout this alternating optimization process, all reasoning modes operate on shared inputs using identical network parameters and are constrained by the same supervisory signals. This inherently embeds diverse CoT reasoning patterns into a unified latent representation. Consequently, our framework naturally supports implicit reasoning, enabling the agent to seamlessly integrate the strengths of multiple reasoning modes without incurring any additional computational overhead during inference. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Cross-Mode Aligned Joint Training

Input: Dataset $\mathcal{D}$, parameters θ, learning rate η, alignment weight λ_{align}
Output: Trained parameters θ^*
while not converged **do**
  $[\mathcal{I}, \{o_{\le t}\}, \mathcal{T}_{t+1}, \mathcal{V}_{t+1}, \mathcal{A}_{t+1}] \sim \mathcal{D}$
  $\widehat{\mathcal{A}}_{t+1} \leftarrow \pi_\theta(\mathcal{I}, \{o_{\le t}\}, g_{\mathcal{T}}{=}0, g_{\mathcal{V}}{=}0)$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{CE}}(\widehat{\mathcal{A}}_{t+1}, \mathcal{A}_{t+1})$
  $\widetilde{\mathcal{A}}_{t+1} \leftarrow \text{sg}\big[\pi_\theta(\mathcal{I}, \{o_{\le t}\}, g_{\mathcal{T}}{=}0, g_{\mathcal{V}}{=}0)\big]$
  $[\widehat{\mathcal{T}}_{t+1}, \widehat{\mathcal{A}}^{\mathcal{T}}_{t+1}] \leftarrow \pi_\theta(\mathcal{I}, \{o_{\le t}\}, g_{\mathcal{T}}{=}1, g_{\mathcal{V}}{=}0)$
  $[\widehat{\mathcal{V}}_{t+1}, \widehat{\mathcal{A}}^{\mathcal{V}}_{t+1}] \leftarrow \pi_\theta(\mathcal{I}, \{o_{\le t}\}, g_{\mathcal{T}}{=}0, g_{\mathcal{V}}{=}1)$
  $[\widehat{\mathcal{M}}_{t+1}, \widehat{\mathcal{A}}^{\mathcal{M}}_{t+1}] \leftarrow \pi_\theta(\mathcal{I}, \{o_{\le t}\}, g_{\mathcal{T}}{=}1, g_{\mathcal{V}}{=}1)$
  Compute $\mathcal{L}^*_{\text{Joint}}$ using Eq. (6)
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^*_{\text{Joint}}$
end while
$\theta^* \leftarrow \theta$
return θ^*

4. Experiments

4.1. Experimental Setup

Benchmark. We evaluate FantasyVLN on LH-VLN [13], a benchmark for multi-stage and long-horizon navigation. We select it because multi-stage tasks assess semantic planning, while long-horizon settings test visual robustness against error accumulation. Following official protocols, we conduct online evaluation on the test set, where both tasks and environments are entirely unseen.

Baselines. We compare FantasyVLN against representative methods across four categories: (i) textual CoT (Aux-Think [16]); (ii) visual CoT (CoT-VLA [31] and World-VLA [27]); (iii) memory-based (MGDM [13]); and (iv)

Table 1. Comparison of navigation accuracy with several advanced VLN methods on LH-VLN. The best and second-best results are marked in bold and underlined, respectively.

CoT Modal	Methods	SR	ISR	CSR	CGT
None/ZS	Random	0	0	0	0
	GLM-4v prompt	0	0	0	0
	GPT-4 + NaviLLM	0	2.19	1.45	2.61
	MGDM	0	2.34	1.65	<u>2.91</u>
Visual	CoT-VLA	0	0	0	0
	WorldVLA	0	0	0	0
Textual	Aux-Think	<u>0.65</u>	<u>3.16</u>	<u>2.04</u>	1.47
Unified Multimodal	FantasyVLN	**2.44**	**11.01**	**9.64**	**8.99**

Table 2. Comparison of inference efficiency across different CoT reasoning methods. The best results are marked in bold.

