WebWorld: Web-Simulated World Model

Ting Huang Zeyu Zhang* Hao Tang†
School of Computer Science, Peking University
*Project lead. †Corresponding author: bjdxtanghao@gmail.com.

"There are a lot of good ideas that can be done without necessarily scaling up (computational resources). That is, how do you do planning with a world model if it evolves not in the physical world, but in, let's say, the world of the internet or some sort of world where an action consists of doing a search in a search engine, interrogating a database, or running a simulation to get a system to actually plan a sequence of actions to give the solution to a problem? Then there is the question of hierarchical planning — for example, planning a trip from New York to Paris." — *Yann LeCun*

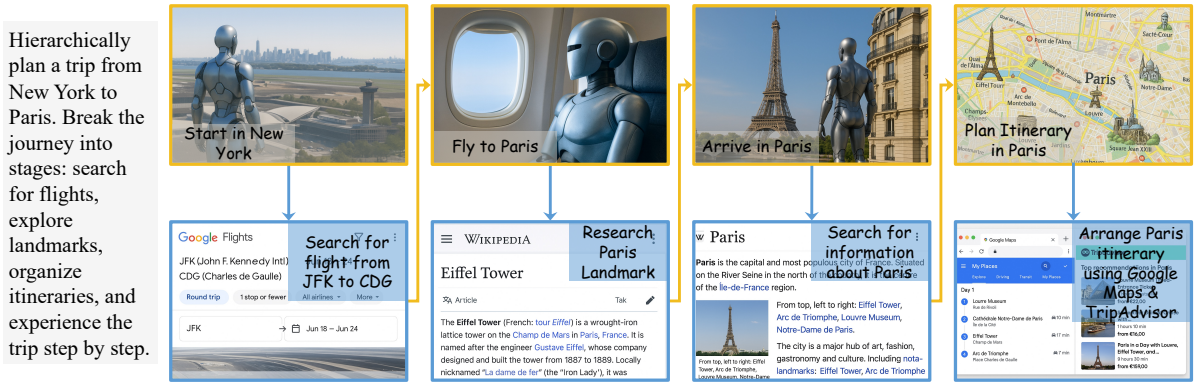


Figure 1: **WebWorld** is a web-simulated world model that addresses complex outdoor reasoning by coupling web navigation, geolocation, and planning within a unified framework.

Abstract

Large language models (LLMs) have recently shown promise as planners across web, spatial, and embodied domains. However, existing agents often handle information seeking, spatial grounding, and action planning as separate capabilities, limiting their ability to solve long-horizon compositional tasks. To address these challenges, we present **WebWorld**, a web-simulated world model for long-horizon reasoning over web evidence, scene observations, and action plans. WebWorld maintains structured intermediate states across heterogeneous modalities and verifies them through schema-guided hierarchical planning. The framework introduces a Mixture-of-Transformers architecture that coordinates website, scene, and motion experts via multimodal self-attention. To support training and evaluation, we develop WebWorld-CoT-55K, a large-scale dataset that integrates heterogeneous modalities into schema-guided reasoning traces, and **WebWorldBench**, a curated benchmark for evaluating long-horizon compositional tasks. Training combines cold-start supervised fine-tuning with offline reinforcement learning based on structural, semantic, and planning rewards; inference uses test-time scaling to aggregate multiple reasoning trajectories. Experiments on WebWorldBench show that WebWorld improves the overall score by up to 8.4% over strong LLM-agent baselines, with gains in outdoor reasoning, geolocation, and travel planning, while still leaving substantial room for progress toward human-level performance. Code: https://github.com/AIGeeksGroup/WebWorld.

1 Introduction

Large language models (LLMs) have recently shown strong potential as planning-oriented agents that reason over heterogeneous information and execute multi-step tasks. On the digital side, web agents enable LLMs to retrieve, navigate, and synthesize knowledge from online environments (Wu et al., 2025; Li et al., 2025b; Wu et al., 2026; Tao

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3592–3608
November 6-10, 2026 ©2026 Association for Computational Linguistics

et al., 2025; Team et al., 2025; Singh et al., 2025). On the embodied side, advances in robotics and embodied AI have extended language-based reasoning to navigation, manipulation, and anomaly detection in physical environments (Huang et al., 2025; Song et al., 2026, 2025a). Complementary to these directions, geolocation benchmarks such as GeoCOT (Song et al., 2025b) and Touchdown (Chen et al., 2019) highlight the importance of spatial inference grounded in realistic visual and geographic signals. Meanwhile, developments in world models (Parker-Holder et al., 2025; Duan et al., 2025; Zhou et al., 2025; Hafner et al., 2025), generalist agents (Reed et al., 2022), and multimodal integration (Driess et al., 2023; Huang et al., 2023) have expanded the design space for linking perception, reasoning, and action.

Despite this progress, these capabilities are often studied and implemented in isolation. Web agents are effective at information seeking but usually lack grounded spatial and embodied context. Embodied agents can operate in physical or simulated environments but often have limited access to external web knowledge. Geolocation systems reason over spatial cues but are typically evaluated in narrower settings that do not require multi-stage web interaction and executable planning. As a result, current agents still struggle with long-horizon compositional tasks that require jointly seeking information, grounding it in scene observations, and converting the resulting plan into action-level decisions.

This gap is particularly evident in web-simulated tasks that combine digital and embodied reasoning. For example, planning a trip from New York to Paris may require searching for flights, identifying landmarks, grounding destinations in map or street-view evidence, organizing an itinerary, and producing step-by-step actions. Such tasks require more than independent web search, visual recognition, or motion prediction. They require structured intermediate states that can be updated and verified across modalities, so that information retrieved from the web can inform spatial grounding and downstream action planning. Figure 1 illustrates this setting, where the agent builds a coherent reasoning trajectory across web, scene, and motion evidence.

To address this challenge, we introduce **WebWorld**, a web-simulated world model for long-horizon reasoning across web navigation, scene understanding, and motion planning. Rather than processing web, scene, and motion signals as isolated inputs, WebWorld organizes them into structured task states that can be updated, grounded, and verified throughout hierarchical reasoning. Concretely, WebWorld consists of four components. First, we propose a Mixture-of-Transformers (MoT) backbone with modality-specific experts for websites, scenes, and motion sequences, fused through multimodal self-attention. Second, we design a schema-guided hierarchical planning strategy that decomposes coarse goals into macro plans, mid-level subgoals, and micro-level operations with verifiable intermediate states. Third, we adopt a two-stage training paradigm, combining cold-start supervised fine-tuning on structured reasoning traces with offline reinforcement learning guided by structural, semantic, and planning rewards. Finally, we incorporate test-time scaling to improve inference robustness by sampling and aggregating multiple reasoning trajectories. To support training and evaluation, we construct two resources. **WebWorld-CoT-55K** is a large-scale dataset that integrates website interactions, scene grounding, and motion demonstrations into schema-guided reasoning traces, covering outdoor reasoning, geolocation inference, hierarchical planning, and embodied action. **WebWorldBench** is a curated benchmark derived from the same data engine but split into disjoint training, validation, and held-out test subsets for controlled evaluation of long-horizon compositional tasks. We further apply schema validation, grounding checks, action-consistency filtering, and human auditing to improve data quality and reduce unsupported reasoning traces. Experiments on WebWorldBench show that WebWorld improves over strong LLM-agent baselines on outdoor reasoning, geolocation, and travel planning, with ablations validating the contributions of the main design components. The key contributions of this paper are summarized as follows:

- We propose **WebWorld**, a web-simulated task-state model with a typed and verifiable cross-modal state interface that organizes web evidence, scene observations, and action states into shared intermediate states that can be consumed, checked, and propagated across long-horizon reasoning steps.
- We design a unified framework that combines a Mixture-of-Transformers backbone with website, scene, and motion experts, schema-guided hierarchical planning, cold-start supervised fine-tuning, offline reinforcement learn-

ing with structural, semantic, and planning rewards, and test-time scaling.

- We construct **WebWorld-CoT-55K**, a large-scale dataset of schema-guided multimodal reasoning traces, and **WebWorldBench**, a curated benchmark for long-horizon compositional reasoning. Experiments show that WebWorld improves over strong LLM-agent baselines, with ablations validating the main design choices.