Decoding Mode	Methods	Model Size	APS
Explicit	CoT-VLA	7B	0.19
Implicit	WorldVLA	7B	1.02
	Aux-Think	8B	0.97
	FantasyVLN	7B	**1.03**

standard baselines from LH-VLN (GLM-4V, NaviLLM, and GPT-4 with NaviLLM). For fair comparison, all models are trained on the LH-VLN training set, with optimal checkpoints selected via the validation set. Due to the absence of official codes, we re-implement Aux-Think and CoT-VLA based on paper details. We adapt the official WorldVLA implementation for LH-VLN compatibility. All other baselines adopt the original LH-VLN codebase.

Metrics. In line with [13], we employ Success Rate (SR), Independent Success Rate (ISR), Conditional Success Rate (CSR), and CSR weighted by Ground Truth (CGT) to measure multi-stage navigation accuracy. Specifically, SR and ISR indicate the success of the entire task and individual subtasks, respectively. CSR penalizes ISR for prior subtask failures, and CGT further weights CSR by the expert trajectory length. Furthermore, we introduce Actions Per Second (APS) to assess inference efficiency:

$$\text{APS} = \frac{N_{\text{act}}}{T_{\text{nav}}}, \tag{9}$$

where N_{act} and T_{nav} denote the total executed actions and overall navigation time in seconds, respectively.

4.2. Main Results

Navigation Accuracy. Table 1 summarizes the quantitative comparison on LH-VLN. Notably, FantasyVLN achieves state-of-the-art performance across all metrics,

Table 3. Comparison of navigation accuracy with different reasoning mode combinations on LH-VLN.

Non-CoT	T-CoT	V-CoT	MM-CoT	SR	ISR	CSR	CGT
✓				0	2.01	1.51	1.55
✓	✓			0.98	8.26	6.60	6.15
✓		✓		<u>1.46</u>	**11.19**	**9.66**	8.84
✓			✓	0.49	7.77	6.48	<u>8.89</u>
✓	✓	✓	✓	**2.44**	<u>11.01</u>	<u>9.64</u>	**8.99**

Table 4. Performance comparison of SR, ISR, CSR, and CGT with and without cross-mode alignment.

Alignment Constraint	SR	ISR	CSR	CGT
✗	0	2.39	1.19	1.28
✓	**2.44**	**11.01**	**9.64**	**8.99**

which validates the benefits of a unified multimodal CoT reasoning framework. Aux-Think ranks second and significantly outperforms other baselines, which confirms the advantages of implicit reasoning. Surprisingly, visual CoT methods (CoT-VLA and WorldVLA) fail completely with zero scores. We attribute this failure to the limited training data in LH-VLN and the inherent difficulty of future scene video generation in navigation tasks, as opposed to fixed-camera manipulation setups. This stark contrast highlights the superiority of our compact visual CoT over pixel-level visual CoT under identical data constraints. Finally, other non-CoT baselines exhibit poor performance on complex navigation tasks because they lack high-level reasoning capabilities.

Inference Efficiency. Table 2 compares the APS across various CoT methods. Implicit reasoning methods, including FantasyVLN, Aux-Think, and WorldVLA, achieve comparable efficiency and substantially outperform the explicit reasoning approach CoT-VLA. This performance gap stems from distinct decoding strategies during inference. Implicit approaches decode a single token per action, whereas explicit methods output lengthy intermediate steps with thousands of tokens. With similar model sizes, implicit CoT operates approximately five times faster than explicit CoT. Consequently, implicit architectures better satisfy the real-time demands of VLN tasks.

4.3. Ablation Studies

Contribution of Each Reasoning Mode. We evaluate various combinations of non-CoT and CoT strategies to verify the contribution of each reasoning mode. Table 3 demonstrates that the addition of any single CoT mode to the non-CoT baseline yields substantial performance gains across all metrics. Furthermore, the unified framework with

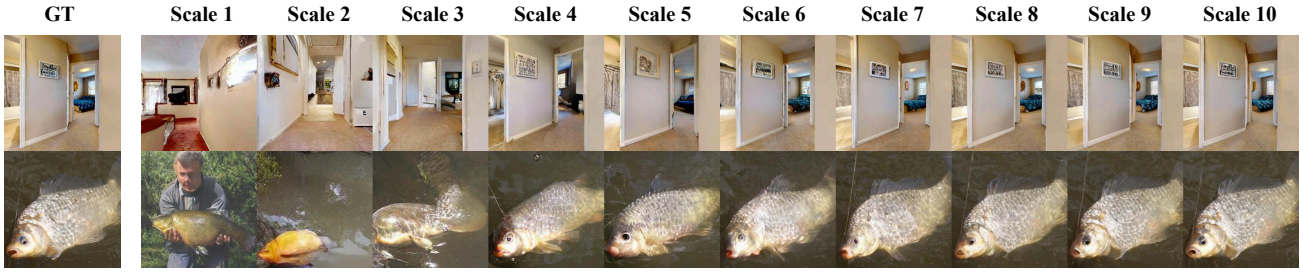

Figure 3. Qualitative comparison of VAR reconstruction across latent scales. Coarse-to-fine reconstruction shows that early scales (1–4) capture the primary structural content of the images.

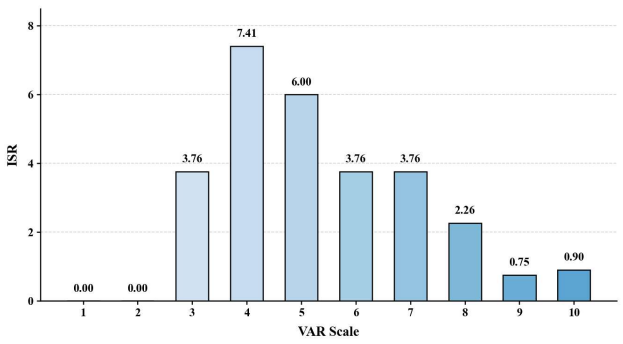

Figure 4. ISR variation with respect to different VAR scales.

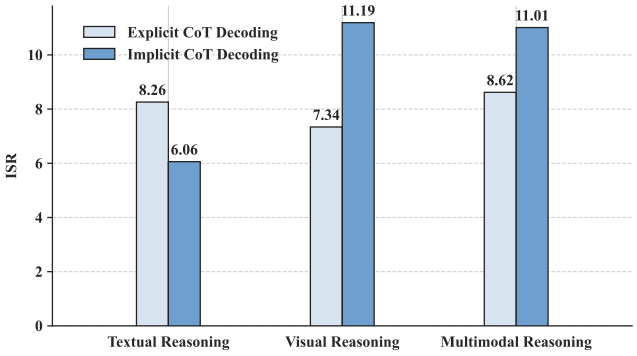

Figure 5. ISR variation with explicit and implicit CoT decoding.

all four modes maximizes overall navigation capabilities, particularly in metrics like SR and CGT. This outcome confirms that diverse CoT modalities provide complementary benefits for complex navigation tasks.

Effect of Cross-Mode Alignment Constraint. We evaluate FantasyVLN with and without the cross-mode alignment to verify its necessity. Table 4 indicates that the inclusion of this constraint drastically elevates all performance metrics. Notably, SR improves from 0 to 2.44, and ISR jumps from 2.39 to 11.01. This massive gap emerges because the constraint facilitates the formation of a unified representation space through decision consistency alignment across diverse reasoning modes. Consequently, explicit feature alignment acts as an indispensable mechanism for unified reasoning architectures.

VAR Scale Selection. We conduct ablation studies to determine the optimal VAR scale for latent visual CoT. Specifically, we train a FantasyVLN variant equipped with non-CoT and compact visual CoT on an LH-VLN subset. This variant utilizes VAR latent tokens across scales 1 through 10 under identical training epochs. Figure 4 illustrates the ISR results. Scale 4 achieves the highest ISR of 7.41, whereas other scales exhibit significant performance drops. We attribute this phenomenon to the information density across scales. Smaller scales, such as 1 and 2, lack essential navigation information, and thus the model merely fits noise. Conversely, larger scales (≥ 6) encode high-frequency textures. These redundant details hinder the association between visual imagination and action decisions. To validate this hypothesis, we design a VAR reconstruction experiment. We feed the ground truth VAR latents up to a specific scale into the model to predict the remaining latents. A decoder then reconstructs the images from these combined latents. Figure 3 visualizes the reconstructed results. Early scales fail to preserve the core layout and semantics. Scale 4 successfully captures essential scene structures. Subsequent larger scales merely append fine-grained visual textures. These qualitative observations perfectly align with the quantitative outcomes from the first experiment.