2 Related Work

World models. World models support prediction and planning by modeling environment dynamics or task-relevant state transitions. Recent work spans physics-informed modeling such as PARC (Xu et al., 2025), representation-based zero-shot planning such as DINO-WM (Zhou et al., 2025), and large-scale generative environments such as Genie 3 (Parker-Holder et al., 2025). Related efforts extend world modeling to embodied navigation (Bar et al., 2025), intrinsically motivated exploration (Lidayan et al., 2025), and interactive multimodal generation (Duan et al., 2025). WebWorld complements these directions by studying web-simulated task states over retrieved evidence, grounded scene observations, and motion plans for long-horizon cross-domain reasoning.

LLM planning. Large language models have shown increasing potential for multi-step planning and structured reasoning. Test-time scaling methods (Muennighoff et al., 2025) improve reasoning by allocating additional inference computation, while hierarchical planning approaches (Puerta-Merino et al., 2025) decompose complex goals into intermediate steps. In embodied settings, 3D-R1 (Huang et al., 2025) improves spatial reasoning for 3D scene understanding, and ManiPLVM-R1 (Song et al., 2026) applies reinforcement learning to embodied manipulation. Other studies explore LLMs for proactive anomaly detection (Song et al., 2025a) and open-ended cooperative decision making (Ma et al., 2025). These works demonstrate the value of structured reasoning for planning-oriented agents, but they typically focus on a single environment or task family. In contrast, WebWorld studies long-horizon reasoning that jointly involves web navigation, scene grounding, and motion planning.

Web agents. LLM-powered web agents aim to navigate, retrieve, and synthesize information from online environments. WebWalker (Wu et al., 2025) studies web traversal and generalization, while WebSailor (Li et al., 2025b), WebDancer (Wu et al., 2026), and WebShaper (Tao et al., 2025) explore complex web navigation, autonomous information seeking, and agentic data synthesis. Related work further connects web interaction with embodied reasoning (Hong et al., 2026), geolocation from human gameplay traces (Song et al., 2025b), and long-horizon web exploration (Team et al., 2025). While these efforts provide important foundations, they typically address web interaction, embodied grounding, or geolocation separately. WebWorld instead studies schema-guided long-horizon reasoning that integrates web evidence, scene observations, and motion plans.

3 Datasets

3.1 Data Engine

To construct **WebWorld-CoT-55K**, we design a data engine that converts heterogeneous web, scene, and motion signals into schema-guided reasoning traces. It integrates website interactions and scene grounding from Embodied Web Agents (Hong et al., 2026), geolocation and spatial reasoning traces from GeoCOT (Song et al., 2025b), and motion sequences from Mixamo (Adobe Systems Incorporated, 2026). Each example is normalized into a hierarchical schema with macro-level intent, mid-level subgoals, micro-level operations, supporting evidence, and action states. Candidate traces are generated by instruction-tuned language models, but the final corpus is filtered rather than directly accepted from raw outputs. The engine controls task templates, decomposition rules, modality-specific evidence fields, and filtering criteria to ensure that web evidence, scene observations, and motion/action plans are connected within a consistent reasoning structure. The resulting dataset covers outdoor reasoning, geolocation inference, hierarchical travel planning, and embodied action prediction.

3.2 WebWorld-CoT-55K

WebWorld-CoT-55K is a large-scale training dataset containing approximately 55K schema-guided multimodal reasoning traces. As summarized in Tab. 1, it combines website interactions, scene observations, and motion demonstrations across outdoor reasoning, geolocation, planning, and action prediction. Each sample includes a

Table 1: **Comparison of datasets.** Compared with prior resources that cover partial modalities or task types, **WebWorld-CoT-55K** integrates website interactions, scene grounding, and motion demonstrations into schema-guided reasoning traces for outdoor reasoning, geolocation, planning, and action prediction.

Dataset	Data Source			Task Type				Number
	Website	Scene	Motion	Outdoor	Geolocation	Planning	Action	
Embodied Web Agent (Hong et al., 2026)	✓	✓	✗	✓	✗	✓	✗	1.5K
GeoCOT (Song et al., 2025b)	✗	✓	✗	✗	✓	✗	✗	3,954K
Mixamo Motion (Adobe Systems Incorporated, 2026)	✗	✗	✓	✗	✗	✗	✓	2.5K
WebWorld-CoT-55K (Ours)	✓	✓	✓	✓	✓	✓	✓	55K

task instruction, modality-specific evidence, a hierarchical reasoning trace, and structured output fields. The trace follows the same schema used by our planner, enabling the model to learn grounded and verifiable intermediate states for long-horizon reasoning. We use WebWorld-CoT-55K as the main supervised fine-tuning resource for WebWorld, while final evaluation is conducted on disjoint WebWorldBench splits. We apply multi-stage quality control and human validation to reduce unsupported reasoning traces, with details provided in *App.* D.

3.3 WebWorldBench

To enable controlled evaluation, we construct **WebWorldBench**, a curated benchmark for long-horizon compositional reasoning. Although WebWorldBench is derived from the same data engine as WebWorld-CoT-55K, it is split before training into disjoint training, validation, and held-out test subsets. The reinforcement learning stage uses only the training split, hyperparameters are selected on the validation split, and all main results are reported on the held-out test split. To reduce potential leakage, we ensure that no identical task instances are shared between training and test splits, and further control overlap in routes, locations, query templates, web evidence, and motion trajectories. The held-out test split is fixed before model training and is not used for reward tuning, model selection, or prompt development. WebWorldBench evaluates tasks that integrate web-based information seeking, geolocation inference, scene grounding, and action planning. Further details are provided in *App.* C.

4 Method

4.1 Overview

Figure 2 illustrates the overall architecture of WebWorld. The framework organizes web evidence, scene observations, and motion plans into structured task states that can be updated and verified during long-horizon reasoning. It consists of four components: (i) a Mixture-of-Transformers (MoT) backbone with modality-specific experts for website, scene, and motion inputs; (ii) schema-guided hierarchical planning that decomposes a coarse goal into macro plans, mid-level subgoals, and micro-level operations; (iii) a two-stage training pipeline with cold-start supervised fine-tuning followed by offline reinforcement learning using structural, semantic, and planning rewards; and (iv) test-time scaling that samples and aggregates multiple reasoning trajectories. Together, these components allow WebWorld to integrate information seeking, spatial grounding, and motion planning in a unified reasoning process.

4.2 WebWorld MoT

The backbone of WebWorld is a Mixture-of-Transformers (MoT) (Liang et al., 2025) that coordinates modality-specific experts through shared multi-modal self-attention. We instantiate three experts. The *Web Expert* encodes textual and structural signals from websites. The *Scene Expert* integrates 2D images, depth maps, and 3D scene representations. The *Motion Expert* processes temporal motion sequences and action-state information. Each expert contains modality-specific transformer layers with feed-forward networks and query-key-value projections. Their outputs are fused by a shared multi-modal self-attention layer, enabling information exchange across web, scene, and motion modalities. On top of the fused representation, we attach task-specific decoders for natural-language response generation, 3D localization, and motion planning. This design allows WebWorld to produce both textual reasoning outputs and structured predictions required by downstream planning tasks.

4.3 Schema-Guided Hierarchical Planning

We operationalize planning as a schema-guided reasoning process rather than free-form chain-of-thought generation. Given a user goal, the planner first produces a macro-level plan, then refines it into mid-level subgoals and micro-level operations. Each reasoning step is constrained by a planning schema specifying: (i) the intent at the current ab-

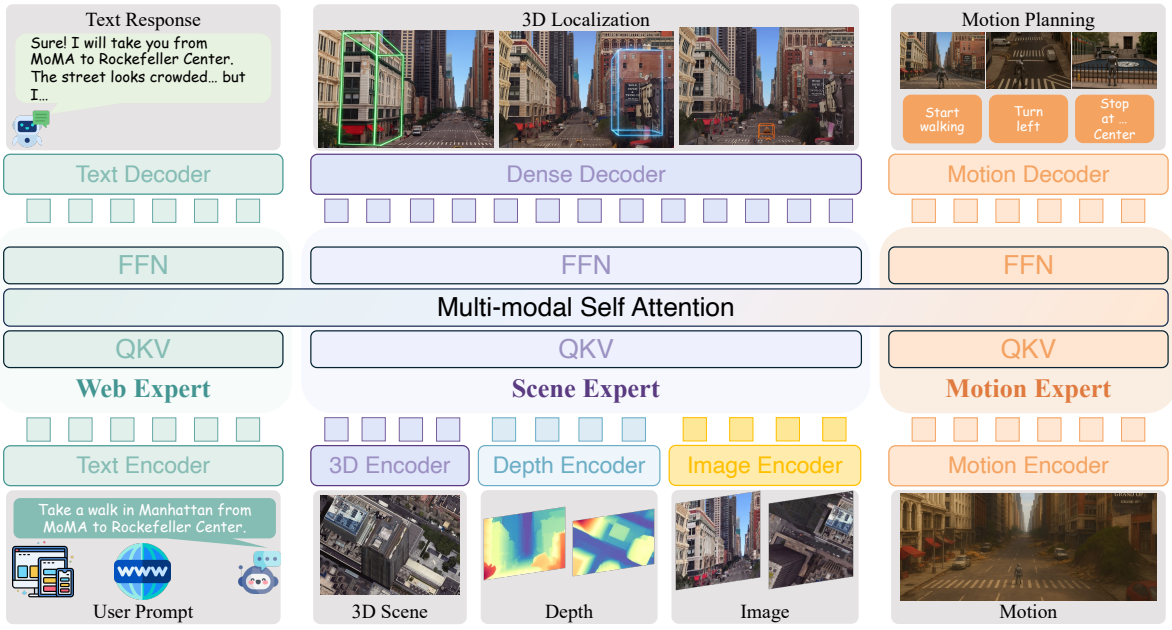


Figure 2: **Architecture of WebWorld.** WebWorld adopts a Mixture-of-Transformers design with website, scene, and motion experts. The experts encode web inputs, 2D and 3D scene signals, and motion sequences, and are fused through multi-modal self-attention. The model outputs text responses, localization predictions, and motion plans within a schema-guided hierarchical reasoning process.

straction level, (ii) admissible operation types tied to the available modalities, (iii) required grounding evidence, (iv) a verifiable intermediate state, and (v) termination or re-planning conditions. Concretely, each step must state its intent, cite or request modality-aligned evidence, and emit a structured intermediate state that can be consumed by subsequent steps. Verification checkpoints are inserted between abstraction levels so that invalid intermediate states can trigger localized repair. Compared with linear CoT, this schema constrains both the format and content of the reasoning trajectory, reducing drift and making intermediate states easier to inspect and evaluate.

4.4 Cold-Start Supervised Fine-Tuning

Training begins with cold-start supervised fine-tuning on WebWorld-CoT-55K. This stage aligns the modality experts and decoders with schema-guided reasoning traces. Given an input x and a reference hierarchical trace y^*, the model is optimized with the standard autoregressive objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|y^*|} \log \pi_\theta(y_t^* \mid x, y_{<t}^*). \tag{1}$$

The SFT stage teaches the model to follow the hierarchical schema, ground intermediate states in modality-specific evidence, and produce structured output fields before reinforcement learning.

4.5 Reinforcement Learning with Exploration

After SFT, we perform an offline reinforcement learning stage to further improve schema adherence, semantic alignment, and action-level consistency. Importantly, this stage uses only the training split of WebWorldBench; validation and held-out test splits are not used for reward tuning or policy optimization. For each input x, the policy samples a group of candidate reasoning trajectories $\{\hat{y}_g\}_{g=1}^{G}$. Each trajectory is scored by a weighted reward:

$$R(\hat{y}_g) = \lambda_{\text{struct}} R_{\text{struct}} + \lambda_{\text{sem}} R_{\text{sem}} + \lambda_{\text{plan}} R_{\text{plan}}. \tag{2}$$

Structural reward. The structural reward measures whether the generated trace follows the required hierarchical schema:

$$R_{\text{struct}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\text{step}_i \in \mathcal{S}\right], \tag{3}$$

where $\mathcal{S}$ denotes the set of valid schema constraints, including intent, operation type, supporting evidence, intermediate state, and termination condition.

Semantic reward. The semantic reward measures whether the generated intermediate states are semantically aligned with the reference trace. We compute it using cosine similarity between frozen sentence embeddings:

$$R_{\text{sem}} = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \cos\left(\text{norm}(E(\hat{s}_t)), \text{norm}(E(s_t^*))\right), \tag{4}$$

where $\hat{s}_t$ and s_t^* are the predicted and reference intermediate states, and $E(\cdot)$ is a frozen text embedding encoder. We report the exact encoder, pooling strategy, normalization rule, and reward weights in the implementation details.

Planning reward. The planning reward evaluates the consistency between predicted and reference action sequences. Both sequences are converted into normalized action-state spans, such as `move_forward`, `turn_left`, `turn_right`, `approach_landmark`, and `stop`. We then compute:

$$R_{\text{plan}} = \frac{|A_{\text{pred}} \cap A_{\text{ref}}|}{|A_{\text{pred}} \cup A_{\text{ref}}|}, \tag{5}$$

where A_{pred} and A_{ref} denote the predicted and reference action-state spans.

Optimization. We optimize the policy with a GRPO-style group-relative objective. For each input, a group of candidate trajectories is sampled and scored by the reward in Eq. 2. The rewards are normalized within the group to obtain relative advantages. The SFT model serves as the reference policy, and a KL penalty regularizes the policy update:

$$\begin{aligned}\mathcal{L}_{\text{RL}} = -\mathbb{E}_{x,\hat{y}_g}\Big[\hat{A}_g \log \pi_\theta(\hat{y}_g \mid x) \\ - \beta D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\Big],\end{aligned} \tag{6}$$

where $\hat{A}_g$ is the normalized group-relative advantage, π_{ref} is the SFT reference policy, β is the KL coefficient, and $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ is computed under the same input x. This offline RL stage encourages trajectories that are schema-valid, semantically grounded, and consistent with the reference action-state sequence.

4.6 Test-Time Scaling

At inference time, WebWorld uses test-time scaling to improve robustness without updating model parameters. Instead of relying on a single reasoning trajectory, the model samples K candidate hierarchical traces with controlled decoding parameters. Each candidate produces intermediate states and final outputs under the same planning schema. The final answer is selected by majority voting for discrete predictions or by semantic consistency scoring for open-ended responses.

This procedure is complementary to schema-guided planning. The schema constrains candidate trajectories to follow comparable intermediate states, while sampling provides diversity for difficult long-horizon tasks. In practice, test-time scaling trades additional inference cost for improved reasoning robustness; we analyze this accuracy–cost trade-off in the experiments.

5 Experiments

5.1 Baselines

We compare WebWorld with strong LLM-agent and multimodal baselines, including GPT-5 (Singh et al., 2025), Gemini 2.5 Pro (Comanici et al., 2025), Qwen-VL 2.5 (Bai et al., 2025), InternVL 3.0 (Zhu et al., 2025), and Embodied Web Agent (Hong et al., 2026). Embodied Web Agent is the closest prior system that connects website interactions with embodied reasoning. For proprietary models, we use the corresponding API versions available during evaluation. For open-source models, we report the exact checkpoints, parameter scales, input modalities, and decoding settings in *App.* B. All prompting-based baselines are evaluated with the same task schema and input evidence when supported by the model interface. We also report human performance as an upper-bound reference rather than a solved-task target.

5.2 Evaluation Metrics

We evaluate models across web-simulated outdoor reasoning, geolocation, hierarchical planning, and action prediction. For the main benchmark, we report **Outdoor Accuracy** (Out. Acc) for high-level outdoor task correctness, **Geo@City** for city-level geolocation accuracy, **Planning Success Rate** (Plan SR) for valid hierarchical planning trajectories, and **Action IoU** (Act IoU) for action-sequence quality. The **Overall** score is a weighted aggregate of these metrics.