4.4. More Results

Explicit vs. Implicit CoT Decoding. We investigate the impact of explicit and implicit CoT decoding on navigation performance. Specifically, we select the optimal scale 4 variant from the ablation study. We conduct inference via explicit and implicit CoT decoding on a test subset proportional to the LH-VLN training split. Figure 5 illustrates the detailed ISR outcomes. Under textual reasoning, explicit decoding achieves an ISR of 8.26, which outperforms

the implicit counterpart at 6.06. Conversely, implicit decoding dominates under visual and multimodal reasoning scenarios. It yields an ISR of 11.19 and 11.01 respectively, whereas explicit decoding only reaches 7.34 and 8.62. We attribute this modality discrepancy to model constraints and task complexity. The base model `Qwen2.5-VL` lacks native image generation capabilities. This deficit complicates the acquisition of future video frame synthesis skills. Furthermore, continuous 3D scene navigation transforms visual imagination into a complex scene generation task. This process is inherently more difficult than textual deduction. Prior work [17] establishes that CoT enhances the sampling probability of correct tokens via intermediate steps. Consequently, flawed CoT deductions suppress correct answer probabilities. Because visual reasoning proves difficult to master, explicit token-level error accumulation becomes severe. Implicit reasoning effectively circumvents this risk.

Training Efficiency. We investigate the fundamental causes for the poor performance of pixel-level visual CoT methods on LH-VLN. We train FantasyVLN and WorldVLA on identical data under their respective optimal parameters. Figure 6 illustrates the train token accuracy across iterations. FantasyVLN achieves full convergence at approximately 3,000 iterations. In contrast, WorldVLA merely reaches a 0.5 accuracy after 13,800 iterations. This stark contrast reveals a 4.6-fold acceleration in convergence speed for FantasyVLN. FantasyVLN executes visual CoT within the VAR latent space. This architecture predicts merely a dozen tokens per image. Conversely, WorldVLA decodes hundreds of tokens for a single frame. Long-sequence learning intrinsically demands extensive time and massive data volumes. Consequently, compact visual CoT in the latent space substantially elevates overall training efficiency.

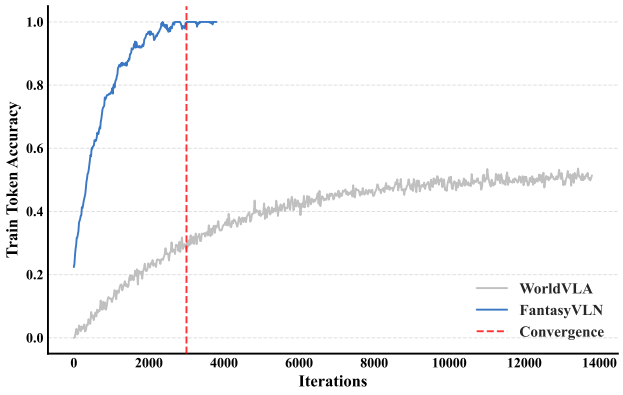

Figure 6. Training efficiency comparison of visual CoT reasoning methods. FantasyVLN converges approximately 4.6x faster than WorldVLA.

5. Limitations and Future Work

5.1. Limitations

Base Model Choice. While FantasyVLN integrates diverse CoT modes, its base model is not natively designed for unified generation and understanding. Thus, current results may not reflect the framework's performance upper bound, particularly for visual CoT requiring future scene imagination.

Limited Data Scale. The LH-VLN benchmark provides inherently limited training data. While compact visual CoT clearly outperforms explicit visual generation in this regime, its scaling behavior on significantly larger datasets remains unexplored.

Training Paradigm. Like concurrent works, we learn CoT primarily through supervised fine-tuning. Alternative paradigms, such as reinforcement learning with environmental feedback, are left for future research.

5.2. Future Work

A natural progression is to explore unified multimodal CoT reasoning within a natively unified generation-and-understanding framework. Furthermore, we plan to verify whether scaling laws hold for the proposed method and whether they can serve as a key to achieving breakthroughs in multi-stage, long-horizon navigation tasks. Beyond this, introducing reinforcement learning within the current training framework to further enhance diverse CoT reasoning capabilities also represents a promising direction. Finally, we believe that developing an adaptive VAR scale selection mechanism could be beneficial for unlocking multi-level visual reasoning capabilities.