For outdoor planning tasks, we additionally report Overall Accuracy, Web-only Accuracy, Embodied-only Accuracy, and Overall Completion Rate. These metrics separate digital information seeking from embodied state completion and measure both final success and partial task progress. For geolocation, we report country-, city-, and street-level accuracy.

Action IoU. Action IoU is computed over normalized action-state spans rather than raw natural-language descriptions. Both model predictions and reference traces are converted into the same discrete action vocabulary, including `move_forward`, `turn_left`, `turn_right`, `approach_landmark`,

Table 2: **Results on the test split of WEBWORLDBENCH.** We report outdoor task accuracy (**Out. Acc**), city-level geolocation accuracy (**Geo@City**), hierarchical planning success rate (**Plan SR**), action sequence quality (**Act IoU**), and the weighted aggregate score (**Overall**). Human performance is included as an upper-bound reference.

Model	Out. Acc ↑	Geo@City ↑	Plan SR ↑	Act IoU ↑	Overall ↑
Human	91.6	85.2	95.0	80.0	88.4
GPT-5 (Singh et al., 2025)	30.4	73.2	25.0	18.0	40.2
Gemini 2.5 Pro (Comanici et al., 2025)	26.5	78.2	27.0	20.0	39.7
Qwen-VL 2.5 (Bai et al., 2025)	13.9	45.1	18.0	12.0	23.6
InternVL 3.0 (Zhu et al., 2025)	11.0	52.1	20.0	14.0	25.3
Embodied Web Agent (Hong et al., 2026)	24.0	60.0	22.0	16.0	33.2
WebWorld (Ours)	**40.8**	**81.3**	**33.5**	**26.9**	**48.6**

and `stop`. We then compute

$$\mathrm{ActIoU} = \frac{|A_{\mathrm{pred}} \cap A_{\mathrm{ref}}|}{|A_{\mathrm{pred}} \cup A_{\mathrm{ref}}|}, \tag{7}$$

where A_{pred} and A_{ref} denote predicted and reference action-state spans. The human Act IoU in Tab. 2 is not a self-comparison; it is computed by comparing one annotator's action trace against the aggregated reference trace from other annotators, reflecting inter-human consistency under the same action abstraction.

5.3 Implementation Details

Our Mixture-of-Transformers backbone instantiates three modality-specific experts. The Web Expert is initialized from WebSailor v2 (Li et al., 2025a). The Scene Expert follows the geospatial scene understanding architecture in (Froehlich et al., 2025), which integrates 2D images, depth maps, and 3D scene representations. The Motion Expert is initialized from MOGO (Fu et al., 2025), which represents motion sequences with residual quantized hierarchical tokens. The experts are fused through shared multi-modal self-attention, with task-specific decoders for natural-language reasoning, localization, and action prediction. Exact checkpoint names, parameter scales, and trainable components are provided in *App.* B.

Training proceeds in two stages. We first conduct supervised fine-tuning on WebWorld-CoT-55K to learn schema-guided hierarchical reasoning across modalities. We then perform offline reinforcement learning only on the training split of WebWorldBench. The validation split is used for hyperparameter selection, and the held-out test split is used only for final evaluation. Unless otherwise specified, we use AdamW with a peak learning rate of 2×10^{-5}, 10% linear warm-up, a global batch size of 8, and 4 NVIDIA H20 GPUs. During reinforcement learning, we use a KL coefficient of $\beta = 0.05$.

At inference time, WebWorld samples up to 16 hierarchical reasoning traces with a maximum length of 256 tokens and aggregates predictions by semantic consistency scoring. For prompting-based baselines, we provide the same task schema and input evidence when supported by the interface; these baselines are not trained on WebWorld-CoT-55K unless explicitly stated in matched-data experiments.

5.4 Result Analysis

Outdoor planning. Tab. 3 reports results on navigation, shopping, and traveling tasks. WebWorld achieves the best performance across all task types and metrics, improving overall accuracy to 41.85% on navigation, 31.02% on shopping, and 37.51% on traveling. Web-only accuracy is consistently higher than embodied-only accuracy, suggesting that current agents handle digital information seeking more reliably than embodied state completion. Completion rates are also higher than final accuracies, indicating that models often make partial progress but fail to complete all cross-domain requirements. Shopping and traveling remain more difficult than navigation because they require longer interactions between web evidence, scene grounding, and action-level planning.

Geolocation. Tab. 4 reports country-, city-, and street-level geolocation accuracy. WebWorld achieves the best performance among model-based agents, reaching 88.3% country accuracy, 80.2% city accuracy, and 8.5% street accuracy. The gains are most visible at city and street levels, where the model must combine visual-spatial cues with retrieved web evidence. However, street-level accuracy remains low for all systems, including humans, showing that fine-grained geolocation remains challenging.

Overall benchmark. Tab. 2 summarizes performance on WebWorldBench. WebWorld obtains the best Overall score of 48.6, compared with 40.2 for GPT-5, 39.7 for Gemini 2.5 Pro, and 33.2 for

Table 3: **Outdoor planning results.** We report final success, completion rate, web-only accuracy, and embodied-only accuracy across navigation, shopping, and traveling tasks. Human performance is included as an upper-bound reference.

Task	Metric	Human	GPT (Singh et al., 2025)	Gemini (Comanici et al., 2025)	Qwen (Bai et al., 2025)	Intern (Zhu et al., 2025)	WebWorld (Ours)
Navigation	Overall Accuracy	90.28	34.72	30.56	15.97	13.19	**41.85**
	Overall Completion Rate	91.32	52.08	48.96	36.81	26.04	**56.33**
	Web-only Accuracy	92.36	69.44	67.36	57.64	38.89	**74.31**
	Embodied-only Accuracy	90.97	48.61	46.53	31.25	23.61	**54.17**
Shopping	Overall Accuracy	92.59	25.46	23.61	13.89	10.65	**31.02**
	Overall Completion Rate	93.52	31.94	30.56	18.52	14.35	**36.53**
	Web-only Accuracy	93.06	39.35	37.50	23.15	17.13	**44.45**
	Embodied-only Accuracy	93.98	34.26	32.41	17.59	12.96	**39.81**
Traveling	Overall Accuracy	91.82	30.91	25.45	11.82	9.09	**37.51**
	Overall Completion Rate	93.64	50.91	48.18	34.55	20.91	**55.82**
	Web-only Accuracy	94.55	57.27	53.64	41.82	25.45	**63.67**
	Embodied-only Accuracy	92.73	47.27	44.55	29.09	19.09	**52.27**

Table 4: **Geolocation results.** We report country-, city-, and street-level localization accuracy on the held-out test split.

Model	Country	City	Street
Human	90.0	85.0	12.0
GPT-5 (Singh et al., 2025)	80.9	75.2	4.5
Gemini 2.5 Pro (Comanici et al., 2025)	75.4	63.2	3.2
Qwen-VL 2.5 (Bai et al., 2025)	69.0	49.3	1.0
InternVL 3.0 (Zhu et al., 2025)	77.3	57.5	2.1
Embodied Web Agent (Hong et al., 2026)	68.0	60.0	3.5
WebWorld (Ours)	**88.3**	**80.2**	**8.5**

Embodied Web Agent. The improvements are consistent across outdoor reasoning, Geo@City, Plan SR, and Act IoU, suggesting that schema-guided planning and multimodal expert integration are useful for long-horizon compositional reasoning. At the same time, the human upper-bound score of 88.4 indicates that WebWorldBench remains far from solved.

5.5 Controlled Comparisons and Additional Analysis

To separate the effect of framework design from the effect of additional task-specific data, we conduct matched-data and data-scaling comparisons. As shown in Tab. 5, under the same 1.5K training budget as Embodied Web Agent, WebWorld improves the Overall score from 33.2 to 38.0. Scaling the training data from 1.5K to 55K further improves the Overall score from 38.0 to 48.6, indicating that WebWorld benefits from both the proposed framework and the WebWorld-CoT data engine.

Table 5: **Matched-data and data-scaling comparison.** Under the same 1.5K training budget, WebWorld outperforms Embodied Web Agent. Increasing the amount of WebWorld-CoT training data further improves performance.

Model / Data	Out. Acc.	Geo@City	Plan SR	Act IoU	Overall
Embodied Web Agent, 1.5K	24.0	60.0	22.0	16.0	33.2
WebWorld, 1.5K	29.6	67.8	25.4	18.9	38.0
WebWorld, 5K	33.1	72.4	28.1	21.5	41.7
WebWorld, 15K	36.7	76.9	30.8	24.2	45.1
WebWorld, 55K	40.8	81.3	33.5	26.9	48.6

Beyond the matched-data and scaling results, we provide additional analyses in the appendix. *App.* A reports ablations on modality experts, hierarchical planning, reward design, and test-time scaling, including a component-level analysis of the planning schema.

5.6 Long-Horizon and Controlled Analysis

To further examine where WebWorld gains its advantage, we analyze performance with respect to executable trajectory length and inference-time scaling. We define the reasoning horizon by the number of executable steps rather than the length of the generated text.

Performance across trajectory lengths. We stratify the test set into short (1–3 steps), medium (4–6 steps), and long ($\geq$7 steps) trajectories. As shown in Tab. 6, WebWorld consistently outperforms Linear CoT and two-level planning, with the improvement becoming larger as the executable horizon increases. Compared with Linear CoT, the gain increases from $+6.4$ points on short trajectories to $+11.9$ points on medium trajectories and $+16.5$ points on long trajectories. This trend suggests that maintaining and verifying structured intermediate states becomes increasingly useful as dependencies extend across more execution steps.

Table 6: **Performance across executable trajectory lengths.** WebWorld shows increasingly larger gains as the reasoning horizon grows.

Trajectory Length	Linear CoT	Two-level Planning	WebWorld
Short (1–3 steps)	34.8	38.5	**41.2**
Medium (4–6 steps)	36.1	42.0	**48.0**
Long ($\geq$7 steps)	36.6	45.3	**53.1**

Effect of test-time scaling. We further vary the number of sampled reasoning chains K to separate learned model capability from inference-time aggregation. As shown in Tab. 7, WebWorld achieves an Overall score of 44.0 with a single reasoning chain, already exceeding GPT-5 (40.2) and Gemini 2.5 Pro (39.7) under our benchmark setting. Increasing K from 1 to 16 further improves the Overall score by 4.6 points, from 44.0 to 48.6, while exhibiting diminishing returns. These results indicate that test-time scaling improves robustness,

Table 7: **Effect of the number of reasoning chains K.** The single-chain model remains strong, while additional chains provide incremental improvements.

K	Overall	Out. Acc	Geo@City	Plan SR	Act IoU
1	44.0	36.9	77.6	29.4	23.1
4	46.1	38.5	79.2	31.0	24.8
8	47.4	39.6	80.4	32.2	25.8
16	**48.6**	**40.8**	**81.3**	**33.5**	**26.9**

but does not solely account for the performance advantage of WebWorld.

Robustness to perturbed web states. WebWorld-Bench uses controlled cached web states for reproducibility. To evaluate robustness and directly examine cross-modal information flow, we perturb only the retrieved web evidence while keeping the scene and motion/action inputs fixed. We consider three controlled perturbations: (i) removing or corrupting a subset of retrieved links, (ii) injecting irrelevant search results or distractor pages, and (iii) shuffling the order of retrieved evidence. The model and evaluation protocol are otherwise unchanged.

Table 8: **Robustness to perturbed web states.** Only web evidence is perturbed while the remaining inputs are kept fixed.

Setting	Out. Acc.	Geo@City	Plan SR	Overall
Original cached web	**40.8**	**81.3**	**33.5**	**48.6**
Broken / missing links	38.9	79.6	31.8	46.7
Distractor web evidence	37.8	78.4	30.9	45.9
Shuffled search results	39.4	80.1	32.2	47.3

As shown in Tab. 8, performance degrades under perturbations but remains relatively stable. The largest drop occurs with distractor web evidence, where the Overall score decreases from 48.6 to 45.9. Notably, Geo@City and Plan SR also decrease when only web evidence is perturbed, indicating that retrieved web information affects subsequent spatial grounding and planning rather than being processed independently. These results do not replace a full live-web evaluation, but provide an initial robustness and cross-modal intervention analysis under controlled web-state changes.

Out-of-domain transfer. We further evaluate WebWorld on compatible out-of-domain settings to test whether the learned schema-guided reasoning transfers beyond WebWorldBench. Because external benchmarks do not always provide all three modalities used by WebWorld, we adapt the interface by disabling unavailable expert inputs and evaluating only compatible task subsets. Therefore, these results should be interpreted as transfer under interface adaptation rather than direct optimization for each benchmark.

As shown in Tab. 9, WebWorld consistently improves over the compared baselines on the evaluated subsets. These results suggest that schema-guided reasoning and multimodal state organization transfer beyond WebWorldBench, although evaluation in fully live and broader environments remains necessary.

Table 9: **Out-of-domain transfer.** Results on compatible external benchmark subsets under interface adaptation.

Benchmark	Metric	EWA	GPT-5	WebWorld
Embodied Web Agents	Success Rate	22.0	25.0	**30.6**
Embodied Web Agents	Completion Rate	38.5	42.1	**47.8**
NaviMaster-style GUI Nav	Success Rate	–	28.4	**33.2**
NaviMaster-style GUI Nav	Step Efficiency	–	41.7	**46.5**

6 Conclusion

We introduced WebWorld, a web-simulated world model for long-horizon reasoning across web navigation, scene understanding, and motion planning. By organizing web evidence, scene observations, and motion plans into structured task states, WebWorld supports schema-guided reasoning over information seeking, spatial grounding, and action-level planning. We further constructed WebWorld-CoT-55K and WebWorldBench to support training and controlled evaluation in this setting. Experiments show that WebWorld improves over strong LLM-agent baselines on outdoor reasoning, geolocation, and travel planning, with ablations validating the main design choices. Future work should further address the remaining gap to human performance and extend the framework toward more dynamic web and real-world environments.

Limitations

Our study has several limitations. First, WebWorld-Bench is built in a controlled web-simulated environment, which improves reproducibility but cannot fully capture live-web dynamics such as broken links, pop-ups, layout shifts, or changing search results. Second, our evaluation focuses on outdoor reasoning, geolocation, travel planning, and action prediction, leaving broader indoor, safety-critical, multi-agent, and real-time interaction settings for future work. Finally, WebWorld-CoT-55K relies on generated and curated traces even after quality control, and hierarchical planning with test-time scaling introduces extra inference cost.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities, Peking University.

References

Adobe Systems Incorporated. 2026. Mixamo Motion Capture Library. https://www.mixamo.com/.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. 2025. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*.

Yinglin Duan, Zhengxia Zou, Tongwei Gu, Wei Jia, Zhan Zhao, Luyi Xu, Xinzhu Liu, Yenan Lin, Hao Jiang, Kang Chen, and 1 others. 2025. Latticeworld: A multimodal large language model-empowered framework for interactive complex world generation. *arXiv preprint arXiv:2509.05263*.

Jon E Froehlich, Jared Hwang, Zeyu Wang, John S O'Meara, Xia Su, William Huang, Yang Zhang, Alex Fiannaca, Philip Nelson, and Shaun Kane. 2025. "does the cafe entrance look accessible? where is the door?" towards geospatial ai agents for visual inquiries. *arXiv preprint arXiv:2508.15752*.

Dongjie Fu, Tengjiao Sun, Pengcheng Fang, Xiaohao Cai, and Hansung Kim. 2025. Mogo: Residual quantized hierarchical causal transformer for high-quality and real-time 3d human motion generation. *arXiv preprint arXiv:2506.05952*.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2025. Mastering diverse control tasks through world models. *Nature*, pages 1–7.