6. Conclusion

We propose FantasyVLN, the first unified multimodal reasoning framework that integrates diverse CoT reasoning modes without incurring additional inference latency. To prevent visual CoT reasoning from becoming a computational bottleneck within the entire framework, we introduce compact visual CoT, which improves computational efficiency by performing CoT decoding within the VAR latent space. Experiments on the challenging LH-VLN benchmark demonstrate that FantasyVLN significantly improves the success rate of long-horizon, multi-stage navigation tasks compared to existing single-modality CoT reasoning methods. Furthermore, the proposed compact visual CoT achieves a convergence rate over $4\times$ faster than conventional pixel-space methods.

Acknowledgment

This work was supported by the Hainan Provincial Joint Project of Li'an International Education Innovation Pilot Zone (Grant No.624LALH008), BUPT Kunpeng&Ascend Center of Cultivation, NSFC (Grant No.61601042), and the Super Computing Platform of BUPT.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1, 3

[2] Qiguang Chen, Libo Qin, Jinhao Liu, Yue Liao, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. Rbf++: Quantifying and optimizing reasoning boundaries across measurable and unmeasurable capabilities for chain-of-thought reasoning. *arXiv preprint arXiv:2505.13307*, 2025. 1

[3] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024. 3

[4] Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025. 3, 4

[5] Chen Gao, Liankai Jin, Xingyu Peng, Jiazhao Zhang, Yue Deng, Annan Li, He Wang, and Si Liu. Octonav: Towards generalist embodied navigation. *arXiv preprint arXiv:2506.09839*, 2025. 1

[6] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022. 1

[7] Yanjia Huang, Mingyang Wu, Renjie Li, and Zhengzhong Tu. Vista: Generative visual imagination for vision-and-language navigation. *arXiv preprint arXiv:2505.07868*, 2025. 1

[8] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120, 2020. 1, 3

[9] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020. 3

[10] Bingqian Lin, Yunshuang Nie, Khun Loun Zai, Ziming Wei, Mingfei Han, Rongtao Xu, Minzhe Niu, Jianhua Han, Liang Lin, Cewu Lu, et al. Evolvenav: Self-improving embodied reasoning for llm-based vision-language navigation. *arXiv e-prints*, pages arXiv–2506, 2025. 1

[11] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[12] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019. 3

[13] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12078–12088, 2025. 1, 5, 6

[14] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 4

[15] Hanqing Wang, Wei Liang, Luc V Gool, and Wenguan Wang. Towards versatile embodied navigation. *Advances in neural information processing systems*, 35:36858–36874, 2022. 3

[16] Shuo Wang, Yongcai Wang, Wanting Li, Xudong Cai, Yucheng Wang, Maiyue Chen, Kaihui Wang, Zhizhong Su, Deying Li, and Zhaoxin Fan. Aux-think: Exploring reasoning strategies for data-efficient vision-language navigation. *Advances in Neural Information Processing Systems*, 2025. 2, 3, 5

[17] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409, 2024. 8

[18] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Vision-language navigation policy learning and adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4205–4216, 2020. 3

[19] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. 3

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022. 3

[21] Qiaoyun Wu, Xiaoxi Gong, Kai Xu, Dinesh Manocha, Jingxuan Dong, and Jun Wang. Towards target-driven visual navigation in indoor scenes via generative imitation learning. *IEEE Robotics and Automation Letters*, 6(1):175–182, 2020. 3

[22] Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316, 2024. 1

[23] Zifan Xu, Bo Liu, Xuesu Xiao, Anirudh Nair, and Peter Stone. Benchmarking reinforcement learning techniques for autonomous navigation. In *ICRA*, 2023. 3

[24] Xinda Xue, Junjun Hu, Minghua Luo, Xie Shichao, Jintao Chen, Zixun Xie, Quan Kuichen, Guo Wei, Mu Xu, and Zedong Chu. Omninav: A unified framework for prospective exploration and visual-language navigation. *arXiv preprint arXiv:2509.25687*, 2025. 1

[25] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024. 3

[26] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *Robotics: Science and Systems*, 2025. 3

[27] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. 3, 5

[28] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. In *Advances in Neural Information Processing Systems*, 2025. 3, 4

[29] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. 3

[30] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024. 3

[31] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 1, 3, 4, 5

[32] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023. 3

[33] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer, 2024. 1, 3