Yining Hong, Rui Sun, Bingxuan Li, Xingcheng Yao, Maxine Wu, Alexander Chien, Da Yin, Ying Nian Wu, Zhecan Wang, and Kai-Wei Chang. 2026. Embodied web agents: Bridging physical-digital realms for integrated agent intelligence. *Advances in Neural Information Processing Systems*, 38.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ting Huang, Zeyu Zhang, and Hao Tang. 2025. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, and 1 others. 2025a. Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning. *arXiv preprint arXiv:2509.13305*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, and 1 others. 2025b. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. 2025. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*.

Aly Lidayan, Yuqing Du, Eliza Kosoy, Maria Rufova, Pieter Abbeel, and Alison Gopnik. 2025. Intrinsically-motivated humans and agents in open-world exploration. *arXiv preprint arXiv:2503.23631*.

Rachel Ma, Jingyi Qu, Andreea Bobu, and Dylan Hadfield-Menell. 2025. Open-universe assistance games. In *First Workshop on Multi-Turn Interactions in Large Language Models*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.

Jack Parker-Holder, Shlomi Fruchter, and 1 others. 2025. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4.

Israel Puerta-Merino, Carlos Núñez-Molina, Pablo Mesejo, and Juan Fernández-Olivares. 2025. A roadmap to guide the integration of llms in hierarchical planning. *arXiv preprint arXiv:2501.08068*.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, and 1 others. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Zirui Song, Guangxian Ouyang, Meng Fang, Hongbin Na, Zijing Shi, Zhenhao Chen, Fu Yujie, Zeyu Zhang, Shiyu Jiang, Miao Fang, and 1 others. 2025a. Hazards in daily life? enabling robots to proactively detect and resolve anomalies. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 7399–7415.

Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Fengxian Ji, Zhenhao Chen, Zhongzhi Li, and Xiuying Chen. 2026. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40:18558–18566.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. 2025b. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *arXiv preprint arXiv:2502.13759*.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, and 1 others. 2025. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, DingChu Zhang, Zekun Xi, Robert Tang, and 1 others. 2026. Webdancer: Towards autonomous information seeking agency. *Advances in Neural Information Processing Systems*, 38:120957–120985.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and 1 others. 2025. Webwalker: Benchmarking llms in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 10290–10305.

Michael Xu, Yi Shi, KangKang Yin, and Xue Bin Peng. 2025. Parc: Physics-based augmentation with reinforcement learning for character controllers. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. 2025. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Forty-second International Conference on Machine Learning*.

Jinguo Zhu and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Ablation Study

Mixture-of-Transformers. We ablate the modality experts in the MoT backbone by enabling different combinations of Web, Scene, and Motion experts. As shown in Tab. 10, single-expert variants perform substantially worse because each modality captures only part of the task signal. Combining two experts improves performance, and using all three experts achieves the best Overall score of 48.6. These results indicate that web evidence, scene observations, and motion signals provide complementary information for long-horizon reasoning.

Table 10: **Effect of modality experts.** We enable different combinations of Web, Scene, and Motion experts in the MoT backbone. The full configuration achieves the best performance, indicating complementary contributions from the three modalities.

Expert Setting			Out. Acc↑	Geo@City↑	Act IoU↑	Overall↑
Web	Scene	Motion				
✓	✗	✗	28.4	72.1	14.8	34.0
✗	✓	✗	24.6	65.3	16.2	32.1
✗	✗	✓	20.9	60.4	12.5	28.3
✓	✓	✗	31.5	76.8	20.4	39.1
✓	✗	✓	29.7	74.5	19.2	37.4
✗	✓	✓	26.8	68.9	18.0	34.7
✓	✓	✓	**40.8**	**81.3**	**26.9**	**48.6**

Hierarchical planning strategy. We compare linear CoT, two-level planning, and full schema-guided planning. As shown in Tab. 11, linear CoT obtains low Plan SR because it lacks explicit intermediate-state verification. Two-level planning improves both Plan SR and Overall score, while the full schema-guided strategy performs best. This suggests that multi-level decomposition and verification help reduce error propagation in long-horizon tasks.

Table 11: **Effect of Hierarchical Planning.** We compare linear chain-of-thought with different levels of structured decomposition. Multi-level planning with schema guidance consistently improves success rate and overall performance by enforcing checkable intermediate states.

Planning Strategy	Out. Acc↑	Plan SR↑	Geo@City↑	Overall↑
Linear CoT (w/o Planning)	28.4	19.2	73.5	36.0
Two-level Planning (Coarse → Fine)	34.6	26.5	77.9	42.3
Schema-guided Full Planning (Ours)	**40.8**	**33.5**	**81.3**	**48.6**

Component-level planning schema ablation. To further analyze the planning grammar, we remove one schema constraint at a time while keeping the same backbone, training data, and evaluation protocol. As shown in Tab. 12, removing verifiable intermediate states causes the largest drop, reducing the Overall score from 48.6 to 41.6. Removing admissible operation types also substantially hurts performance, indicating that typed operations are important for preventing modality-incompatible steps. The results show that the gains of schema-guided planning come not only from decomposition depth, but also from explicit constraints on evidence, state verification, and re-planning.

Table 12: **Component-level ablation of the planning schema.** Each row removes one schema constraint while keeping the same backbone, training data, and evaluation protocol.

Planning Variant	Out. Acc.	Plan SR	Act IoU	Overall
Full schema-guided planning	**40.8**	**33.5**	**26.9**	**48.6**
w/o intent specification	37.6	30.7	24.8	45.2
w/o admissible operation types	35.8	28.4	22.6	42.8
w/o grounding evidence	36.9	29.5	23.7	44.0
w/o verifiable intermediate state	34.7	27.2	21.9	41.6
w/o termination / re-planning	38.1	30.1	24.4	45.0
Linear CoT w/o planning	28.4	19.2	18.7	36.0

Reward design. We ablate the structural, semantic, and planning rewards used in offline RL. As shown in Tab. 13, SFT alone obtains an Overall score of 39.2. Each individual reward improves performance, and pairwise combinations further increase Plan SR and Overall score. The full reward design performs best, indicating that schema adherence, semantic alignment, and action-state consistency provide complementary supervision signals.

Table 13: **Effect of reward components.** We ablate structural (R_{struct}), semantic (R_{sem}), and planning (R_{plan}) rewards. The full reward combination achieves the best performance.

Reward Configuration			Plan SR↑	Act IoU↑	Out. Acc↑	Overall↑
R_{struct}	R_{sem}	R_{plan}				
✗	✗	✗	24.6	19.1	34.0	39.2
✓	✗	✗	27.0	20.5	35.6	41.0
✗	✓	✗	26.3	20.1	35.0	40.5
✗	✗	✓	25.8	20.8	34.7	40.1
✓	✓	✗	30.2	23.1	37.9	44.7
✓	✗	✓	29.6	22.7	37.4	44.0
✗	✓	✓	29.9	22.9	37.7	44.3
✓	✓	✓	**33.5**	**26.9**	**40.8**	**48.6**

Test-time scaling. We evaluate the effect of sampling multiple reasoning trajectories at inference time. As shown in Figure 3, increasing the number of sampled chains improves Overall accuracy, especially for long-horizon planning, but also increases inference cost. Moderate scaling with 16 chains provides a strong accuracy–cost trade-off in our setting.

B Implementation and Baseline Details

Model initialization. Tab. 14 summarizes the initialization of WebWorld components. The Web Expert is initialized from WebSailor v2, the Scene Expert follows the geospatial scene understanding backbone, and the Motion Expert is initialized from MOGO-style motion modeling weights. The

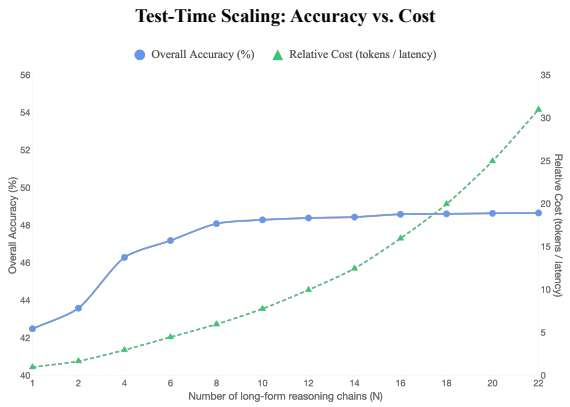


Figure 3: **Effect of test-time scaling.** Sampling more reasoning chains improves Overall accuracy but increases inference cost. In our setting, 16 chains provide a favorable accuracy–cost trade-off.

shared multi-modal self-attention layers and task-specific decoders are trained during WebWorld SFT and offline RL.

Table 14: **Initialization of WebWorld components.** We report the initialization source and whether each component is trainable during SFT/RL.

Component	Initialization	Trainable
Web Expert	WebSailor v2-style checkpoint	Yes
Scene Expert	Geospatial scene backbone	Yes
Motion Expert	MOGO-style checkpoint	Yes
Multi-modal self-attention	Random initialization	Yes
Text / localization / action decoders	Task-head initialization	Yes

Baseline settings. For proprietary models, we use the API versions available during evaluation. For open-source baselines, we use the official checkpoints and evaluate them with the same task schema and input evidence when supported by the model interface. Tab. 15 summarizes the baseline settings.

Table 15: **Baseline settings.** We report model source, input modality, and evaluation protocol for reproducibility.

Model	Source	Inputs	Protocol
GPT-5	API	Text + Image	Prompting
Gemini 2.5 Pro	API	Text + Image	Prompting
Qwen-VL 2.5	Official checkpoint	Text + Image	Prompting
InternVL 3.0	Official checkpoint	Text + Image	Prompting
Embodied Web Agent	Official setting	Web + Scene	Reported / reproduced
WebWorld	Ours	Web + Scene + Motion	SFT + offline RL

C More Details about Data Collection

C.1 Outdoor Data Collection

Task intents and templates. To better organize and summarize outdoor task intents, we design a comprehensive set of task templates, where each template corresponds to several specific tasks across navigation, shopping, travel planning, and Wikipedia query scenarios. These templates are instantiated during data generation by filling placeholders ({ }, {{ }}) with concrete content sampled from our environment. For example, "Show me the fastest route from {origin} to {destination}" becomes "Show me the fastest route from Times Square to Central Park." Figure 4 summarizes representative templates used in WebWorld-CoT-55K.

Task and location generation. For each template, we first prompt a vision-language model to generate diverse task intents and corresponding physical locations. The system samples start and end coordinates from the outdoor environment graph constructed from the collected map and street-view sources. These coordinates are projected onto map and street-view views to ensure plausible paths between origins and destinations. Each generated sample therefore contains both (1) a textual task instruction and (2) a grounded 3D environment reference, forming the foundation for schema-guided reasoning. The prompt to generate location can be seen in section E.1.

Annotation and visualization. We develop a custom graphical annotation tool for interactive visualization and correction of trajectories, enabling manual review of agent plans, motion paths, and retrieved web results. Annotators can inspect both textual CoT traces and spatial trajectories overlaid on 2D maps or 3D environment views. Edits are automatically synchronized to JSON-formatted backend files, ensuring reproducible modification tracking.

C.2 Geolocation and Web Data Collection

Dataset curation. The geolocation subset of WebWorld-CoT-55K is derived from the Breadth portion of FairLocator, containing rich human gameplay traces for spatial reasoning. We randomly sample 150 locations across five continents, followed by manual filtering to retain those with sufficient visual cues - such as building facades, vegetation, and signage - suitable for reasoning-based localization.

Image observations. For each selected coordinate, we collect panoramic image observations via the Google Street View API. Each location includes the ground-truth viewpoint and several adjacent nodes, supporting short-range embodied exploration. We exclude isolated nodes without adjacent viewpoints to maintain continuity during navigation-based reasoning.

Website queries. Web-based reasoning data are collected through the WebSailor v2 environment (Li et al., 2025a), which provides a controllable web-browsing API and dynamic page

Templates for Outdoor Tasks

Navigation:
* Show me the fastest route from {origin} to {destination}.
* Plot a walking path from {origin} to {destination}, avoiding highways.
* Starting at {origin}, guide me step-by-step to {destination}.
* If I leave {origin} now, what's the quickest way to reach {destination}?
* ...

Shopping:
* Add all the items with {{quality}} on this page into my cart.
* Buy me a {{product}} with {{detail}}.
* Compare {{item_A}} and {{item_B}} and recommend one with higher rating.
* What size of {{item}} should I buy if {{condition}}?
* ...

Travel Planning:
* Plan a 3-day trip from {city_A} to {city_B} including hotels and attractions.
* Suggest nearby landmarks when staying at {hotel}.
* Create an itinerary that starts from {location} and covers {landmark}.
* ...

Wikipedia:
* Search for "{query}" and tell me more about it.
* I would like to know more about "{query}".
* Provide me a summary of "{query}" from Wikipedia.
* ...

Figure 4: Templates for outdoor tasks, covering navigation, shopping, travel planning, and Wikipedia queries.

simulation. For each task, the system generates a lightweight query configuration at runtime to evaluate the agent's ability to formulate and execute its own search. The agent primarily interacts with Wikipedia-like domains and open knowledge sources to ensure factual retrieval. All retrieved HTML structures and textual contents are cached as structured documents and subsequently used in hierarchical chain-of-thought reasoning within WebWorld-CoT-55K.

C.3 Motion Data Collection

Motion source and preprocessing. The motion modality in WebWorld-CoT-55K is derived from the Mixamo Motion Capture Library (Adobe Systems Incorporated, 2026), which provides diverse human motion clips such as walking, turning, waving, sitting, and gesturing. We select 2,500 sequences that are visually relevant to outdoor and navigation scenarios. Each motion sequence is represented by 3D joint trajectories with a temporal resolution of 30 FPS and normalized joint positions centered on the pelvis to ensure consistency across characters and scenes.

Tokenization and representation. To integrate motion signals into the Mixture-of-Transformers (MoT) backbone, we adopt a residual quantized hierarchical causal transformer inspired by MOGO (Fu et al., 2025). Each motion clip is first encoded into multi-level discrete tokens using motion-scale adaptive vector quantization. The resulting motion tokens capture coarse-to-fine temporal dynamics, enabling efficient fusion with textual and scene modalities via the shared multimodal self-attention layer. During data construction, we ensure synchronization between motion steps and textual chain-of-thought reasoning (e.g., "Start walking", "Turn left", "Stop at the intersection").

Alignment with scene and web modalities. To align motion with spatial and web signals, we map the start and end poses of each motion sequence to corresponding locations in the outdoor scene mesh. Each action step is paired with a localized 3D coordinate and a natural-language description extracted from the reasoning trace. This yields temporally aligned triplets of {text, scene, motion} at every step. For example, the instruction "walk from MoMA to Rockefeller Center" is grounded to motion frames traversing the corresponding 3D route and linked to retrieved web information about the landmarks.

D Data Quality Control and Human Validation

Because WebWorld-CoT-55K contains generated reasoning traces, we apply a multi-stage quality-control pipeline before constructing the final dataset. The goal is to ensure that each retained example is structurally valid, grounded in available evidence, consistent with the corresponding action trajectory, and free from unsupported entities. Tab. 16 summarizes the filtering statistics and human validation results.

Automatic quality control. We first apply *schema validation* to check whether each generated trace follows the required hierarchical planning format.

Location Generation for Outdoor Tasks

System Prompt:
You are an AI assistant familiar with cities all around the world. For a given city, provide an iconic location suitable for outdoor reasoning tasks. Each entry must include: (1) `location` — a searchable address or latitude-longitude coordinate; (2) `instruction` — a concise description capturing the site's visual or cultural features **without directly naming it**. Ensure that the description is distinctive enough for an expert to infer the location. Return the results as a JSON list, where each element contains `location` and `instruction`. Cities are sampled from: `New York, Paris, Tokyo, London.`

Example:
["location": "48.8584, 2.2945 (Paris, France)", "instruction": "I'm in Paris and would like to visit the wrought-iron tower built for the 1889 Exposition, visible from almost anywhere in the city."]

Notes: This prompt grounds textual hints to geographic coordinates and establishes multi-modal anchors for subsequent web and motion task generation.

Figure 5: LLM prompt for location generation in outdoor tasks, producing landmark descriptions and coordinates without explicitly naming the location.

Specifically, a valid trace must contain macro-level intent, mid-level subgoals, micro-level operations, supporting evidence fields, and structured action states. Examples with missing fields, invalid ordering, or malformed intermediate states are removed.

Second, we conduct *grounding validation*. For each trace, we verify whether the referenced locations, landmarks, retrieved web evidence, and scene observations are supported by the corresponding environment state. Traces are filtered out if they mention unavailable webpages, unsupported landmarks, inconsistent geolocation evidence, or scene elements that cannot be matched to the provided observation.

Third, we perform *action-consistency validation*. This step checks whether the predicted motion or action sequence is compatible with the planned route and intermediate states. Examples are removed if they contain impossible transitions, inconsistent turn or movement commands, missing stop conditions, or action sequences that do not correspond to the planned navigation or interaction goal.

Table 16: **Quality control and human validation for WebWorld-CoT-55K.** Generated traces are filtered through schema, grounding, action-consistency, and hallucination checks. We further audit a stratified subset manually to estimate the quality of the retained data.

Stage	# Samples or Subset	Result
Raw generated traces	71,420	100.0%
Schema-valid traces	64,870	90.8% retained
Grounding-valid traces	60,215	84.3% retained
Action-consistent traces	57,640	80.7% retained
Hallucination-filtered traces	55,930	78.3% retained
Final WebWorld-CoT-55K	55,000	77.0% retained
Human audit	1,000 samples	91.8% pass
Major errors	1,000 samples	4.7%
Minor edits	1,000 samples	3.5%

Finally, we apply *hallucination filtering*. This stage removes traces that introduce unsupported entities, non-existent landmarks, unavailable pages, or reasoning steps that are not justified by either the retrieved web evidence or the grounded scene context. After these automatic checks, the dataset is reduced from 71,420 raw generated traces to 55,000 final examples, corresponding to a 77.0% retention rate.

Human validation. In addition to automatic filtering, we conduct human-in-the-loop auditing on a stratified subset of 1,000 examples. The subset is sampled across navigation, shopping, travel planning, geolocation, and action-prediction tasks to cover the major task categories in WebWorld-CoT-55K. Annotators inspect the task instruction, retrieved web evidence, scene context, hierarchical reasoning trace, and action sequence through an annotation interface.

For each audited example, annotators judge whether the reasoning trace is supported by the available evidence, whether the intermediate states are consistent, and whether the final action sequence matches the intended task. We categorize errors into major and minor types. Major errors include unsupported landmarks, invalid action mappings, inconsistent route-state transitions, or reasoning steps that cannot be grounded in the provided evidence. Minor edits mainly involve wording normalization, schema-format correction, or small phrasing inconsistencies that do not affect the semantic validity of the trace.

As shown in Tab. 16, the audited subset achieves a 91.8% human pass rate, with 4.7% major errors and 3.5% minor edits. These results indicate that the final dataset is not accepted directly from raw model outputs, but is filtered and audited to reduce hallucinated or structurally inconsistent reasoning traces.

Hierarchical Planning Prompt

System Prompt:
You are a schema-guided planner that decomposes a high-level goal into structured multi-level reasoning steps. Given a user query, generate reasoning chains in three abstraction levels: (1) *Macro-level* (goal formulation), (2) *Mid-level* (subgoal reasoning), and (3) *Micro-level* (execution alignment). Each reasoning step must follow the schema: `(Intent, Evidence, Intermediate State, Next Action)`.

Example:

Macro-level (Goal Formulation)

User: `Plan a trip from New York to Paris.`

Model:

```
Step 1 - Search for flights;
Step 2 - Find hotels near city center;
Step 3 - Explore landmarks;
Step 4 - Organize itinerary.
```

Mid-level (Subgoal Reasoning)

```
Intent: Search for flights from JFK to CDG.
Evidence: Retrieved flight schedule webpage.
Intermediate State: Available flights with duration and price.
Next Action: Select the shortest-duration flight.
```

Micro-level (Execution Alignment)

```
Web: Query "JFK to CDG flights" on travel site.
Scene: Visualize airport terminal and boarding gate.
Motion: Walk from check-in counter to gate A21.
```

Output Schema:

$$[\text{Goal}] \rightarrow [\text{Plan}] \rightarrow [\text{Observation}] \rightarrow [\text{Action}] \rightarrow [\text{Reflection}],$$

ensuring verifiable intermediate states and cross-modal alignment.

Figure 6: LLM prompt for hierarchical planning, illustrating three-stage reasoning (macro, mid, micro) for structured and cross-modal goal decomposition.

E LLM Prompts

E.1 Outdoor

For outdoor scenarios, we list two key system prompts following the schema-guided chain-of-thought (CoT) framework used in WebWorld-CoT-55K. These prompts are designed to integrate spatial grounding with hierarchical reasoning across web, scene, and motion modalities.

Location generation. As shown in Figure 5, this prompt instructs the model to describe outdoor landmarks and generate their approximate coordinates without explicitly naming the location. This encourages the model to rely on visual and geographic cues rather than directly copying place names. The generated landmarks are later verified for geographic and semantic plausibility.

Hierarchical planning. Figure 6 illustrates the hierarchical planning prompt, which decomposes high-level outdoor goals into three reasoning stages—*macro*, *mid*, and *micro*. Each stage refines the previous one, coupling web reasoning (e.g., retrieving travel or shopping information) with embodied navigation steps. This structured prompt enforces verifiable intermediate states and maintains cross-modal coherence throughout the reasoning process.

E.2 Geolocation

For geolocation reasoning, we design a structured system prompt shown in Figure 7, which emulates human spatial inference in GeoGuessr-style tasks. The model is instructed to analyze panoramic visual inputs and infer the most plausible location purely from environmental evidence such as language on signs, architectural style, vegetation type, and traffic patterns, without relying on any metadata. The output schema contains three explicit components: step-by-step reasoning describing the observed clues, estimated geographic coordinates in latitude–longitude format, and a hierarchical location prediction consisting of continent, country,

Geolocation

System Prompt:
You are an expert geolocation analyst.
In this task, you are given one or more panoramic images captured from a single coordinate on Earth. Your objective is to infer the most precise location of the viewpoint using only visual and contextual cues from the images. You must output only the following: (1) *Reasoning Steps*: Explain the visual and semantic evidence that led to your localization. (2) *Estimated Coordinate*: format as 'latitude, longitude'. (3) *Estimated Location*: format as 'Continent, Country, City, Street or Landmark'.
Rules:

- Do not use any external tools or metadata beyond the provided visual content.
- Use clues such as language on signs, architecture, vegetation, vehicles, and road markings.
- Be as specific as possible in your conclusions. When uncertain, make your best educated guess.

Example Output:
Reasoning Steps: The signage contains Japanese characters and the street pattern is narrow with utility poles visible, typical of Japanese urban residential areas. Vegetation and building density suggest a temperate climate. The architecture and colors resemble the Shinjuku district in Tokyo.
`Estimated Coordinate: 35.6926, 139.7074`
`Estimated Location: Asia, Japan, Tokyo, Shinjuku`

Figure 7: LLM prompt for geolocation reasoning, guiding the model to infer precise coordinates and locations solely from visual cues.

city, and landmark.

This prompt formulation constrains the model to produce grounded and verifiable spatial reasoning instead of memorized place recognition. In WebWorld-CoT-55K, such schema-guided prompts generate the geolocation subset of WebWorld-CoT-55K, with each result verified through human inspection and spatial consistency checks. During evaluation on WEBWORLDBENCH, the same schema ensures unified comparison between WebWorld and baseline LLM agents, supporting consistent assessment of visual–language–spatial reasoning under real-world uncertainty.

F Future Work

Future work will extend WebWorld to more dynamic and realistic environments, including live-web interaction, indoor and safety-critical tasks, and multi-agent settings. Another direction is efficiency-aware reasoning, such as budget-adaptive test-time scaling, early stopping, and distillation to reduce inference cost. We also plan to incorporate more human-generated interactions and richer modalities such as audio and video to improve ecological validity.